Power-aware Remote Replication for Enterprise-level Disaster Recovery Systems

Kazuo Goda　　　　　Masaru Kitsuregawa

Institute of Industrial Science, the University of Tokyo, Japan

{kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract

Electric energy consumed in data centers is rapidly growing. Power-aware IT, recently called 'green IT', is widely recognized as a significant challenge. Disk storage is a non-negligible energy consumer. Rather, in light of recent data-intensive systems where a number of disk drives are incorporated, the disk storage may be what we must consider primarily. Yet, all of the disk drives are not used for primary datasets, but rather larger portion of them are utilized for storing a variety of *copies* such as backups and snapshots. Saving the energy of such storage resources that manage copies is a promising approach. The paper presents a power-aware disaster recovery system, in which the reflection of transferred updated information can be deferred through eager compaction technique. Great energy saving of storage systems is expected in the remote secondary site. Our experiments using a commercial database system show that 80-85% energy of the secondary-site disk storage can be saved with small penalties of possible service breakdown time.

1 Introduction

Many attentions are paid on the energy consumption of IT systems, which has been grown up by 25% every year [9]. The recent analysis [2] reports that the annual electricity cost paid by the system owner will go up to twice higher than the annual server expense in 2009. More and more powerful cooling systems and power-supply equipments are being installed into data centers for accommodating the increase of energy consumption; accordingly, the electric energy and the related equipments account for 44% of TCO in a typical system [1]. In addition to the cost issue, energy and heat management has become a key of data center design and operation. The exploding energy consumption might strictly constrain the design space of modern IT systems [32]. Hence, energy saving is a grand challenge for IT research and development.

Storage systems are non-negligible energy consumer in IT systems. Storage systems present 27% of total energy consumption in a typical data center [23]. As the digital data volume is explosively increasing [16], extremely many disk drives are being incorporated into an enterprise system to improve the throughput. Thus, much larger portion of the total energy may be consumed by the storage system in high-performance data-intensive systems. Q. Zhu et al. in paper [34] points out an interesting example, where disk drives account for 71% of the total energy consumption in a large-scale OLTP system. Therefore, energy saving of the disk storage is rather essential as well as server processors and network devices.

Interestingly, all the disk drives of recent enterprise disk storage are not necessarily used for primary datasets. Rather, larger portion of the disk drives are utilized for storing a variety of *copies* to improve the system performance and availability. Suppose a simple IT system, which holds a single snapshot in a local data center and a backup copy in a remote data center. Two thirds of all the disk drives equipped in the total system are used for copies. Modern enterprise systems may use much more disk drives for copies [18]. Saving the energy of such storage resources should be a natural idea.

The paper proposes a power-aware disaster recovery system. It has been widely recognized that the business breakdown due to unpredictable disasters such as terrors and hurricanes provides a nation and a society with terrible damage [7, 25]. Business continuity is being enforced by nation-level legal systems as well as by enterprise-level internal disciplines [3, 26, 29]. The disaster recovery system [8, 15, 17] is a practical solution which places a remote secondary site and, in case of disaster, continues the business on the secondary site. By concentrating transferred update information through eager compaction technique, our proposed system can derive longer idle time of the data volume in order to reduce significantly the energy consumption of disk storage of the secondary site with small penalties of service break-
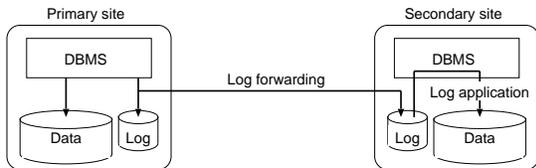
Figure 1: A disaster recovery system based on database log forwarding.

Figure 2: Deferred log application.

down time. In the real disaster recovery system, many resources of the secondary site may not be fully utilized when the primary site is normally operating. Great energy saving is expected in practice. To the best of our knowledge, similar researches have not been published.

This paper is organized as follows. Section 2 concisely describes disaster recovery systems that are currently deployed in many enterprise systems. Section 3 proposes novel techniques for energy saving of disaster recovery systems and Section 4 evaluates the proposal through the experiments using a commercial database system. Section 5 briefly summarizes related works and finally Section 6 concludes the paper.

2 Disaster recovery system

A disaster recovery system comprises two or more sites. Business is usually operated in the primary site and, once a disaster damages the primary site, the business is continued in the remote secondary site. For enabling such disaster recovery, up-to-date data of the primary site must be always copied into the secondary site. Many solutions have been proposed in papers and deployed into real systems, but the basic idea is similar in that they are composed of the following two steps: (1) transferring only updated information of the primary site to the secondary site and (2) reflecting it to the storage in the secondary site. Here the updated information means queries or transactions for the conventional logical database replication, changed blocks for the storage-level physical block forwarding [8, 17], and database log entries for the log forwarding [15, 27].

The recovery capability of such a disaster recovery system can be defined by two metrics: recovery point objective (RPO) and recovery time objective (RTO). RPO denotes possibility of data loss, i.e. how latest data can be recovered in case of disaster. RTO means inter-site takeover overhead, i.e. how soon the business can start again in the secondary site in case of disaster. It is preferable that RPO and RTO would be small. This paper focuses on enterprise-level systems such as brokerage and e-commerce, which accept only small service breakdown time and do never allow any data loss even in case of disaster. Thus, we assume here that inter-site data transfer
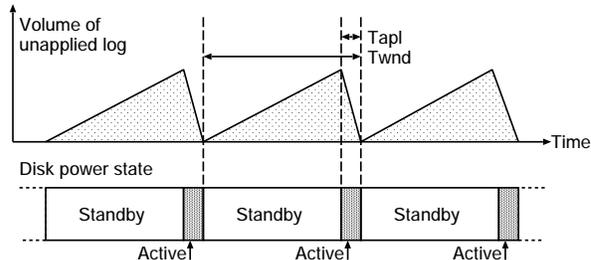
is done in the synchronous fashion; specifically, RPO is always zero. Needless to say, our contribution can be directly applied to more relaxed asynchronous modes.

3 Power-aware disaster recovery system

To save the energy consumption of disaster recovery systems, we focus on the disk storage in the secondary site, in which most of the storage resources are used only for storing backup copies.

Let us describe a scenario based on database log forwarding, which is recently deployed in high-end disaster recovery systems. Figure 1 illustrates a disaster recovery system based on database log forwarding, where physical database log is shipped from the primary site to the secondary site and the forwarded log is applied in the secondary site.

3.1 Deferred log application

Our idea is to *defer* the application of transferred database log to derive longer idle time of the data volume in the secondary site. Figure 2 illustrates a batch application scheme for realizing such deferred log application. In the secondary site, the transferred log is stored in the log volume and is not immediately applied to the data volume. While the database log is not being applied, the data volume is idle, so that the energy consumption of the data volume can be saved by spinning down the volume. Longer standby time gives greater energy saving, but provides larger amount of unapplied log stored in the log volume. In case of disaster, the secondary site must apply all the unapplied log before starting the business again. Therefore, deferability of log application is mainly determined by RTO requirements.

Let us discuss the relationship between RTO requirements and deferability. Let R_{gen} and R_{apl} be the log generation rate in the primary site and the maximum log application rate in the secondary site respectively. We assume that the recovery time of the secondary site is proportional to the amount of unapplied log[1]. The fol-

---

[1]Fast log application is a key to quick recovery [19]. Other marginal

Table 1: Typical OLTP systems

Rank	Vendor	System	tpmC	Database	# of disks (data volume)	# of disks (log volume)
1	IBM	System p5 595	4,033,378	IBM DB2 9	6400	360
2	IBM	eServer p5 595	3,210,540	IBM DB2 UDB 8.2	6400	140
3	IBM	eServer p5 595	1,601,784	Oracle Database 10g	3200	96
4	Fujitsu	PRIMEQUEST 540 16p/32c	1,238,579	Oracle Database 10g	1920	224
5	HP	Integrity Superdome 64p/64c	1,231,433	Microsoft SQL Server 2005	1680	56

Quoted from *Top Ten TPC-C by Performance Version 5 Results* disclosed at `http://www.tpc.org/` as of December 11, 2006.

lowing formulae give the optimal batch configuration: a batch interval T_{wnd} and a necessary log application time in the interval T_{apl}, the combination of which can gain maximum energy conservation. For the page limitation, we have to omit the mathematical proof.

$$T_{wnd} = \frac{R_{apl}^2}{(R_{apl} - R_{gen}) \cdot R_{gen}} T_{RTO}$$

$$T_{apl} = \frac{R_{apl}}{R_{apl} - R_{gen}} T_{RTO}$$

Here T_{RTO} denotes a RTO requirement, i.e. given allowance of service breakdown time. Implicitly, $R_{apl} > R_{gen}$ and $T_{RTO} > T_{up} + T_{down}$, where T_{up} and T_{down} denote time penalties of spinning up and down respectively. The secondary site can concentrate log application based on the above batch configuration and proactively spin up and down the data volume in order to save energy consumption of the disk storage.

3.2 Eager log compaction

Obviously, larger $\frac{R_{apl}}{R_{gen}}$ leads to greater energy saving. In this section, we introduce *eager compaction* technique to improve the log application throughput significantly.

In a disaster recovery system based on log forwarding, the transferred log entries are applied to the data volume in a way similar to database redo operation. In the normal redo operations, log entries are applied strictly in log sequence number (LSN) order. On the contrary, our proposed eager techniques can compact the log sequence in a window buffer and apply the compacted sequence to the data volume. The compaction process comprises *log folding* and *log sorting*. Log folding is a technique to reduce the number of log entries to be applied. The log entries which manipulate the identical record are coalesced in the window buffer. For example, assuming that three log entries, `insert(data1)`, `update(data1 → data2)` and `update(data2 → data3)`, are given in the sequence, these three entries, manipulating the same record, can be logically folded into a single entry, `insert(data3)`. On the other hand, log sorting reorders log entries in the window buffer to im-

prove disk access sequentiality. In many cases, an entry of database log has a physical reference to the target record. Log sorting leverages such physical information. Both the methods together can improve the log application throughput. Accordingly, larger energy saving can be expected.

3.3 Discussion

Here we would like to briefly discuss our contribution to the total energy consumed by the secondary-site storage. Table 1 shows top-five systems quoted from "Top Ten TPC-C by Performance Version 5 Results". Four systems used only 2.2-5.6% of all the disk drives for database log, and the other system used only 11.6% for log. That is, in the secondary site, only very few disk drives must be always spinning actively and the other disk drives can be spun down by the combination of deferred log application and eager log compaction. The contribution of our idea is still significant on the whole storage system in the secondary site.

Although this section discusses the problem mainly based on the database log forwarding, the proposed method can be easily extended to other remote replication methods such as logical database replication and physical block forwarding. Specifically, eager compaction technique can be directly applied to physical block forwarding. Forwarded blocks can be folded and sorted similarly based on physical block address. On the other hand, slight modification is necessary for logical database replication, since queries and transactions described in SQLs are not aware of physical addresses. Queries and transactions should be scheduled with assistance of batch query scheduling techniques [21, 24]. So far, such compaction techniques of updated information were intended for reducing inter-site traffic [17]. In contrast, our attempt is focused on deriving long idle period for energy saving.

4 Evaluation

This section presents validating experiments using TPC-C benchmark, showing that disk power consumption of the secondary site can be saved with slight degradation of the quality of business continuity. Online transactions

---
recovery overheads are out of the scope of this paper.

Table 2: Basic parameters of disk drive models.

Model	IBM Ultrastar 36Z15	HGST Deskstar T7K250
Capacity	18.4 GB	250 GB
Rotational speed	15000 rpm	7200 rpm
Avg. seek time	3.4 ms	8.5 ms
Transfer rate	55 MB/s	61 MB/s
Active power	39.0 W	9.7 W
Idle power	22.3 W	5.24 W
Standby power	4.15 W	(U) 4.04 W
		(L) 2.72 W
		(N) 0.93 W
Spin-down penalties (time and energy)	15.0 s 62.25 J	(U) 0.7 s, 3.5 J (L) 17.0 s, 19.0 J (N) 0.7 s, 3.5 J
Spin-up penalties (time and energy)	26.0 s 904.8 J	(U) 0.7 s, 3.5 J (L) 17.0 s, 19.0 J (N) 0.7 s, 3.5 J

(U): unloaded mode, (L): low-rpm mode, (N): non-spinning mode

(a) 16 warehouses

(b) 160 warehouses

Figure 3: Power saving of secondary-site disk storage with high-end disk drives.

Table 3: Batch interval for 160 warehouses and high-end disks under 100 seconds of RTO requirement.

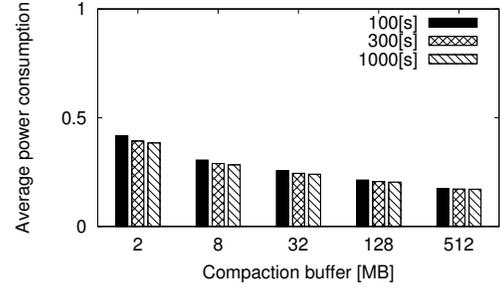| Window buffer	32 MB	128 MB	512 MB
Batch interval	536 s	1070 s	2150 s

are typical workloads that are seen in enterprise-level disaster recovery systems.

We prepared a hybrid simulation environment for measuring the potential energy saving due to the proposed system. In the experiment, we used a disk drive simulator which can calculate energy consumption based on a disk drive model. We implemented deferred log application and eager log compaction on the top of the disk drive simulator. The developed log applier can apply database log generated by HiRDB [15], a commercial database system.

The experiments were done on a Linux server with dual Xeon processors and 2GB main memory. We set up TPC-C benchmark with 16 and 160 warehouses respectively, and we generated database log on each configuration by executing one million transactions using HiRDB plus 512MB database buffer with no think time. At this execution, we also traced IO behavior by using a kernel-level IO tracer. Then, by replaying the traced IOs using a disk drive model in the simulation environment, we simulated log generation at the primary site. Here, we assumed that the primary site processed transactions at the maximum rate on the specified disk drive model. Next, we applied the generated database log by the log applier, and measured the power reduction effect at the secondary site. Throughout this experiment, we followed the storage system configuration of "IBM System p5 595" in Table 1. That is, we assumed that 94.4% of disk drives were used for the data volume and the same type of disk storage was used both in the primary and secondary sites. The experiments were conducted for different RTO requirements and different window buffer lengths. For validation, we compared the energy saving of the proposed power-aware system with the conventional system in which the transferred update information is immediately reflected.
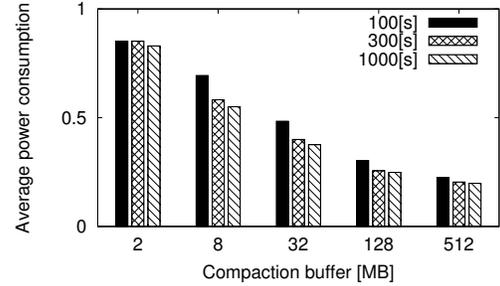
Figure 3 summarizes the results obtained with a high-end disk drive model. Basic parameters of the model are presented in Table 2. This model, which is based on IBM Ultrastar 36Z15, may not be new, but has been used in many previous papers [4, 20, 33, 34]. In the graphs, each bar, denoting the average power consumption of the disk storage in the secondary-site storage, is normalized by that of the conventional system. Larger window buffer could accelerate the log application throughput more, and accordingly, greater power saving was gained. Note that, by using 512 MB window buffer, which is as large as the database buffer of the primary site, 85% of the power could be conserved totally in the secondary-site storage. Such great saving was supported by accelerated log application; $\frac{R_{apl}}{R_{gen}}$ could speed up to 20.5 (W=16) and 49.3 (W=160) at maximum by eager log compaction. On the other hand, more tolerant RTO requirements could lead to more energy saving, but its contribution was slight. In our experiments, only short RTOs (such as 30 seconds) failed because RTOs were shorter than time penalties of spinning up and down the volume, but moderate RTOs (100 seconds and more) could conserve the energy so much.

Let us consider the batch interval T_{wnd}. Frequent transition of energy modes affects the life time of disk drives. Table 3 summaries sampled values of T_{wnd}. With small buffer, the data volume had to change its modes frequently, but with as large buffer as 512 MB, the disk mode changes only 40 times a day. Note that this frequency was given when the primary site generates database log at top speed. Thus, it looks almost acceptable, since many high-end disk drives support at least 50,000 cycles of starts/stops. Of course, more tolerant RTO gives larger intervals. This analysis reveals that eager compaction is a key technique to the longevity of disk drives.

We conducted the experiments using a recent midrange disk drive, which has new energy-efficient features [14]. Basic parameters of the model are presented in Table 2 too. The disk drive, which is based on HGST Deskstar T7K250, has three standby states: unload, lowrpm and non-spinning (equal to conventional standby mode). Basically, the proposed disaster recovery system could work with mid-range disk drives similarly. But, since recent mid-range disk drives can change the energy modes with much smaller time penalties than high-end disk drives, the proposed system could work for such small RTOs as 30 seconds. Figure 4 compares these three standby modes with 160 warehouses. By using the non-spinning standby mode, 80% saving was gained at maximum in comparison with the conventional system. However, we cannot observe the substantial benefit of using new energy-saving functions such as unload and low-rpm.

In summary, the proposed power-aware disaster recovery system can achieve great energy saving of the secondary-site disk storage without little harm to the recovery capability. Only 100 seconds and 30 seconds of RTO allowance were needed for high-end disks and midrange disks respectively. This observation is surprising, since strict high-availability systems, as known as five nines (99.999%), allow only 315 seconds of breakdown per year. Our proposal can be promising in such topdrawn disaster recovery systems.

5 Related works

Research communities have presented various attempts for energy-efficient disk storage.

The simplest approach is to transition disk drives to a low-power mode after the predetermined time period has elapsed after the last disk access. This technique is widely deployed in commercial disk drives. More sophisticated techniques that try to tune the threshold adaptively have been also studied [6, 11]. Such threshold based techniques work effectively for battery-operated mobile and laptop computers. However, it looks difficult
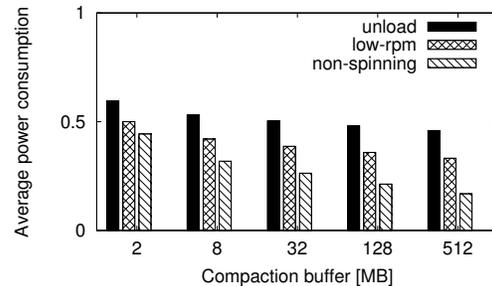

Figure 4: Power saving of secondary-site disk storage with mid-range disk drives under 30 seconds of RTO requirement.

to directly apply these techniques to enterprise systems.

Massive Array of Idle Disks (MAID) [5] and Popular Data Concentration (PDC) [4] are alternative approaches that migrate/replicate popular blocks on specific disk drives to create long idle period of the other disk drives. These techniques leveraging access locality are deployed in real archival storage systems.

Exploiting redundancy information and large cache space that RAID capability holds seems a reasonable approach. Energy Efficient RAID (EERAID) [20] and RIMAC [33] can arrange IO requests at RAID controllers so as to avoid evicting out blocks that are originally stored in spun-down disk drives as much as possible. Power-Aware RAID (PARAID) [30] introduces an asymmetric parity placement on the legacy RAID-5 so that the system can dynamically change the number of actively spinning disk drives.

Other researchers [12, 34] have actively studied on multi-speed disk drives which have the capability of changing the rotational speeds. These attempts look very effective. However, to our knowledge, such multi-speed disk drives are still limited in experimental prototypes and not yet seen in the market.

Recently several application-assisted approaches for storage energy conservation have been reported. Cooperative IO [22, 31] is a set of power-aware IO system calls, by which the user can specify deferability and abortability to each IO. Compiler-based application transformation [10, 13, 28] tries to arrange IO commands in source code levels in order to concentrate IO requests.

Our work differs from these previous works in that we are trying to fully leverage the characteristics of the secondary-site disk storage. That is, the disk storage there manages only copies and its resources are not necessarily busy when the primary site is alive. Our eager strategy of concentrating database log can provide long idleness to many disk drives, accordingly obtaining substantial energy saving opportunities.

6 Conclusion

The paper proposes a power-aware disaster recovery system, in which the reflection of transferred updated information can be deferred through eager compaction technique, so as to gain great energy saving of storage systems in the remote secondary site. Experiments with a commercial database system showed that 80-85% energy consumption can be conserved in the secondary-site disk storage with small penalties of possible service breakdown time.

In this paper, we focus on the energy consumption of disk drives which are main components of recent disk storage. Further, we would like to extend our approach so as to provide a system-wide analysis considering RAID controllers and cache memory.

References

[1] APC. Determining Total Cost of Ownership for Data Center and Network Room Infrastructure. White paper, 2002.

[2] B. RUDOLPH. Storage In an age of Inconvenient Truths. Storage Network World Spring 2007, 2007.

[3] BRITISH STANDARDS INSTITUTION. BS25999: Business Continuity Managemenet, 2006.

[4] CARRERA, E. V., PINHEIRO, E., AND BIANCHINI, R. Conserving Disk Energy in Network Servers. In *Proc. Int'l Conf. on Supercomputing* (2003), pp. 86–97.

[5] COLARELLI, D., AND GRUNWALD, D. Massive Arrays of Idle Disks for Storage Archive. In *Proc. ACM/IEEE Conf. on Supercomputing* (2002), pp. 1–11.

[6] DOUGLIS, F., KRISHNAN, P., AND BERSHAD, B. Adaptive disk spin-down policies for mobile computers. In *Proc. USENIX Symp. on Mobile and Location-Independent Computing* (1995), pp. 121–137.

[7] EAGLE ROCK ALLIANCE. Online Survey Results: 2001 Cost of Downtime, Contingency Planning Research, 1996.

[8] EMC CORP. Symmetrix Remote Data Facility product description guide, 2000.

[9] F. MOORE. More power needed. Energy User News, 2002.

[10] GNIADY, C., HU, Y. C., AND LU, Y.-H. Program Counter-Based Prediction Techniques for Dynamic Power Management. *IEEE Trans. Comput. 55*, 6 (2006), 641–658.

[11] GOLDING, R. A., BOSCH, P., STAELIN, C., SULLIVAN, T., AND J.WILKES. Idleness is not sloth. In *Proc. USENIX Tech. Conf.* (1995), pp. 201–212.

[12] GURUMURTHI, S., SIVASUBRAMANIAM, A., KANDEMIR, M., AND FRANKE, H. Reducing Disk Power Consumption in Servers with DRPM. *IEEE Computer 36*, 12 (2003), 59–66.

[13] HEATH, T., PINHEIRO, E., HOM, J., KREMER, U., AND BIANCHINI, R. Application Transformations for Energy and Power-Aware Device Management. In *Proc. Int'l Conf. on Parallel Arch. and Compilation Tech.* (2002), pp. 121–130.

[14] HGST INC. Quietly cool. White Paper, HGST, 2004.

[15] HITACHI LTD. Hitachi Relational Database Management System Solutions for Disaster Recovery to Support Business Continuity. Review Special Issue, Hitachi Technology, 2004.

[16] IDC. The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010. An IDC White Paper sponsored by EMC.

[17] JI, M., VEITCH, A., AND WIKES, J. Seneca: remote mirroring done write. In *Proc. USENIX Conf. on File and Storage Tech.* (2003), pp. 253–268.

[18] KLEIMAN, S. Trends in Managing Data at the Petabyte Scale. Invited talk, USENIX Conf. on File and Storage Tech., 2007.

[19] LAHIRI, T., GANESH, A., WEISS, R., AND JOSHI, A. Fast-Start: Quick Fault Recovery in Oracle. White paper, 2001.

[20] LI, D., WANG, J., AND VARMAN, P. Conserving Energy in Conventional Disk based RAID Systems. In *Proc. Int'l Workshop on Storage Network Arch. and Parallel I/Os* (2005), pp. 65–72.

[21] LU, H., AND LEE TAN, K. Batch Query Processing in Shared-Nothing Multiprocessors. In *Proc. Int'l. Conf. on Database Syst. for Advanced Applications* (1995), pp. 238 – 245.

[22] LU, Y., BENINI, L., AND MICHELI, G. Power-aware operating systems for interactive systems. *IEEE Trans. Very Large Scale Integration Syst. 10*, 2 (2002), 119–134.

[23] MAXIMUM THROUGHPUT INC. Power, heat, and sledgehammer. White paper, 2002.

[24] MEHTA, M., SOLOVIEV, V., AND DEWITT, D. J. Batch Scheduling in Parallel Database Systems. In *Proc. IEEE Int'l. Conf. on Data. Eng.* (1993), pp. 400 – 410.

[25] NATIONAL CLIMATE DATA CENTER, U.S. DOC. Climate of 2005 Atlantic Hurricane Season. Online Report available at http://www.ncdc.noaa.gov/oa/climate/research/2005/hurricanes05.html, 2005.

[26] NATIONAL FIRE PROTECTION ASSOCIATION. NFPA1600: Standard on Disaster/Emergency Management and Business Continuity Programs (2004 Edition), 2004.

[27] ORACLE CORP. Oracle Data Guard 11g. The Next Era in Data Protection and Availability. White paper, 2007.

[28] SON, S. W., MANDEMIR, M., AND CHOUDHARY, A. Software-Directed Disk Power Management for Scientific Applications. In *Proc. IEEE Parallel and Distributed Processing Symp.* (2005), p. 4b.

[29] U.S. SEC. General Rules and Regulations promulgated under the Securities Exchange Act of 1934, 2005.

[30] WEDDLE, C., OLDHAM, M., QIAN, J., WANG, A. A., REIHER, P., AND KUENNING, G. PARAID: A Gear-Shifting Power-Aware RAID. In *Proc. USENIX Conf. on File and Storage Tech.* (2007), pp. 245–260.

[31] WEISSEL, A., BEUTEL, B., AND BELLOSA, F. Cooperative I/O – A Novel I/O Semantics for Energy-Aware Applications. In *Proc. USENIX Symp. on Operating Syst. Design and Imple.* (2002), pp. 117–130.

[32] WORTH., S. Green Storage. SNIA Education, 2006.

[33] YAO, X., AND WANG, J. RIMAC: A Novel Redundancy-based Hierarchical Cache Architecture for Energy Efficient, High Performance Storage System. In *Proc. EuroSys* (2006), pp. 249–262.

[34] ZHU, Q., CHEN, Z., TAN, L., ZHOU, Y., KEETON, K., AND WIKES, J. Hibernator: Helping Disk Arrays Sleep through the Winter. In *Proc. ACM Symp. on Operating Syst. Principles* (2004), pp. 177–190.