This paper appeared in the Proceedings of the 12th IEEE Symposium on Computers and Communications (ISCC'08), Marrakech, Morocco, July 2008.

Improving Reliability and Energy Efficiency of Disk Systems via Utilization Control

Kiranmai Bellam[†], Adam Manzanares[†], Xiaojun Ruan[†], Xiao Qin[†]*, and Yiming Yang[‡]
[†]*Department of Computer Science and Software Engineering*
Auburn University
Auburn, Alabama 36830, USA
{kzb0008, acm0008, xzr0001}@eng.auburn.edu, xqin@auburn.edu

[‡]*Intel Corporation*
Rio Rancho, NM 87124
yiming.yang@intel.com

Abstract

As disk drives become increasingly sophisticated and processing power increases, one of the most critical issues of designing modern disk systems is data reliability. Although numerous energy saving techniques are available for disk systems, most of energy conservation techniques are not effective in reliability critical environments due to their limitation of ignoring the reliability issue. A wide range of factors affect the reliability of disk systems; the most important factors – disk utilization and ages – are the focus of this study. We build a model to quantify the relationship among the disk age, utilization, and failure probabilities. Observing that the reliability of a disk heavily relies on both disk utilization and age, we propose a novel concept of safe utilization zone, where energy of the disk can be conserved without degrading reliability. We investigate an approach to improving both reliability and energy efficiency of disk systems via utilization control, where disk drives are operated in safe utilization zones to minimize the probability of disk failure. In this study, we integrate an existing energy consumption technique that operates the disks at different power modes with our proposed reliability approach. Experimental results show that our approach can significantly improve reliable while achieving high energy efficiency for disk systems.

1. Introduction

Disk drives started the new era in the computer dictionary. Therefore, it is crucial to have highly reliable, energy efficient and cost effective disks. Extensive research has been done (and continues to be done) to reduce the energy consumption of disks. Literature proves that very little or no research has been done to maximize reliability and energy efficiency of the disks. Reliability may be the most important characteristic of the disk drives if the data stored on the disk drives is mission critical. Disk drives are generally very reliable but may fail. In this paper disk age and utilization are studied as the major factors that affect disk drive reliability. In this paper we studied the failure probabilities with respect to disk age and utilization. This relationship is made explicit in the form of graphs. These graphs are used to estimate safe utilization levels for disks of differing ages. Safe utilization zones are the range of utilization levels where the probabilities of disk failures are minimal. Simulations prove that when disks are operated in the safe utilization zones, the probabilities of failures are effectively minimized.

It is evident that disk drives consume large amounts of energy. Reducing the energy consumption not only lowers electricity bills but also tremendously reduces the emissions of air pollutants. Energy consumption is reduced by operating the disks at three power modes: Active, Idle and Sleep. Mirroring disks (a.k.a., RAID1) is considered for the experiments, which uses a minimum of two disks; one primary and one back up. The data is mirrored to the backup disk from the primary disk. Traditional methods wake up the backup disk when the utilization of the primary disk exceeds 100 percent. Load balancing technique keeps both the primary and back up disks always active to share the load. These methods consume a massive amount of energy, as the disks stay active even when there are no

requests to serve for a long period of time. The reliability of these disks is also ignored most of the time. We designed a policy where we operate the disks at different power modes based on the utilization of the disk. The utilization of the disk is calculated a priori. Along with the power conservation we also aim at achieving high reliability. Thus, we must operate the disks only in safe utilization zones.

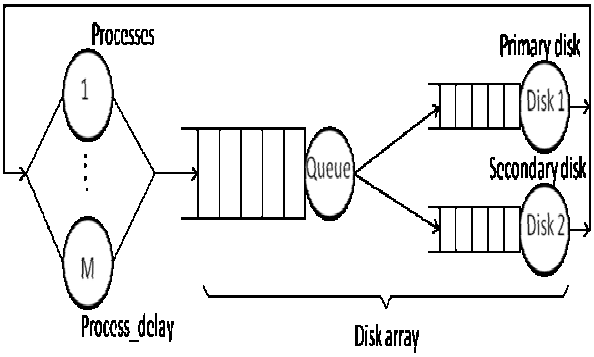

Fig 1. Queuing model of RAID 1 with read workloads

In the policy we defined, the processor generates the read requests (See Fig 1) that should be processed by the disks. These read requests are queued up in a buffer and the utilization levels are calculated. These utilization levels are checked against the limits of the safe utilization zone and then the disk modes are changed accordingly. This approach is not only reliable but also saves a significant amount of energy. This approach saves more energy for Travelstar disks when compared with Ultrastar disks, because the energy levels of the disk parameters like spin up, spin down, active power, idle power and etc. are much higher for Ultrastar disks when compared to Travelstar disks. It should be noted here that the disk parameters play a major role in the energy consumption of the disk.

The rest of the paper is organised as follows. In the next section we discuss the related work and motivation. In section 3, we describe the disk failure model. Reliability aware power conservation model is explained in section 4. In section 5, we evaluate the performance RARE based on real world traces. Section 6 concludes the paper with the summary.

2. Related Work

Extensive research has been carried out in developing energy efficient storage systems. Dynamic voltage scaling [3][9][13], dynamic power management[17], compiler directed energy optimizations [15][16] are some of the state of the art energy conservation techniques.

Du el al. studied the dynamic voltage scaling technique with a real-time garbage collection mechanism to reduce the energy dissipation of flash memory storage systems [18]. A dynamic spin down technique for mobile computing was proposed by Helmbold et al. [4]. A mathematical model for each Dynamic Voltage Scaling - enabled system is built and their potential in energy reduction is analyzed by Lin Yuan and Gang Qu [13]. Carrera et al. [7] proposed four approaches to conserving disk energy in high-performance network servers and concluded that the fourth approach, which uses multiple disk speeds, is the one that can actually provide energy savings.

An energy saving policy named eRAID [2] for conventional disk based RAID-1 systems using redundancy is given by Li et al. Energy efficient disk layouts for RAID-1 systems [12] have been proposed by Lu et al. Yue et al. investigated the memory energy efficiency of high-end data servers used for supercomputers [10]. Son et al. proposed and evaluated a compiler-driven approach to reduce disk power consumption of array-based scientific applications executing on parallel architectures [15][16][9].

Pinheiro et al presented failure statistics and analyzed the correlation between failures and several parameters generally believed to impact longevity [5]. Four causes of variability and an explanation on how each is responsible for a possible gap between expected and measured drive reliability, are elaborated by Elerath and Shah [8]. Dempsey, a disk simulation environment that includes accurate modeling of disk power consumption is presented by Zedlewski et al [11]. They also demonstrated that disk power consumption can be simulated both efficiently and accurately. Optimal power management policies for a laptop hard disk are obtained with a system model that can handle non-exponential inter-arrival times in the idle and the sleep states [17]. Schroeder and Gibson presented and analyzed the field-gathered disk replacement data from five systems in production use at three organizations [1]. They found evidence that failure rate is not constant with age, and that there was a significant infant mortality effect. The infant failures had a significant early onset of wear-out degradation. Significant levels of correlation between failures, including autocorrelation and long-range dependence, were also found.

Gurumurthi et al. [14] provided a new approach called DRPM to modulate disk speed (RPM) dynamically, which gives a practical implementation to exploit this mechanism. They showed that DRPM can provide significant energy savings without heavily

compromising performance. Rosti et al. presented a formal model of the behavior of CPU and I/O interactions in scientific applications, from which they derived various formulas that characterize application performance [6]. All of the previously mentioned work either concentrated on the power conservation or on the disk reliability. Not many researchers address both energy efficiency and reliability. It is very important for a data disk to be very reliable, while consuming less power. The importance of energy efficiency and reliability, and the lack of research of their relationship, motivates the research conducted in this paper.

3. Disk Failure Model

Let *Z+* be the set of positive integers. Without loss of generality, we consider a workload condition where there are $m \in Z+$ disk I/O phases. The utilization U_i of the *i*th ($1 \le i \le m$) I/O phase is a constant that can be straightforwardly derived from the disk I/O requirements of data-intensive tasks running within the *i*th I/O phase. Let ϕ_i be the number of data-intensive tasks running in the *i*th phase. Let $\lambda_{ij}, 1 \le j \le \phi_i$, denote the arrival rate of disk request submitted by the *j*th data-intensive task to the disk system. Let s_{ij}, $1 \le j \le \phi_i$ be the average data size of disk requests of the *j*th task. The disk I/O requirement of the *j*th task in the *i*th phase is a product of the task's request arrival rate and the average data size of disk requests issued by the task, i.e., $\lambda_{ij} \cdot s_{ij}$. The accumulative disk I/O requirements R_i, measured in terms of MByte/Sec., of all the tasks running in the *i*th phase can be written as:

$$R_i = \sum_{j=1}^{\phi_i} \left(\lambda_{ij} \cdot s_{ij} \right). \quad (1)$$

Note that R_i in Eq. (1) can be environed as accumulative data amount access per time unit.

The utilization of a disk system within a given I/O phase equals to the ratio the accumulative disk requirement R_i and the bandwidth of the disk system. Thus, the utilization U_i of the *i*th I/O phase can be expressed as

$$U_i = \frac{Ri}{B_{disk}} = \frac{\sum_{j=1}^{\phi_i} \left(\lambda_{ij} \cdot s_{ij} \right)}{B_{disk}}, \quad (2)$$

where B_{disk} is the bandwidth of the disk system.

The utilization *U* of the disk system during the *m* I/O phases is the weighted sum of the disk utilization of all the *m* phases. Thus, the utilization *U* is expressed by Eq. (3) as follows:

$$U = \frac{R_i}{\sum_{i=1}^{m} R_i} \cdot U_i = \frac{\left[\sum_{j=1}^{\phi_i} \left(\lambda_{ij} \cdot s_{ij} \right) \right]^2}{B_{disk} \cdot \sum_{i=1}^{m} \sum_{j=1}^{\phi_i} \left(\lambda_{ij} \cdot s_{ij} \right)}. \quad (3)$$

Given I/O requirements of data-intensive tasks issuing disk request to a disk system, one can leverage the above model to quantify utilization of the disk system.

To determine the failure rate for a given utilization rate, we took the points from the google study [5] and used the cubic spline interpolation method to approximate annual failure rate of a disk with certain utilization and age. The disk failure model can be modeled as an n+1dimensional vector $\vec{\theta} = [\theta_0, \theta_2, ..., \theta_n]$, where θ_i, $0 \le i \le n$, is the vector $\theta_i = (x_i, y_i)$ that captures the correlations between utilization x_i and disk failure rate y_i.

To develop the disk failure rate model, we have to determine n+1dimensional vector $\vec{\theta}$. Thus, given the value of utilization x_i, one can make use of the failure rate model to calculate the failure rate y_i in component θ_i of vector $\vec{\theta}$. To achieve this goal, we adopted the cubic spline interpolation to generate failure rates n+1dimensional vector $\vec{\theta}$ such that it results in a smooth curve for a given disk age.

Cubic spline generates $y = f(x) = ax3 + bx2 + cx + d$

That is a cubic function for each interval is used to plot the annual failure rate percentiles for disk systems of ages 2 years (see Fig. 2.) and as functions of disk utilizations.

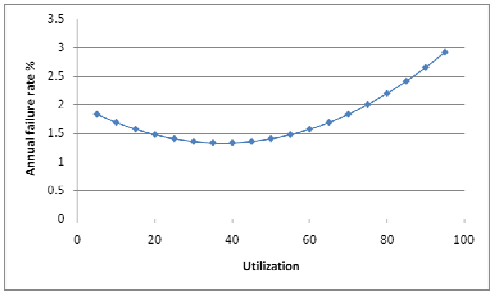

Fig. 2. AFR for 2 year old disk with respect to utilization.

Interestingly, results shown in Figs. 2- contradict the findings of the previous studies that indicate that lower utilization levels produce lower failure rates whereas

higher utilization levels correspond to higher failure rates. The trend of a disk annual failure rate as the disk utilization grows is different from that reported in the literature. Specifically, our disk failure rate model built from real-world data of disk failures suggests that the probability that a disk fails under either very low or very high utilization levels is very high. In contrast, when the utilization levels stay within a certain range (e.g., between 20% and 60% when the disk age is 2 years old, see Fig. 2), the probability of failures is far smaller than a specified threshold (e.g., smaller than 2%). We term this range of utilization levels as the safe utilization zone, which largely depends on disk ages. Given a disk system, its safe utilization zone is a function of the disk's age and the specified low failure rate threshold.

4. Reliability-aware power conservation model

RAID 1 is popular and is widely used for disk drives. RAID 1 is implemented with a minimum of two disks, which are the primary and back disks. Initially the data is stored to the primary disk and then it is mirrored to the backup disk. This mirroring helps to recover the data when there is a failure in the primary disk. It also helps to increase the performance of the RAID 1 system by sharing the workload between the disks. We considered RAID 1 for all of our experiments.

The processor in the system generates the I/O stream, which is queued to the buffer. The utilization of the disk is calculated using the request arrival rate. Please refer to Section 3 for details of the description for the disk utilization model.

It should be noted that all requests here are considered as read requests. At any given point of time the disks can be in the following three states.

- State 1: Both the disks in sleep mode
- State 2: Primary disk active and backup disk in sleep mode
- State 3: Both the disks in active mode and share the load.

Let us consider that the disks are in state 1 at the beginning. Once the utilization is calculated, it is compared with the safe utilization zone range. If the calculated value falls below the range then disks stay in state 1. If the calculated value is within the range, then the primary disk is made active while the backup disk continues to stay in the sleep mode. This represents a transition to state 2. If the calculated value is beyond the range then both the disks are made active and both of them share the load, which corresponds to state 3. Transition of states from one power mode to another involves disk spin up and/or spin down. The disk spin ups and spin downs also consume a lot of energy.

5. Performance Evaluation

RAID 1 is used in our experiments. RAID 1 uses a minimum of two disks, one as a primary and one as the backup. We conducted the experiments on three types of disks from IBM.

The experimental results are compared against two traditional state of the art methods. In the first method, load balancing, both the disks are always made active. Load balancing achieves very high performance because both the disks share the load. The second method, traditional method, is where the primary disk is made always active and the backup disk is kept in sleep mode. The backup disk is made active only when the utilization of the primary disk exceeds 100%, also known as saturation. In what follows, we term our approach as RAREE (Reliability aware energy efficient approach).

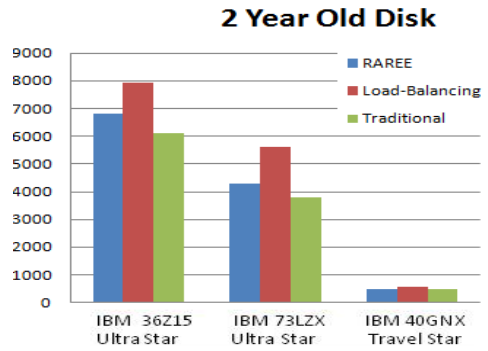

Fig. 3. Energy consumed by a 2 year old disk

The experimental data generated from the simulations is plotted in figure 3. Figure 3 represents the energy consumed by the 2 year old disk respectively. In the above figure RAREE represents the algorithm that is presented in this paper. RAREE is compared against load balancing and the traditional method. From fig 3 it is observed that for the IBM 36Z15 disk the power consumed by RAREE falls in between the load balancing and traditional techniques. Even for the IBM 73LZX the trend is similar, but the difference in values is not as high as IBM 36Z15. For the IBM 40GNX the power consumed by RAREE is smaller than the traditional and load balancing power consumption values because disk spin up and spin down values are much smaller for the IBM 40GNX when compared with the other two disks. It should be observed that the disk spin down and disk spin up values play a vital role in the energy consumption.

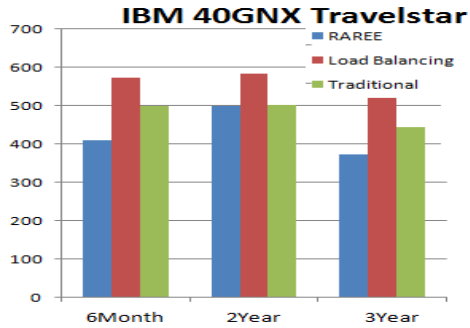

Fig. 4. Energy consumed by a IBM 40GNX disk with different ages

Fig 4 shows the performance of RAREE on Travelstar disks of different ages. It can be observed from figure that RAREE consumes less energy when compared to traditional and load balancing.

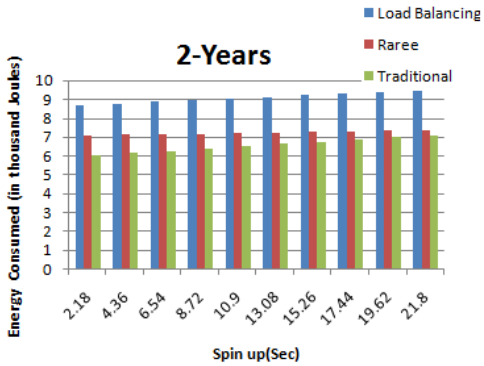

Fig. 5. Impact of spin up power on energy

Fig 5 shows the effects on energy when the spin up energy is varied for a 2 year old disk. The RAREE energy consumption falls in between traditional and load balancing techniques. Though the energy consumed by RAREE is a little higher than traditional technique, here we are also gaining good amount reliability.

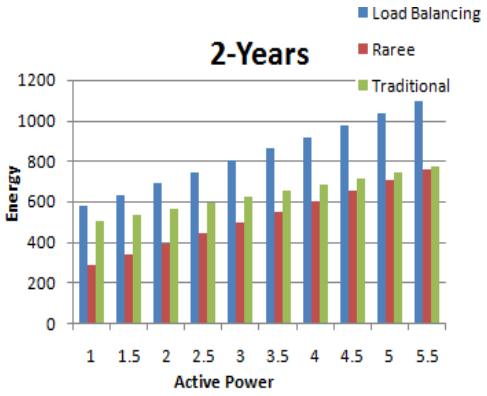

Fig. 6. Impact of active power on energy

Fig 6 shows the change in energy as the active power is varied. Here RAREE energy consumption is definitely less than the two existing techniques, because RAREE makes the disks go to sleep mode as soon there are no requests unlike the other techniques. When idle power is changed unlike the active power the energy consumed by the RAREE again falls between the two techniques. This is because RAREE makes the system go to sleep mode very often depending on the conditions. It should be observed here though the RAREE consumed a bit higher energy than traditional it can be neglected as we are achieving reliability.

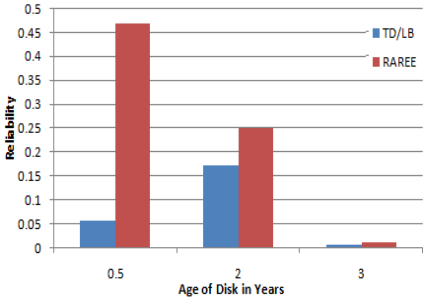

Fig. 7. Reliability vs. disk ages

Fig 7 is a very important graph here it shows the reliability in terms of annual failure rate percentile. It can be observed from the graph that RAREE achieves a very high reliability when compared to load balancing and traditional. Only one bar is shown for load balancing and traditional techniques because both have the same reliability levels as they don't pay special attention to reliability. We also found an interesting observation that when RAREE is applied to IBM40GNX, which is a travelstar, it definitely consumes much less energy than the other two Ultrastars, which are high performance disks. This makes it clear that RAREE gives best results when it is used on mobile disks instead of high performance disks. This doesn't limit the usage of RAREE to mobile disks because though Ultrastar consumes a little more energy than traditional technique we still get a good reliability at a marginal cost of energy.

Simulation results prove that on an average roughly 20% of energy can be saved when RAREe is used instead of load balancing. When RAREe is used instead of the traditional method an excess of 3% of energy is saved, it is not a very significant amount but along with a very little energy saving we are also achieving high reliability which makes it significant.

6. Summary

Although an array of energy conservation schemes have been proposed for disk systems, most energy conservation techniques are implemented at cost of

disk reliability. In order to solve this problem, we first build a model to quantify the relationship among the disk age, utilization, and failure probabilities. In this study, we focused on mirroring disk systems, where data sets are mirrored to backup disks from primary disks. Traditional methods wake up the backup disks when the utilization of the primary disks exceeds 100 percent. Load balancing technique keeps both the disks always active to balance the load between primary and backup disks to achieve high performance. These two methods consume a massive amount of energy. Hence, we aimed at developing a mechanism to reduce the energy consumption where we determine the safe utilization levels for the disks to operate with minimum probability failure rates while conserving energy.

After proposing a novel concept of safe utilization zone where energy of the disk can be conserved without degrading reliability, we designed and implemented a utilization control mechanism to improve both reliability and energy efficiency of disk systems. The utilization control mechanism ensures that disk drives are operated in safe utilization zones to minimize the probability of disk failure. We integrate the energy consumption technique that operates the disks at different power modes with our proposed reliability approach. Experimental results show that our approach can significantly improve reliable while achieving high energy efficiency for disk systems.

Acknowledgements

The work reported in this paper was supported by the US National Science Foundation under Grants No. CCF-0742187, No. CNS-0757778, No. CNS-0831502, No. OCI-0753305, and No. DUE-0621307, and Auburn University under a startup grant.

References

[1] B. Schroeder, and G.A. Gibson, “Disk failures in the real world: what does an MTTF of 1,000,000 hours mean to you?” *Proc.5th Conf. on USENIX Conf. on File and Storage Technologies*, vol.5, San Jose, CA, Feb 2007.

[2] D. Li, and J. Wang, "Conserving Energy in RAID Systems with Conventional Disks," *Proc. 3rd Int'l Workshop on Storage Network Architecture and Parallel I/Os*, 2005.

[3] D. Zhu, R. Melhem, and D. Mosse, "The effects of energy management on reliability in real-time embedded systems", *Proc. IEEE/ACM Int'l conf. Computer-aided design*, pp. 35-40, 2004.

[4] D.P. Helmbold, D.D. Long, and B. Sherrod, "A dynamic disk spin-down technique for mobile computing," *Proc. 2nd annual Int'l Conf. on Mobile computing and networking*, ACM New York , USA , 1996, pp. 130-142.

[5] E. Pinheiro, W. Weber, and L. A. Barroso, “Failure trends in a large disk drive population,” *Proc. 5th Conf. on USENIX Conf. on File and Storage Technologies*, vol.5, San Jose, CA, Feb 2007.

[6] E. Rosti, G. Serazzi, E. Smirni and M.S. Squillante, "Models of Parallel Applications with Large Computation and I/O Requirements," *IEEE Trans. Software Engineering* ,vol. 28 , no.3, pp. 286-307.

[7] E. V. Carrera, E. Pinheiro, and R. Bianchini, "Conserving Disk Energy in Network Servers," *Proc. 17th Int'l Conf. on Supercomputing*, ACM Press, pp. 86-97.

[8] J. G. Elerath and S. Shah, “Server class disk drives: how reliable are they,” *IEEE Reliability and Maintainability Symp.*, pp. 151-156, Jan 2004.

[9] J. Mao, C. G. Cassandras, Q. Zhao, “Optimal Dynamic Voltage Scaling in Energy-Limited Nonpreemptive Systems with Real-Time Constraints,” *IEEE Trans. Mobile Computing*, vol.6,Jun 2007,pp. 678-688.

[10] J. Yue, Y. Zhu, and Z. Cai, "Evaluating Memory Energy Efficiency in Parallel I/O Workloads", *Proc. IEEE International Conf. on Cluster Computing*, Austin, Texas, 2007, pp. 21-30

[11] J. Zedlewski, S. Sobti, N. Garg, F. Zheng, A. Krishnamurthy, and R. Wang, “Modeling Hard-Disk Power Consumption,” *Proc. 2nd USENIX Conf on File and Storage Technologies*, San Francisco, CA, March 2003.

[12] L. Lu, P. Varman, and J. Wang, "DiskGroup: Energy Efficient Disk Layout for RAID1 Systems," *Int'l Conf. on Networking, Architecture, and Storage*, IEEE Press, pp. 233-242.

[13] L. Yuan, and G. Qu, "Analysis of energy reduction on dynamic voltage scaling-enabled systems," *IEEE Tran. Computer-Aided Design of Integrated Circuits and Systems* , vol. 24, no. 12, pp. 1827-1837.

[14] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, “DRPM: dynamic speed control for power management in server class disks,” *Proc. 30th Int’l Symp. Computer Architecture*, pp. 169-18, May 2003.

[15] S. Son, G. Chen, M. Kandemir, and A. Choudhary," Exposing disk layout to compiler for reducing energy consumption of parallel disk based systems," *Proc. ACM SI Symp. Principles and practice of parallel programming*, pp. 174-185.

[16] S. W. Son, G. Chen, O. Ozturk, M. Kandemir, and A. Choudhary, "Compiler-Directed Energy Optimization for Parallel Disk Based Systems," *IEEE Trans. Parallel and Distributed Systems, vol. 18, no.*9, 2007, pp. 1241-1257.

[17] T. Simunic, L. Benini, and G. De Micheli, "Dynamic Power Management of Laptop Hard Disk," *Proc. Design Automation and Test*, Europe, 2000.

[18] Y. Du, J. Dong, and M. Cai, "Dynamic Voltage Scaling of Flash Memory Storage Systems for Low-Power Real-Time Embedded Systems," *Proc. Second Int'l Conf.Embedded software and systems*, 2005, pp. 152-157.