This paper appeared in the Proceedings of the 12th IEEE International Symposium on Distributed Simulation and Real Time Applications (DS-RT'08), Vancouver, British Columbia, Canada, Oct. 2008.

An Adaptive Energy-Conserving Strategy for Parallel Disk Systems

Mais Nijim
School of Computing
University of Southern Mississippi
Hattiesburg, MS 39406
mais.nijim@usm.edu
http://orca.st.usm.edu/~mais

Adam Manzanares, Xiao Qin[†]
Department of Computer Science and
Software Engineering
Auburn University, Auburn, AL 36849
{acm0008,xqin}@auburn.edu
http://www.eng.auburn.edu/~xqin

Abstract

In the past decade parallel disk systems have been highly scalable and able to alleviate the problem of disk I/O bottleneck, thereby being widely used to support a wide range of data- intensive applications. Optimizing energy consumption in parallel disk systems has strong impacts on the cost of backup power-generation and cooling equipment, because a significant fraction of the operation cost of data centres is due to energy consumption and cooling. Although a variety of parallel disk systems were developed to achieve high performance and energy efficiency, most existing parallel disk systems lack an adaptive way to conserve energy in dynamically changing workload conditions. To solve this problem, we develop an adaptive energy-conserving algorithm, or DCAPS, for parallel disk systems using the dynamic voltage scaling technique that dynamically choose the most appropriate voltage supplies for parallel disks while guaranteeing specified performance (i.e., desired response times) for disk requests. We conduct extensive experiments to quantitatively evaluate the performance of the proposed energy-conserving strategy. Experimental results consistently show that DCAPS significantly reduces energy consumption of parallel disk systems in a dynamic environment over the same disk systems without using the DCAPS strategy.

1. Introduction

In the last decade, parallel disk systems have been widely used to support data-intensive applications, including but not limited to video surveillance [1], remote-sensing database systems, and digital libraries [5], The performance of data-intensive applications deeply relies on the performance of underlying disk systems due to the rapidly widening gap between CPU and disk I/O speeds [7]. Parallel disk systems play an important role in achieving high-performance for data-intensive applications, because the high parallelism and scalability of parallel disk systems can alleviate the disk I/O bottleneck problem.

A growing number of data centers introduce a momentous problem – a substantial amount of energy is consumed by hardware resources in data centers. For example, the power consumption of today's data center ranges from 75 W/ft^2 to 150-200 W/ft^2 . Since this trend will undoubtedly continue in the near future [8], the energy-consumption problem in data centers will become even more serious. Growing evidence shows that among various hardware resources in a data center, storage systems (e.g., parallel disk systems) are one of the biggest consumers of energy. A recent industry report reveals that storage devices account for almost 27% of the total energy consumed by a data center[14]. This problem is exacerbated by the availability of faster disks with higher power needs. Therefore, it desirable to design energy-efficient parallel disk systems by extensively investigate energy-conservation software techniques.

Modern data-intensive applications are likely to dynamically change their disk I/O patterns and performance requirements. As such, it is imperative for next-generation parallel disk systems to flexibly and adaptively reduce energy consumption during the course of the execution of a data-intensive application. Adaptively conserving energy in parallel disk systems becomes particularly critical for data-intensive applications in which disk requests need to be completed within specified response times or desired response times. Hence, energy-efficient parallel disk systems will have to aim at achieving two major goals: low energy dissipation and high guarantee of specified performance.

Disk scheduling algorithms play an important role in reducing the performance gap between processors and disk I/O [9]. The shortest seek time first (*SSTF*) algorithm is efficient in minimizing seek times; SSTF is starvation-bound and unfair in nature[10]. The SCAN scheduling algorithm can solve the unfairness problem while optimizing seek times [10]. Reist and Daniel proposed a parameterized generalization of the SSTF and SCAN algorithms [11]. Most existing disk scheduling algorithms are inadequate for adaptive energy conservation in parallel disk systems. To remedy this problem, in this study we develop an adaptive energy-conservation scheme or DCAPS using the dynamic voltage scaling (DVS) technique for parallel disks systems. More importantly, our scheme can provides significant energy savings while guaranteeing desired response times of disk requests by seamlessly integrating the DVS technique with disk scheduling mechanisms.

Disk I/O parallelisms can be provided in forms of both inter-request and intra-request parallelisms. The inter-request parallelism allows multiple independent requests to be served simultaneously by an array of parallel disks, whereas the intra-request parallelism enables a single disk request to be processed by multiple disks in parallel. A parallelism degree of a data request is the number of disks where the requested data resides. The DCAPS strategy developed in this research are capable of dealing with both types of parallelisms.

The rest of the paper is organized as follows. We summarize related work in the next section. Section 3 describes a system architecture for energy-efficient parallel disk systems. In Section 4, we propose the adaptive energy-conservation scheme. Section 5 evaluates the performance of the proposed energy-saving technique by comparing an existing approach. Section 6 concludes the paper with summary and future directions.

2. Related Work

Disk I/O has become a performance bottleneck for data-intensive applications due to the widening gap between processor speeds and disk access speeds [13]. To help alleviate the problem of disk I/O bottleneck, a large body of work has been done on parallel disk systems. For example, Kallahalla and Varman designed an on-line buffer management and scheduling algorithm to improve performance of parallel disks[14]. Scheuermann *et al.* addressed the problem of making use of striping and load balancing to tune performance of parallel disk systems. Rajasekaran and Jin developed a practical model for parallel disk systems [15]. Kotz and Ellis proposed investigated several write back policies used in a parallel file system implementation [16]. Our research is different from the previous studies in that we focused on energy savings for parallel disk systems. Additionally, our strategy is orthogonal to the existing techniques in the sense that our scheme can be readily integrated into existing parallel disk systems to substantially improve energy efficiency and performance of the systems.

Most of the previous research regarding conserving energy focuses on single storage system such as laptop and mobile devices to extend the battery life. Recently, several techniques proposed to conserve energy in storage systems include dynamic power management schemes [9], power aware cache management strategies [17], power aware perfecting schemes [18], software-directed power management techniques [19], redundancy techniques [19], and multi-speed settings[20]. However, the research on energy-efficient parallel disk systems is still in its infancy. It is imperative to develop new energy conservation techniques that can provide significant energy savings for parallel disk systems while maintaining high performance.

The dynamic voltage scaling technique or DVS is a widely adopted approach to conserving energy in processors. The DVS technique can dynamically reduce the voltage supplies of processors to conserve energy consumption in processors (see, for example,[21]). Thus, processor voltage supplies are scaled down to the most appropriate levels, thereby quadratically reducing power whenever possible. Compared with traditional systems with fixed voltage supply, systems with DVS can achieve high energy efficiency. Our approach differs from the conventional DVS methods, because ours is the first technique of its kind designed exclusively for energy-efficient parallel disk systems aiming to guarantee specified performance of data-intensive applications. Our adaptive energy-conserving strategy makes use of the DVS technique to achieve extremely low energy consumption in parallel disk systems while guaranteeing desired response times of disk requests.

3. System Architecture and Model

3.1 System Architecture

First of all, let us describe a framework within which we can develop an adaptive energy-conservation technique for parallel disk systems. The framework for energy-efficient parallel disk system is delineated in Fig.1. It is worth noting that the framework designed in this study is general enough to accommodate a wide range of storage systems, including both network attached storage devices (NAS) and storage area networks (SAN).

The framework embraces a parallel disk system, networks, an adaptive energy-conserving mechanism, a response time estimator, and a data partitioning mechanism. The energy-conserving mechanism, which is at the heart of the proposed system framework, is responsible for adaptively saving energy consumption in parallel disks without significantly degrading performance of the parallel disk system. Thus, the energy-conserving mechanism aims to achieve the best tradeoff between energy efficiency and performance.

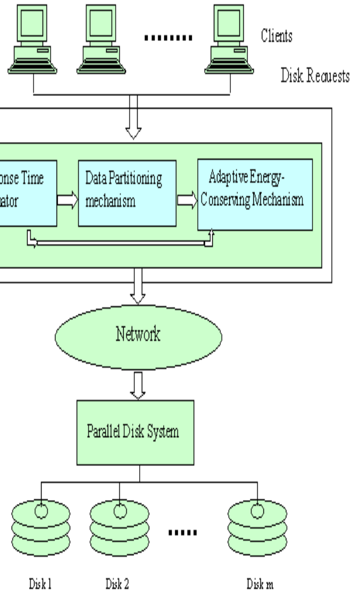


Fig. 1 The framework for adaptive energy saving techniques.

More specifically, the adaptive energy-conserving mechanism reduce energy consumption by making use of the dynamic voltage scaling technique to judiciously lower voltage supply levels of disks as long as specified performance requirements can be met. Section 4.3 gives much greater detail on the design and development of the adaptive energy-conserving mechanism. The data partitioning mechanism is geared to divide a large amount of data into fixed-size of data units stored on a number of disks. In this study we consider file striping, which is a generic method for a vast variety of data types. To determine an optimal parallelism degree (also known as stripe unit size) for each disk request, the data partitioning mechanism has to leverage the response time estimator to predict the response time of the request. The process of data partitioning is described in sufficient detail in Section 4.1. Moreover, the response time estimator is indispensable for the adaptive energy-conserving mechanism in the sense that estimating response times make it possible to save energy by dynamically adjust voltage supply levels without violating timing requirements (i.e., desired response times) Section 4.2 outlines a means of estimating response times of disk requests submitted to a parallel disk system.

3.2 Energy Consumption Model

Before developing the adaptive energy-conserving mechanism, we first introduce a power consumption model for parallel disk systems. We consider a sequence of disk requests $R = \{r_1, r_2, \cdots, r_n\}$ submitted to a parallel disk system. Each disk request $r_i \in R$ has an arrival time a_i, a desired response time t_i, and data size d_i. Ideally, request r_i needs to be completed within the desired response time t_i.

A multiple-voltage disk system has a number of discrete voltages; the disk system can instantaneously switch from one voltage to another. Without loss of generality, we assume the parallel disk system can be operated at a finite set $V = \{v_1, v_2, \cdots, v_{\max}\}$ of voltage supply levels. Given a disk voltage v_i, we can accordingly determine the bandwidth b_i of the disk.

Because energy dissipation in disks quadratically proportional to supply voltages, voltage scaling can achieve significant energy savings for disks. Thus, the energy consumption rate P_i of the ith disk can be expressed as below:

$$P_i = C_1 \cdot v_{i,dd}{}^2 \cdot \frac{\left(v_{i,dd} - v_t\right)^\alpha}{C_2}, \quad v_{i,dd} \in V, v_{i,dd} \geq v_t;$$

(1)

where C_1, C_2, and $\alpha \in [1,2]$ are constants depending on physical characteristics of disk devices, $v_{i,dd}$ is the supply voltage, and v_t is the threshold voltage. Let D_i denote a set of disk requests to be processed by the ith disk in the parallel disk system. Given a disk request r_j to be processed by the ith disk, we can calculate the energy consumption of the request as below:

$$E_{i,j} = P_i\left(v_{i,dd}^j\right) \cdot \theta_j\left(v_{i,dd}^j\right) \qquad (2)$$

where $v_{i,dd}^j$ is the voltage supply level determined for the disk request, $P_i\left(v_{i,dd}^j\right)$ is the disk's energy consumption rate, and $\theta_j\left(v_{i,dd}^j\right)$ is the processing time of the disk request. Both $P_i\left(v_{i,dd}^j\right)$ and $\theta_j\left(v_{i,dd}^j\right)$

largely rely on the supply voltage $v_{i,dd}^j$ of the disk; $P_i\left(v_{i,dd}^j\right)$ can be straightforwardly derived from Eq. (1).

The energy consumption E_i of disk i is written as a summation of energy consumption caused by each disk request handled by the disk. Thus, we have

$$E_i = \sum_{r_j \in D_i} E_{i,j} = \sum_{r_j \in D} P_i\left(v_{i,dd}^j\right)\cdot \theta_j\left(v_{i,dd}^j\right) \qquad (3)$$

Suppose there are m disks in the parallel disk system, the total energy consumption E of the disk system can be expressed as:

$$E = \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \sum_{r_j \in D_i} E_{i,j} = \sum_{i=1}^{m} \sum_{r_j \in D} P_i\left(v_{i,dd}^j\right)\cdot \theta_j\left(v_{i,dd}^j\right)$$

(4)

We can now obtain the following non-linear optimization problem formulation to compute the energy consumption of a parallel disk system

Minimize $E = \sum_{i=1}^{m} \sum_{r_j \in D} P_i\left(v_{i,dd}^j\right)\cdot \theta_j\left(v_{i,dd}^j\right)$

Subject to (a) $v_{i,dd}^j \in \{v_1, v_2, \cdots, v_{\max}\}$

(b) $f_j \le t_j$ (5)

where f_j is the response time of the jth disk request. $f_j \le t_j$ in Expression (5) signifies that the desired response time constraints must be met.

4. Adaptive Energy-Conserving Strategy

The proposed adaptive energy-conserving strategy encompasses three components, namely, a data partitioning technique, response time estimation method, and an adaptive DVS algorithm. In this section, we describe the design of these three components in more detail.

4.1 Data Partitioning

One of the major components in the proposed framework (see Section 3.1) is the method of data partitioning that determines the optimal parallelism degrees for disk requests. Dynamic data partitioning is of importance for our adaptive energy-conserving strategy, because the data partitioning method helps in minimizing the response times of requests, thereby creating more space to reduce energy consumption by scaling down disk supply voltages. As such, in the first place our strategy aims to shorten the response times by adaptively determining the optimal parallelism degree of each request (see Step 3 in Fig. 2).

We denote the parallelism degree and data size of a request r_i by p_i and d_i, respectively. Before proceeding to the analysis of optimal parallelism degrees, let's first formally derive the disk service time $T_{disk}(d_i, p_i)$ of request r_i. Thus, the disk service time can be computed as

$$T_{disk}(d_i, p_i) = T_{seek}(p_i) + T_{rot}(p_i) + T_{trans}(d_i, p_i),$$

(6)

where $T_{seek}(p_i), T_{rot}(p_i), \text{and}\, T_{trans}(d_i, p_i)$ are the seek time, rotation time, and transfer time of the disk request The seek time can be approximated as below, where C is the number of cylinders on a disk, a and b are two disk-type-independent constants, whereas e and f are disk-type-dependent constants.

$$T_{seek}(p_i) = eC\left(1 - a - b\ln(p_i)\right) + f \qquad (7)$$

The value of rotation time can be expressed as Eq. (8), where T_{ROT} is the rotation time of a disk.

$$T_{rot}(p_i) = \frac{p_i}{p_i + 1}\cdot T_{ROT} \qquad (8)$$

The transfer time can be approximated by Eq. (9), where B_{disk} is the disk bandwidth.

$$T_{trans}(d_i, p_i) = \frac{d_i}{p_i}\cdot\frac{1}{B_{disk}} \qquad (9)$$

Substituting Eqs. (7)-(9) into Eq. (6), we obtain the value of disk service time as

$$T_{disk}(d_i, p_i) = eC\left(1 - a - b\ln(p_i)\right) + f + \frac{p_i}{p_i + 1}\cdot T_{ROT} + \frac{d_i}{p_i}\cdot\frac{1}{B_{disk}}.$$

(10)

Now we are positioned to calculate the optimal parallelism degree of request r_i by determining the minimum of the function $T_{disk}(d_i, p_i)$. Thus, we can obtain the optimal value of p_i by solving Eq. (11).

$$\frac{dT_{disk}(d_i, p_i)}{d(p_i)} = \frac{T_{ROT}}{p_i + 1} - \frac{p_i \cdot T_{ROT}}{(p_i + 1)^2} - \frac{eCb}{p_i} - \frac{d_i}{p_i^2}\cdot\frac{1}{B_{disk}} = 0.$$

(11)

The parallelism degree determined by Eq. (11) can not exceed m, which is the number of disks in the system. Therefore, the optimal parallelism degree is given by $\min(p_i, m)$.

4.2 Response Time Estimator

To adaptively adjust the voltage supply of disk requests, we need to estimate each request's maximum response time, which is defined as an interval between the time a request submitted and the time the parallel disk system completes corresponding disk I/O operations. Given a newly issued request *r*, the response time of *r* is estimated by Eq. (12).

$$T(r, p, \sigma) = T_{queue} + T_{partition} + \max_{i=1}^{p}\left\{T^{i}_{proc}(r, p, \sigma_i)\right\} \quad (12)$$

where *p* is the parallelism degree determined by the data partitioning mechanism, $v = (v_1, v_2, \cdots, v_p)$ is the request's vector of the supply voltage for *p* stripe units, T_{queue} is the queueing delay at the client side, $T_{partition}$ is the time spent in data partitioning, and T^{i}_{proc} is the system processing delay experienced by the *i*th stripe unit of the request. With respect to the *i*th stripe unit of the request, the system processing delay T^{i}_{proc} can be expressed as

$$T^{i}_{proc}(r, p, v_i) = T^{i}_{network}(r, p, v_i) + T^{i}_{disk}(r, p, v_i) \quad (13)$$

where $T^{i}_{network}$, and T^{i}_{disk} are the delays at the network subsystem, and parallel disk subsystems, respectively.

We assume that when the *i*th stripe unit of a request arrives at the network queue, there are *k* stripe units waiting to be delivered to the parallel disk sub-system. Suppose stripe units are transmitted in a first-in-first-out order, all the stripe units that are already in the queue prior to the arrival of the *i*th stripe unit must be transmitted earlier than the *i*th stripe unit. Hence, the delay in the network subsystem $T^{i}_{network}(r, p, v_i)$ can be written as

$$T^{i}_{network}(r, p, v_i) = \frac{i \cdot \frac{d}{p} + \sum_{j=1}^{k} d_j}{B_{network}} \quad (14)$$

where d_j is the data size of the *j*th stripe unit in the network queue, and $B_{network}$ is the effective network bandwidth. It is worth noting that *k* in Eq. (14) is the optimal parallelism degree determined by the data partitioning mechanism (see Eq. 11 in Section 4.1).

Similarly, it is assumed that when the *i*th stripe unit of the request arrives at disk *j*, there are *k* disk requests must be processed by disk *j* before handling the stripe unit. Thus, the delay in the disk subsystem $T^{i}_{disk}(r, p, v_i)$ is given by the following formula

$$T^{i}_{disk}(r, p, v_i) = T_{disk,j}(d/p) + \sum_{l=1}^{k} T_{disk,j}(d_l) \quad (15)$$

where $T_{disk,j}(d)$ is the disk processing time of a request containing *d* bytes of data. We can quantify $T_{disk,j}(d)$ as follows

$$T_{disk,j}(d) = T_{seek} + T_{rot} + \frac{d}{B_{disk}}, \quad (16)$$

where T_{seek} and T_{rot} are the seek time and rotational latency, and $\frac{d}{B_{disk}}$ is the data transfer time depending on the data size *d* and disk bandwidth B_{disk}.

4.3 The Adaptive Energy-Conservation Algorithm

Now we are positioned to design the adaptive energy-conservation algorithm or DCAPS for parallel disk systems. The DCAPS algorithm aims at judiciously lower the parallel disk system voltage using dynamic voltage scaling technique or DVS, thereby reducing the energy consumption experienced by disk requests running on parallel disk systems. The processing algorithm separately repeats the process of controlling the energy by specifying the most appropriate voltage for each disk request. Thus, the algorithm is geared to adaptively choose the most appropriate voltage for stripe units of a disk request while warranting the desired response time of the request.

Specifically, the algorithm is carried out in three phases: dynamic data partitioning (see Section 4.1), response time estimation (see Section 4.2), and adaptive energy consumption controller. To diminish the energy consumption of the disk systems, DCAPS endeavors to minimize the supply voltage of a request. Hence, the first phase dynamically calculates the optimal parallelism degree of the request, thereby reducing delays at the parallel disk subsystems (see Eq. 18). During the second phase of the algorithm, the response time of each stripe unit is estimated (see Eqs. 15 and 16). Phase three, guided by the estimated response time obtained and desired response time, adaptively reduce the supply voltage for each stripe unit provided that the request's response time does not exceed the request's desired response time. The complete algorithm of DCAPS is outlined in Fig. 2.

When a disk request is issued to the system, the DCAPS strategy inserts the newly arrived requests into the waiting queue based on the earliest desired response time first policy (see Step 1). After the data portioning of each request in the queue, DCAPS initializes the voltage of all the stripe units of request r_i to the maximum supply voltage v_{max} (see Step 6).

In doing so, DCAPS are more likely to guarantee desired response times under heavily loaded conditions. Using the dynamic voltage scaling technique, the DCAPS strategy adaptively makes disks operate at low voltage supply levels for all the stripe units to conserve the total energy consumption of the parallel disk system. Assume that the maximum supply voltage v_{max}. is 3.3 Volts, the supply voltage can be reduced as long as the disk request can be accomplished within its desired response time or the supply voltage reach the minimum voltage v_{min}. In this study, we assume that the threshold voltage is 0.8 [online]. In an effort to steadily reduce the voltage of stripe units, DCAPS guarantees that all requests will be completed before their desired response times. Thus, the following property needs to be satisfied in DCAPS.

1. The reduced supply voltage v_{ij} is greater than the minimum voltage v_{min};
2. $T_j(r_i, p_i, \sigma_i) \le t_i$, where T_j is the response time of the *j*th stipe unit, t_i is the desired response time of the request, and

$$T_j(r_i, p_i, \sigma_i) = T_{queue} + T_{partition} + T_{proc}^{ij}(r_i, p_i, \sigma_{ij})$$

3. The reduced supply voltage v_{ij} is greater than the minimum voltage v_{min};
4. $T_j(r_i, p_i, \sigma_i) \le t_i$, where T_j is the response time of the *j*th stipe unit, t_i is the desired response time of the request, and

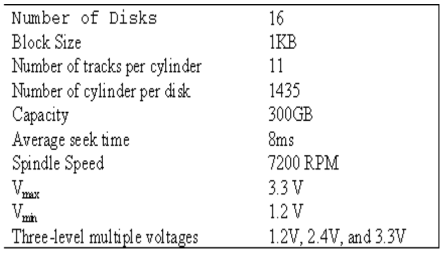
$$T_j(r_i, p_i, \sigma_i) = T_{queue} + T_{partition} + T_{proc}^{ij}(r_i, p_i, \sigma_{ij})$$

Steps 10-11 are repeatedly performed to scale down disk voltage until a request's desired response time cannot be guaranteed (see Step 12) or the supply voltage are approaching the threshold voltage. Consequently, DCAPS adaptively reduce the supply voltage while making the best effort to complete all the disk requests before their desired response time.

5. Experimental Results

To evaluate the performance of the DCAPS strategy in an efficient way, we simulated a parallel disk system with all the functions that are necessary to implement our system. Table 1 summarizes important parameters used to resemble real world disks. In addition, we implemented a data-partitioning algorithm to optimize parallelism degrees of large disk I/O requests. We will first compare the performance of a parallel disk system with DCAPS with that of another system without employing DCAPS. We will then study effects of varying arrival rates, data size, and disk bandwidth on the performance of the two disk systems. Next, we will compare and evaluate the two disk systems based on varying the voltage. Finally, we will also analyse the performance impacts of parallelism degrees on the parallel disk systems.

Table 1. Disk parameters of the simulated parallel disk system



Number of Disks	16
Block Size	1KB
Number of tracks per cylinder	11
Number of cylinder per disk	1435
Capacity	300GB
Average seek time	8ms
Spindle Speed	7200 RPM
V_{max}	3.3 V
V_{min}	1.2 V
Three-level multiple voltages	1.2V, 2.4V, and 3.3V

In our simulation experiments, we made use of the following three performance metrics to demonstrate the effectiveness of the DCAPS scheme. (1) Satisfied ratio is a fraction of total arrived disk requests that are found to be finished within their desired response times. (2) Energy consumption is the total energy consumed by the parallel disk systems. And (3) Energy conservation ratio

5.1 Impact of Arrival Rate

This experiment is focused on comparing a parallel disk system with the DCAPS strategy against a standard parallel disk system with a fixed voltage supply level. We study the impacts of disk request arrival rate on the satisfied ratio and normalized energy consumption. To achieve this goal, we increased the arrival rate of disk requests from 0.1 to 0.5 No./Sec with an increment of 0.1 No./Sec.

Figs. 3 and 4 plot the satisfied ratios, normalized energy consumption, and energy conservation ratio of the parallel disk systems with and without DCAPS. Figs 3(a) reveals that the DCAPS scheme yields satisfied ratios that are very close to those of the parallel disk system without employing DCAPS. This is essentially because DCAPS endeavors to save energy consumption at the marginal cost of satisfied ratio. More importantly, Figs. 3(b) and 4 show that DCAPS significantly reduces the energy dissipation in the parallel disk system by up to 71% with an average of 52.6%. The improvement in energy efficiency can be attributed to the fact that DCAPS reduces the disk supply voltages in the parallel disk system while making the best effort to guarantee desired response times of the disk requests. Furthermore, it is observed that as the disk request arrival rate increases, the energy consumption of the both parallel disk systems soars. Fig. 4 shows that as the load increases, the energy conservation ratio tends to decrease.

This result is not surprising because high arrival rates lead to heavily utilized disks, forcing the DCAPS to boos disk voltages to process larger number of requests within their corresponding desired response times. Increasing number of disk request and scaled-up voltages in turn give rise to the increased energy dissipations in the parallel disk systems.

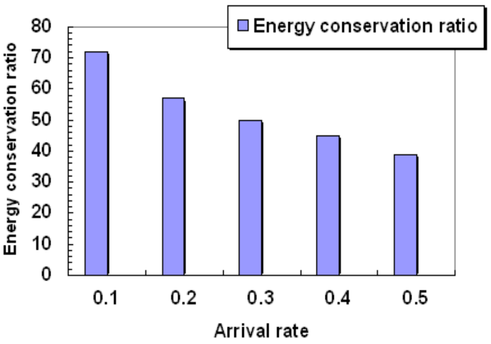


Fig. 4. Impact of request arrival rate on energy conservation ratio.

6. Conclusions and Future Work

Parallel disk systems play an important role in achieving high-performance for data-intensive applications, because the high parallelism and scalability of parallel disk systems can alleviate the disk I/O bottleneck problem. However, growing evidence show that a substantial amount of energy is consumed by parallel disk systems in data centers. It is therefore highly desirable to design energy-efficient parallel disk systems by extensively investigate energy-conservation software techniques. Adaptively conserving energy in parallel disk systems becomes particularly critical for data-intensive applications in which disk requests need to be completed within specified response times or desired response times. In this paper, we focused on the design of novel parallel disk systems that can achieve both great energy efficiency and high guarantee of specified performance. Specifically, we developed an adaptive energy-conserving strategy, which dynamically scaled down disk voltage supplies to the most appropriate levels, thereby significantly reduce energy dissipation in parallel disk systems. The experimental results have confirmed that our scheme can achieve up to 70% energy savings compared with standard parallel disk systems with fixed supply voltage.

Our approach is the first technique of its kind designed exclusively for energy-efficient parallel disk systems aiming to guarantee specified performance of data-intensive applications. As a future direction, we will propose a dynamic voltage scaling technique at the level of data-intensive applications. Further, we plan to extend our approach by considering overhead of scaling disk supply voltages.

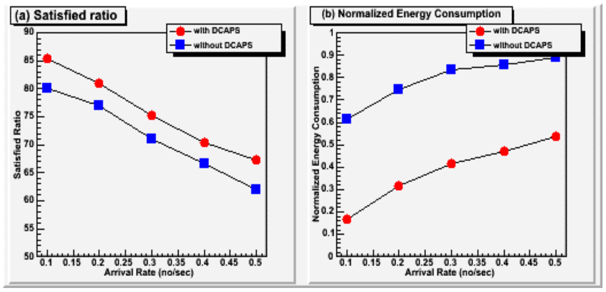


Fig. 3. Impact of request arrival rate on satisfied ratio and normalized energy consumption when disk bandwidth is 30MB/sec.

Acknowledgments:

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

References

[1] Avitzour, "Novel scene calibration procedure for video surveillance systems," *IEEE Trans. Aerospace and Electronic Systems*, Vol. 40, No. 3, pp. 1105-1110, July 2004.

[2] Ali Manzak and Chaitali Chakrabarti, 'variable Voltage Task Scheduling Algorithms for Minimization Energy/Power", IEEE Trans very Large Scale Integration Sys, vol.11, no. 2, April

2003.

[3] C. Chang, B. Moon, A. Acharya, C. Shock, A.Sussman, and J. Saltz. "Titan: a High-Performance Remote-Sensing Database," *Proc. 13th Int'l Conf. Data Eng.*, Apr 1997.

[4] J. Coffman and M. Hofri, "Queueing Models of Secondary Storage Devices," *Stochastic Analysis of Computer and Comm. Sys.*, Ed. Hideaki Takagi, North-Holland, 1990.

[5] T. Sumner and M. Marlino, "Digital libraries and educational practice: a case for new models," *Proc. ACM/IEEE Conf. Digital Libraries*, pp. 170 – 178, june 2004

[6] J. Coffman and M. Hofri, "Queueing Models of Secondary Storage Devices," *Stochastic Analysis of Computer and Comm. Sys.*, Ed. Hideaki Takagi, North- Holland, 1990.

[7] X. Qin, "Performance Comparisons of Load Balancing Algorithms for I/O-Intensive Workloads on Clusters,"*Journal of Network and Computer Applications*, 2007.

[8] B. Moore, Taking the data center power and cooling challenge. Energy User News, August 27th, 2002.

[9] F. Douglis, P.Krishnan, and B. Marsh, "Thwarting the Power-Hunger Disk," *Proc. Winter USENIX Conf.*, pp.292-306, 1994.

[10] P. J. Denning, "Effects of Schuling on File Memory Operations," *Proc. AFIPS Conf.*, pp.9-21, April 1967.

[11] .R. Reist and S. Daniel, "A Continuum of Disk Scheduling Algorithms," *ACM Trans. on Computer Sys.*, pp.77-92, Feb. 1987

[12] Power, heat, and sledgehammer. White paper Maximum institution Inc., http://www.max-t.com/downloads/whitepapers/SledgehammerPowerheat20411.pdf, 2002.

[13] S. Rajasekaran, "Selection algorithms for parallel disk systems," *Proc. Int'l Conf. High Performance Computing*, pp.343-350, Dec. 1998.

[14] M. Kallahalla and P. J. Varman, "Improving parallel-disk buffer management using randomized writeback,"*Proc. Int'l Conf. Parallel Processing*, pp. 270-277, Aug. 1998.

[15] S. Rajasekaran and X. Jin, "A practical realization of parallel disks Parallel Processing," *Proc. Int'l Workshop Parallel Processing*, pp. 337-344, Aug. 2000.

[16] D. Kotz and C. Ellis, "Cashing and writeback policies in parallel file systems," *Proc. IEEE Symp. Parallel and Distributed Processing*, pp. 60-67, Dec. 1991.

[17] Q. Zhu, F.M David, C.F. Devaaraj, Z. Li, Y.Zhou, and P. Cao, Reducing Energy Consumption Of Disk Storage Using Power Aware Cache Management," *Proc. High Performance Computer Framework*, 2004.

[18] S.W. Son and M. Kandemir, " Energy Aware data perfecting for multi-speed disks,"Proc. ACM International Conference on Computing Frontiers, Ischia, Italy, May 2006.

[19] S.W. Son, M. Kandemir, and A. Choudhary, "Software-directed disk power management for scientific applications," *Proc. Int'l Symp. Parallel and Distr. Processing*, April 2005.

[20] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Fanke, "DRPM: Dynamic Speed Control for Power Management in Server Class Disks," *Proc. Int'l Symp. of Computer Architecture*, pp. 169-179, June 2003.

[21] T. Kuroda, et al., " Variable supply voltage scheme for low power high speed CMOS digital design," *IEEE J. Solid-State*, vol. 33, pp.454-462, Mar. 1998.

[22] M. Nijim, X. Qin, and T. Xie, "Modeling and Improving Security of a Local Disk System for Write-Intensive Workloads," *ACM Transactions on Storage*, in press, 2007.

```
Input: $r$: a newly arrived disk request
       $t_i$: desired response time of the $i$th request
       $v_{max}$: the maximum supply voltage
       $v_{min}$: the minimum supply voltage
       $Q$, a waiting queue at the client side
Insert $r$ into $Q$ based on the earliest desired response
  time first policy
for each request $r_j$ in the waiting queue $Q$ do
  /* Phase 1: dynamic data partitioning */
  Calculate the optimal parallelism degree $p_i$ of $r_i$
  Partition the request into $p_i$ stripe units
  for each stripe unit of $r_i$ do
    Initialize $v_{ij}$ of the $j$th stripe unit to the maximum
    supply voltage $v_{max}$
    /* Phase 2: response time estimation */

    Apply Eqs. (15) and (16) to estimate the response
         time of the $j$th stripe unit;
    /* Phase 3: Dynamic Voltage Scaling */
    while (estimated response time < desired response time $t_i$) do
      if $v_{ij} > v_{min}$ then /* $v_{ij}$ can be further reduced *
     / (see property 1)
        scale the voltage $v_{ij}$ down to the next level;
        Apply Eqs. (15) and (16) to estimate the response
         time of the $j$th stripe unit;
      else break  /* $v_{ij}$ can not be further reduced */
    end while
     Deliver the $j$th unit through the network subsystem to
     the parallel disk system;
end for
```

Fig. 2. The adaptive energy-conserving strategy (DCAPS)