Applied Mathematics & Information Sciences
An International Journal

© 2013 NSP
Natural Sciences Publishing Cor.

Transition Probability: A Novel Modeling Approach of Energy Consumption for Storage Subsystem

Qiang Zou[1] *and Yong Li*[2]

[1] School of Computer and Information Science, Southwest University, Chongqing 400715 China
[2] School of Mathematics, Chongqing Normal University, Chongqing 400047 China

Received: 10 Jan. 2012; Revised 28 Aug. 2012; Accepted Oct. 6, 2012
Published online: 1 Jan. 2013

Abstract: In this paper, considering the transitions among different power states such as active, idle, and standby, we define the transition probability mathematically from active mode to idle mode, and propose a novel analytical approach to evaluate energy consumption and performance metrics, i.e., queue length, throughput, and response time. Simulation results indicate that our proposed model might motivate storage researchers to exploit a quick analysis toolkit and give some insights for the design of power-aware storage systems.

Keywords: power consumption, queue length, throughput.

1. Introduction

The disk-based storage subsystems have been considered as one of the major parts which consume energy in computer systems, especially in data centers, since it is proven that the energy consumed by them surpasses the one by the rest in computer systems [1,2]. Recent studies have demonstrated that hard disk drives "can" draw more energy than all the processors together [2]. Typically hard drives consume 15-30% of the total power consumed by the whole computer.

In order to efficiently utilize energy, new power management techniques are proposed to achieve the goals of low-power disk subsystems [3–9]. The main and ultimate goals of low-power disk subsystems and techniques are: 1) to switch as many disks from the active power state to the standby mode as possible; 2) and keep the disks in standby mode as long as possible. This motivates us to propose an analytical performance and energy model for hard disks. The ability to operate the hard disk in a low-power state, combining with intelligent management, will reduce the power consumption in disk subsystems without a significant performance degradation.

Based on the definition of multiple power states, a lot of efforts focus on power consumption management mechanisms [10–13], which can be developed to allow system software to control transitions between the power states. But, little work studies energy consumption and performance metrics of I/O systems through the transition probability among different disk modes. Given that multiple power states have been defined, this paper proposes a mathematical power model aimed at a specific disk subsystem, to model the transition process among those modes. To the best of our knowledge, most of existing power management model studies are for mechanisms developed to allow system software to control transitions between these states [7,8,15], and little research has been done on the definition of transition probability which can be further developed to efficiently analyze the transition process of power state, and characterize power consumption as well as I/O performance.

In this paper, our proposed model provides a series of explicit expression to calculate the performance metrics of I/O systems. This disk state model provides a novel theoretical approach to evaluate I/O performance and energy consumption, and might be exploited the useful insight to design an efficient power management system.

The rest of this paper is organized as follows. Section 2 gives an overview of the related works. Section 3 briefly describes the characteristics and operating modes of a representative hard disk. Section 4 develops analytical model

[*] Corresponding author: e-mail: qzoucs@gmail.com

to evaluate performance metrics. Section 5 presents the analytical results of our proposed model. At last, we conclude this paper in Section 6.

2. Related Work

Many researches on power-aware storage systems focus on saving power with new low-power storage medium [20], new power-efficient storage architectures [21], and clever data organizations [22,19]. However, only a few literatures concentrate on the performance-power modeling of power-aware disk systems [18]. Generally, they analyze the impact of aggressively spinning down disks since the service time of last I/O request exceeds some thresholds [23–25]. Ref. [26] develops an adaptive power management algorithm called ABLE (i.e., Adaptive Battery Life Extender), which has been successfully incorporated in IBM 2.5-inch Travelstar drives and IBM Micro drives. Douglis [23] used a disk simulator with an assumed-fixed average response time for all requests, except for some requests which lie within a neighborhood of the previous request.

Greenawalt et al [27] propose and effect an analytical model which assumes that the I/O request arrivals yield to a Poisson distribution. This simplification relies on two implicit assumptions. One is that a disk has only two distinct power levels: active and idle. And the other is that an active disk always consumes energy at a same consumption rate. In this paper, the two implicit assumptions above also be used to develop our model.

Single performance measurement is one of the three major problems with current server-side disk energy management policies [14,15]. The traditional performance metrics of I/O systems are throughput and response time. To the best of our knowledge, all the current disk energy management algorithms only consider one of the two traditional measures for the tradeoff [16,17]. In this paper, we consider throughput and queue length.

3. Hard Disk Model

In this paper, we refer to the disk characteristic parameters of IBM Ultrastar 36ZX disk model [29], and summarize the corresponding performance and power parameters in Table 1.

For IBM Ultrastar 36ZX disk model, one power-mode switch cycle is illustrated in Fig. 1. As shown in Fig. 1, disk keeps the active mode with a power consumption of 39W when it is accessed. If no I/O request is waiting for respondence as the last one has been completed, disk will switch to the idle mode automatically, and power consumption decreases to 22.3W.

If the duration of idle mode is less than 15 seconds as new I/O requests arrive, disk then seeks and switches to active mode again. If the duration of idle mode reaches 15 seconds and no I/O request arrives or waits, disk will spin down and switch to the standby mode. A spin-down operation lasts 15 seconds, with a power consumption of 12.72W.

Table 1 Disk Characteristic Parameters.

Parameter	Value
Disk Model	IBM Ultrastar 36ZX
Storage Capacity	33.6 GB
Rotation Speed	15,000 RPM
Power (active)	39 W
Power (idle)	22.3 W
Power (standby)	12.72 W
Power (seek)	39 w
Energy(spindown:idle→standby)	12.72 W
Time(spindown:idle→standby)	15 secs
Energy(spinup:standby→active)	34.8 w
Time(spinup:standby→active)	26 secs

In the standby mode, power consumption further decreases to 12.72W. However, once new I/O requests arrive, disk spins up again, and returns to the active mode. A spin-up operation continues 26 seconds, with a power consumption of 34.8W.

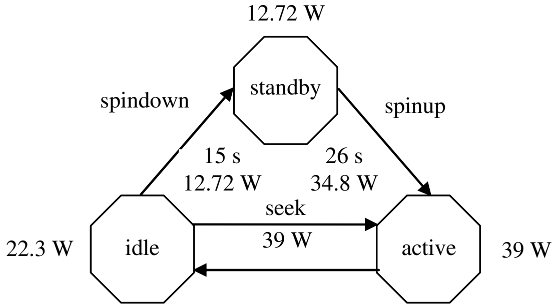


Figure 1 Power Consumption Mode.

In the following section, according to the disk characteristic parameters summarized in Table 1 and power consumption mode shown in Fig. 1, we will construct an analytical model based on Poisson assumption to evaluate energy consumption and I/O performance metrics.

4. Energy Consumption Modeling

Considering a simplified example of one single disk drive in this section, the conclusions can be easily extended to the multi-disk cases as well as the Redundant Array of Independent Disk (RAID).

In this paper, a workload is called heavy if I/O arrival rates are above $1/15$, and is called light if I/O arrival rates are below $1/15$. So, the power consumption mode in Figure 1 is composed of two power consumption state models as follows: heavy workload and light workload, respectively. Therefore, the probability of I/O workload being

Table 2 The symbols and notations

Symbol	Definition
P_l	The probability of I/O workload to be light
P_h	The probability of I/O workload to be heavy
P_d	The probability of I/O defection
G	Service load
p_w	The probability of I/O request agree to wait
β	The degree of I/O requests' patience
T	Throughput
N_{hc}	The average queue length per heavy-load cycle
N_{lc}	The average queue length per light-load cycle
N_c	The average queue length per cycle
$\bar{E}_l$	The energy consumption per light-load cycle
$\bar{E}_h$	The energy consumption per heavy-load cycle
$\bar{E}$	The average total energy consumption per cycle
$\bar{t}_c$	The average response time per cycle
t_i	Time in idle mode
$\bar{t}_i$	The average time in idle mode
t_s	Time in standby mode
$\bar{t}_s$	The average time in standby mode
$\bar{t}_{lc}$	The average response time per light-load cycle
$\bar{t}_{hc}$	The average response time per heavy-load cycle

heavy in an interval is $P_h = P(t_i < 15) = 1 - e^{-15\lambda}$, while the probability of I/O workload being light in an interval is $P_l = P(t_i \geq 15) = e^{-15\lambda}$. The disk characteristics shown in Table 1 will be used to calculate the power consumption and performance metrics. In order to effectively introduce the analytical model, some symbols are defined in Table 2.

4.1. Heavy-workload Modeling

As can be seen from Section 3, disk will spin down and switch to standby mode if idle time is over 15 seconds. Then disk will spin up again as new I/O requests arrive. However, both spin-down and spin-up operations will increase the energy consumption sharply. For heavy workload, disk I/O arrivals usually exhibit very bursty, which makes idle time is often less than 15 seconds. In the heavy-workload case, I/O requests are intensive enough so that idle time is never over fifteen seconds with the probability of $P = P(t_i < 15) = 1 - e^{-15\lambda}$.

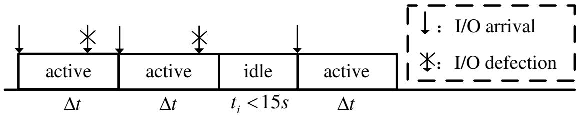


Figure 2 Disk states and transitions.

According to the queueing theory, we consider storage subsystem as a queueing system which consists of two components, i.e., disk and I/O request series. This paper tracks the disk modes shown in Fig. 1, and temporally plots the disk states and transitions in Fig. 2.

We call the interval between the starting times of two consecutive services as a cycle. As can be seen from Fig. 2, on the one hand, disk will continue to serve another I/O request after completing the last one if there is still I/O request existing in the queueing system. In this case, a cycle just consists of the service interval (Δt).

On the other hand, disk will switch to the idle mode after completing the last one if there is no I/O request existing in the queueing system. Therefore, a new I/O request arrives and can be served immediately if and only if the time in idle mode for disk, t_i, is below 15 unit times, i.e., $t_i < 15$. Then disk switches to the active mode again. In this case, a cycle will consist of two components: the service and idle intervals ($\Delta t + t_i$). The average idle interval due to the Poisson request arrival assumption is given as: $\bar{t}_i = 1/\lambda$.

As shown in Fig. 2, for intensive disk I/O workloads, I/O request with little patience arrives and will be canceled with a certain probability if disk is not available immediately. This is called "I/O defection". Our discussion focuses on the case that exists a certain degree of patience for I/O requests. On the one hand, some of I/O requests may wait for a while due to the fact that the disk is active and might not respond immediately. On the other hand, it is also impossible for each I/O request to wait for a long time. Once disk is not available, with the probability of I/O defection given as $P_d = \frac{\lambda \Delta t}{1+\lambda \Delta t}$, I/O request may be canceled immediately or after a waiting delay. And the corresponding disk throughput is $T = \lambda \Delta t(1 - P_d)$, where $\lambda \Delta t = G$ is service load. According to Ref. [28], the probability of which an I/O request is determined to wait for more than t time units is $p_w = p_{waite}(t, \bar{t}) = e^{-t/\bar{t}}$, where $\bar{t}$ is the average length of waiting time. The probability of that a random I/O request will wait for service is $\beta(1 - e^{-1/\beta})$.

If the number of I/O arrivals within an interval is n, then the probability of that k of them ($k \leq n$) will wait for service can be given as: $P_n(k) = C_n^k p_w^k (1 - p_w)^{n-k}$. For any of n, we note Ψ_k^* as the probability of k requests surviving at the end of the service interval, and Ψ_k^* can be represented as:

$$\begin{aligned}
\Psi_k^* &= \sum_{n=k}^{\infty} e^{-\lambda \Delta t} \frac{(\lambda \Delta t)^n}{n!} C_n^k p_w^k (1 - p_w)^{n-k} \\
&= e^{-\lambda \Delta t} \sum_{n=k}^{\infty} \frac{(\lambda \Delta t)^n}{n!} \frac{n!}{k!(n-k)!} p_w^k (1 - p_w)^{n-k} \\
&= e^{-\lambda \Delta t} \frac{(\lambda \Delta t)^k p_w^k}{k!} \sum_{n=k}^{\infty} \frac{(\lambda \Delta t)^{n-k} (1 - p_w)^{n-k}}{(n-k)!} \\
&= e^{-\lambda \Delta t} \frac{(\lambda \Delta t)^k p_w^k}{k!} \sum_{l=n-k=0}^{\infty} \frac{[(\lambda \Delta t(1 - p_w)]^l}{l!} \\
&= e^{-\lambda \Delta t} \frac{(\lambda \Delta t p_w)^k}{k!} e^{\lambda \Delta t(1-p_w)} \\
&= e^{-\lambda \Delta t p_w} \frac{(\lambda \Delta t p_w)^k}{k!}
\end{aligned}$$

© 2013 NSP
Natural Sciences Publishing Cor.

Note that $\sigma_{\Delta t} = \lambda \Delta t p_w = G\beta(1 - e^{-1/\beta})$, thus $\Psi_k^* = e^{-\sigma_{\Delta t}} \frac{(\sigma_{\Delta t})^k}{k!}$. Thus, the number of surviving I/O requests yields to a Poisson distribution with the parameter of $\sigma_{\Delta t}$. Disk will enter the idle mode with the transition probability of Ψ_0^*.

The average queue length of I/O requests existing in the queueing system per heavy-load cycle, N_c, can be represented as follows:

$$N_{hc} = \Omega_0^* + \sum_{i=1}^{\infty} i\Omega_i^* = e^{-\sigma_{\Delta t}} + \sigma_{\Delta t}$$
$$= e^{-G\beta(1-e^{-1/\beta})} + G\beta(1 - e^{-1/\beta}) \qquad (1)$$

The average response time per heavy-load cycle can be expressed as $\bar{t}_{hc} = (1 - \Psi_0^*)\Delta t + \Psi_0^*(\Delta t + \bar{t}_i)$, and the average power consumption per heavy-load cycle can be represented as $\bar{E}_h = 39\Delta t + 22.3\Psi_0^* \bar{t}_i$.

4.2. Light-workload Modeling

As can be seen from Figure 1, disk will switch to the idle mode after completing the last one if there is no I/O request existing in the queueing system. In the light-workload case, I/O requests are not intensive enough so that idle time is over fifteen seconds with the probability of $P = P(t_i \geq 15) = e^{-15\lambda}$. Disk will further spin down and switch to standby mode if idle time is over 15 seconds. After a random interval during standby, i.e., $\bar{t}_s = \frac{1}{\lambda}$, disk will spin up again as new I/O requests arrive, and a light-workload cycle is completed once disk switches to the active mode again. However, both spin-down and spin-up operations will increase disk energy consumption rapidly.

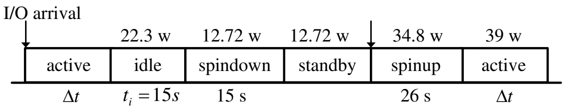


Figure 3 Disk states and transitions assumed in light workload.

In the light-workload case, a cycle will consist of five components: active, idle, spin-down, standby and spin-up intervals. So, in this cycle, it is reasonable to consider the average queue length as zero, i.e., $N_{lc} = 0$.

The average response time per light-load cycle is:

$$\bar{t}_{lc} = \Delta t + 15 + 15 + \bar{t}_s + 26$$
$$= \Delta t + \frac{1}{\lambda} + 56$$

The average power consumed per light-load cycle is:

$$\bar{E}_l = 39\Delta t + 12.72\frac{1}{\lambda} + 1430.1$$

The average queue length of I/O requests existing in the queueing system per cycle, N_c, can be represented as $N_c = N_{hc}$.

The average duration per cycle is:

$$\bar{t}_c = e^{-15\lambda}\bar{t}_{lc} + \bar{t}_{hc}(1 - e^{-15\lambda})$$

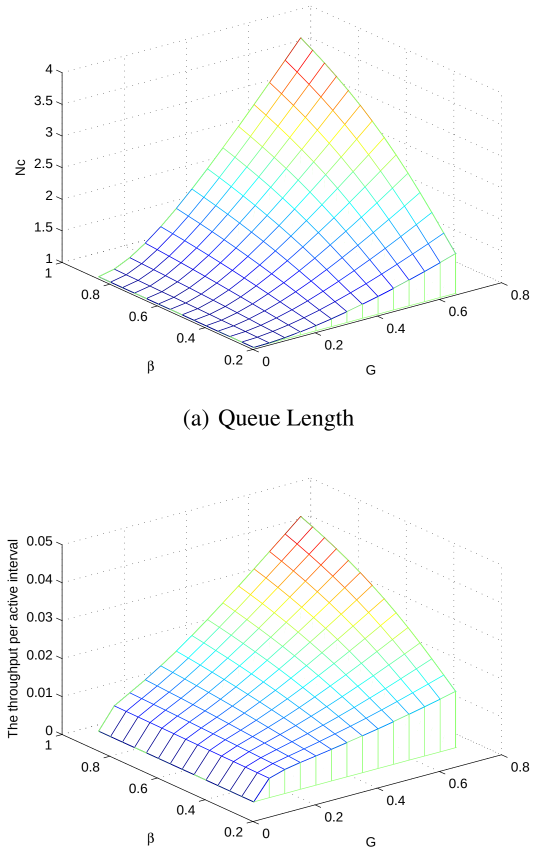
The total power consumption is the sum of the power used in each power mode of disk. The total power consumedcan be given as follows:

$$\bar{E} = e^{-15\lambda}\bar{E}_l + \bar{E}_h(1 - e^{-15\lambda})$$

The throughput per cycle can be represented as follows:

$$T = \frac{e^{15\lambda}[Ge^{-G\beta(1-e^{-1/\beta})} + G^2\beta(1 - e^{-1/\beta})]}{G + e^{-G\beta(1-e^{-1/\beta})} + (e^{15\lambda} - 1)(G + 56\lambda + 1)} \qquad (2)$$

5. Analytical Results



(a) Queue Length

(b) Throughput

Figure 4 (a), (b): Queue Length and Throughput.

© 2013 NSP
Natural Sciences Publishing Cor.

In order to further understand the analytical model proposed in this paper, and provide some useful insight to evaluate energy consumption as well as performance metrics for disk systems, we show some analytical results in this section.

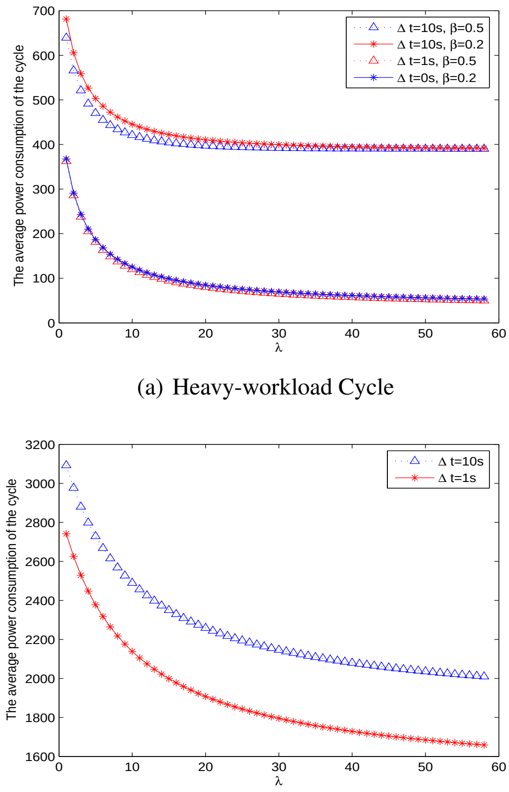
First, we plot the queue length curve in Fig. 4(a), as a function of service load, $G = \lambda\Delta t$, and the degree of I/O requests' patience, β. As shown in Fig. 4(a), the x-axis shows the service load, the y-axis denotes the degree of I/O requests' patience, and the z-axis indicates the average queue length of I/O requests surviving in the queueing system within a cycle. The throughput curve is also illustrated in Fig. 4(b), as a function of G and β. As shown in Fig. 4(b), the x-axis shows the service load, the y-axis denotes the degree of I/O requests' patience, and the z-axis indicates the throughput per cycle. As can be seen from Fig. 4, the degree of I/O requests' patience evidently effects the variety of queue length as $G > 0.2$.

We plot the power curves for the heavy-load case as a function of λ in Figure 5(a), as well as the light-load case as a function of λ in Figure 5(b), where the x-axis shows the arrival rate, λ, and the y-axis denotes the average energy consumption per cycle. As shown in Fig. 5(a) and (b), the average power consumption curve decreases slowly and approaches to a constant with the increase of I/O arrival rate. For example, in Fig. 5(a), the average power consumption approaches to 50W as the service interval, Δt, equals to 1 second, and 400W as the service interval is 10 seconds, respectively.

As can be seen from Fig. 5(a), for the service interval $\Delta t = 10s$, higher the degree of I/O requests' patience is, smaller the value of power consumption is. However, for the service interval $\Delta t = 1s$, the power consumption almost is not obviously influenced by the difference of I/O requests' patience. It indicates that, larger the size of accessed file is, longer the service interval is, then I/O requests' patience effects the value of power consumption much more.

As can be seen from Fig. 5(a), for the same patience such as $\beta = 0.5$, the difference between the power consumption as $\Delta t = 10s$ and the power consumption as $\Delta t = 1s$ is nearly a constant, about 300W. That might be explained by the fact that disk will stay at the active mode with a constant power consumption as I/O arrival rate is high enough. However, as can shown in Fig. 5(b), the difference between the power consumption as $\Delta t = 10s$ and the power consumption as $\Delta t = 1s$ approaches to a constant, about 350W. This observation can be explained by the different sizes of I/O requests. In fact, Δt represents service time that depends on the accessed file size.

Furthermore, comparing the former plot to the latter plot in Figure 5, we find that the power consumption in the former is higher than the latter for a magnitude. This observation indicates that the cost of power consumption in the light-workload case is far larger than the heavy-workload case because disk will go over all of states per cycle.



(a) Heavy-workload Cycle

(b) Light-workload Cycle

Figure 5 (a), (b): The average power consumption per cycle.

6. Conclusions

With the ever increase of power consumption, a lot of research work focused on improving energy-efficiency. However, there is little work to explore the performance-energy combined impact of disk-based storage systems. In this paper, based on the transition probability, we propose an analytical approach to model the energy consumption and performance metrics of I/O systems. Our model can provide the storage researchers a quick analysis toolkit and give some insights for the design of power-aware storage systems.

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper. This project is supported by the National Key Technology R&D Program under Grant No. 2012CB318208, the Fundamental Research Funds for the Central Universities under Grant No. XDJK2012A006, and the Ph.D. Foundation of Southwest University under Grants No. SWU111015.

© 2013 NSP
Natural Sciences Publishing Cor.

References

[1] E. Pinheiro and R. Bianchini. Energy Conservation Techniques for Disk Array-Based Servers. In Proceedings of the 18th International Conference on Supercomputing (ICS), June 2004.

[2] Q. Zhu and Y. Zhou. Power-Aware Storage Cache Management. IEEE Transactions on Computers, 54(5), 2005.

[3] X. Ge, D. Feng and D. H.C. Du. DiscPOP: Power-aware buffer management for disk accesses. Sustainable Computing: Informatics and Systems, In Press, Corrected Proof, Available online 3 April 2012.

[4] L. Prada, J. Garcia, J. D. Garcia and J. Carretero. Power saving-aware prefetching for SSD-based systems. Journal of Supercomputing, 2011, 58(3): 323-331.

[5] M. Sharifi, H. Salimi and M. Najafzadeh. Power-efficient distributed scheduling of virtual machines using workload-aware consolidation techniques. Journal of Supercomputing, 2012, 61(1): 46-66.

[6] M. Nijim, X. Qin, M. Qiu and K. Li. An adaptive energy-conserving strategy for parallel disk systems. Future Generation Computer Systems, January 2013, 29(1): 196-207.

[7] C. Weddle and et al. PARAID: A Gear-Shifting Power-Aware RAID. In Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST), 2007.

[8] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke. DRPM: Dynamic Speed Control for Power Management in Server Class Disks. In Proceedings of the ISCA, June 2003.

[9] Advanced Power Management-Next Generation. Intel Corp. and Microsoft Corp., August, 1991.

[10] Y. Won, J. Kim and W. Jung. Energy-aware disk scheduling for soft real-time I/O requests. Multimedia Systems, 2008, 13: 409-428.

[11] Y. Deng, F. Wang and N. Helian. EED: Energy Efficient Disk drive architecture. Information Sciences, November 2008, 178(22): 4403-4417.

[12] T. Xie and Y. Sun. Understanding the relationship between energy conservation and reliability in parallel disk arrays. Journal of Parallel and Distributed Computing, February 2011, 71(2): 198-210.

[13] J. Kim and D. Rotem. FREP: Energy proportionality for disk storage using replication. Journal of Parallel and Distributed Computing, August 2012, 72(8): 960-974.

[14] E. Pinheiro and et al. Exploiting Redundancy to Conserve Energy in Storage Systems. In Proceedings of SIGMETRICS/Performance'06, Saint Malo, France.

[15] D. Li and J. Wang. eRAID: A Queueing Model Based Energy Saving Policy. In Proceedings of the 14th MASCOTS, 2006.

[16] V. W. Freeh and et al. Exploring the Energy-Time Tradeoff in MPI Programs on a Power-Scalable Cluster. In Proceedings of the 19th IEEE IPDPS, 2005.

[17] R. Ge and K. W. Cameron. Power-Aware Speedup. In Proceedings of the 21st IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2007.

[18] Z. John, S. Sumeet, G. Nitin and et al. Modeling Hard-Disk Power Consumption. In Proceedings of the 1st USENIX Conference on File and Storage Technologies (FAST'03), 2003.

[19] S. W. Son, M. Kandemir and et al., Software-Directed Disk Power Management for Scientific Applications. In Proceedings of the 19th IEEE IPDPS, 2005.

[20] S. W. Son et al. Disk Layout Optimization for Reducing Energy Consumption. In Proceedings of ICS'05, Boston, MA, June 20-22, 2005.

[21] S. W. Son, G. Chen, M. Kandemir et al. Exposing Disk Layout to Compiler for Reducing Energy Consumption of Parallel Disk Based Systems. In Proceedings of PPoPP'05, Chicago, IL.

[22] E. V. Carrera, E. Pinheiro, and R. Bianchini. Conserving Disk Energy in Network Servers. In Proceedings of the 17th International Conference on Supercomputing, June 2003.

[23] F. Douglis, P. Krishnan, and B. Marsh. Thwarting the power-hungry disk. In Proceedings of the Winter USENIX Conference, 1994.

[24] K. Li, R. Kumpf, P. Horton, and T. E. Anderson. A quantitative analysis of disk drive power management in portable computers. In Proceedings of the Winter USENIX Conference, 1994.

[25] A. Weissel, B. Beutel, and F. Bellosa. Cooperative I/O: A novel I/O semantics for energyaware applications. In Proceedings of the Fifth OSDI, 2002.

[26] Adaptive power management for mobile hard drives. Tech. rep., IBM Corporation, April 1999.

[27] P. Greenawalt, Modeling power management for hard disks. In Proceedings of the Symposium on Modeling and Simulation of Computer Telecommunication Systems (MASCOTS), 1994.

[28] S. Ross. Applied Probability Models with Optimization Applications. UMI, Out-of-Print Books on Demand.

[29] IBM Hard Dish Drive-Ulvastar 36ZX, http://www.storage.ibm.com/hdd/ultra/ul36zx.html

© 2013 NSP
Natural Sciences Publishing Cor.

Dr. Qiang Zou, received his MS degree in applied mathematics and PhD degree in computer architecture, from Huazhong University of Science and Technology (HUST), Wuhan, China in 2005 and in 2009, respectively. He then worked as an Associate Professor in School of Computer and Information Science, Southwest University (SWU), China. His main research interests focus on workload characterization, Markov chain, storage system and performance evaluation. He has published several papers in journals and international conferences.

Dr. Yong Li, received his MS degree in applied mathematics and PhD degree in probability and statistics, from Huazhong University of Science and Technology (HUST), Wuhan, China in 2005 and in 2008, respectively. He then worked as an Assistant Professor in School of Mathematics Science, Chongqing Normal University, China. His current research interests include stochastic analysis and application, and stability theory.