A reliability model of energy-efficient parallel disk systems with data mirroring

Fangyang Shen* and Andres Salazar

Department of Engineering,
Northern New Mexico College,
921 Paseo De Onate,
Espanola, NM 87532, USA
E-mail: fshen@nnmc.edu
E-mail: asalazar@nnmc.edu
*Corresponding author

Xiao Qin

Department of CSSE,
Auburn University,
Auburn, AL 36849, USA
E-mail: xqin@auburn.edu

Min-Te Sun

Department of CSIE,
National Central University,
Chung-Li, Tao-Yuan 320, Taiwan
E-mail: msun@csie.ncu.edu.tw

Abstract: In the last decade, parallel disk systems have increasingly become popular for data-intensive applications running on high performance computing platforms. Conservation of energy in parallel disk systems has a strong impact on the cost of cooling equipment and backup power-generation. This is because a significant amount of energy is consumed by parallel disks in high performance computing centres. Although a wide range of energy conservation techniques have been developed for disk systems, most energy saving schemes have adverse impacts on the reliability of parallel disk systems. To address this deficiency, we must focus on reliability analysis for energy-efficient parallel disk systems. In this paper, we make use of a Markov process to develop a quantitative reliability model for energy-efficient parallel disk systems using data mirroring. With the new model in place, a reliability analysis tool is developed to efficiently evaluate reliability of fault-tolerant parallel disk systems with two power modes. More importantly, the reliability model makes it possible to provide good trade-offs between energy efficiency and reliability in energy-efficient and fault-tolerant parallel disk systems.

Keywords: storage systems; RAID 1; disk reliability model.

Reference to this paper should be made as follows: Shen, F., Salazar, A., Qin, X. and Sun, M-T. (2010) 'A reliability model of energy-efficient parallel disk systems with data mirroring', *Int. J. High Performance Systems Architecture*, Vol. 2, Nos. 3/4, pp.240–249.

Biographical notes: Fangyang Shen is currently an Assistant Professor at Northern New Mexico College. He received his PhD in Computer Science from the Samuel Ginn College of Engineering at Auburn University in Alabama, USA. His research interests include storage systems, wireless networks and information assurance. He has four years of industrial experience in networking and database management at Guangdong Fuel Company in Guangzhou, China.

Andres Salazar is currently a Professor at Northern New Mexico College (NNMC). Prior to his appointment at NNMC in 2007, he was the PNM Chair and a Professor on Microsystems, Commercialisation and Technology at the University of New Mexico, Albuquerque (UNM). Before starting his academic career at UNM, he had been in the industry for over 30 years as a Researcher, Manager and Executive at high-tech companies in New Jersey, Massachusetts, Florida, California and Georgia. He received his PhD in EE from Michigan State University.

Xiao Qin is an Assistant Professor in the Department of Computer Science and Software Engineering at Auburn University. He received his PhD in Computer Science from the University of Nebraska-Lincoln in 2004. He won an NSF CAREER award in 2009. His research interests include parallel and distributed systems, real-time computing, storage systems and fault tolerance. His research is supported by the US National Science Foundation and Intel Corporation. He had served as a Subject Area Editor of IEEE Distributed System Online. He has been on the programme committees of various international conferences, including IEEE Cluster, IEEE IPCCC and ICPP.

Min-Te Sun is currently an Assistant Professor in the Department of CSIE at National Central University in Taiwan. Before his current appointment, he has worked as an Assistant Professor in the Department of CSSE at Auburn University for six years. His research interests include ad hoc networks, wireless LAN, Bluetooth, sensor networks and wireless ATM. He received his PhD in Computer Science from Ohio State University, USA.

1 Introduction

In the last decade, parallel disk systems have been widely used to support a variety of data-intensive applications running on high performance computing platforms. The reason behind the popularity of parallel disk systems is that parallel I/O is a promising avenue to bridge the performance gap between processors and I/O systems (Lee and Lui, 2002). For example, redundant arrays of inexpensive disks or RAID can offer high I/O performance with large capacities (Lee and Lui, 2002). In storage systems field, reliability is a property of some disk arrays (most commonly in RAID systems) which provides fault tolerance, so that all or part of the data stored in the array can be recovered in the case of disk failure (http://en.wikipedia.org/wiki/Data_reliability). The RAID systems improve reliability because they prevent data loss using disk redundancy.

Recent studies show that a significant amount of energy is consumed by parallel disks in high performance computing centres (Qin, 2007; Zong et al., 2007). Energy conservation in parallel disk systems not only lowers electricity bills, but also leads to reductions of emissions of air pollutants from power generators. Moreover, the storage requirements of modern data-intensive applications and the emergence of disk systems with higher power needs make it desirable to design energy-efficient parallel disk systems. Energy conservation techniques developed for parallel disk systems include dynamic power management schemes (Douglis et al., 1994; Li et al., 1994), power-aware cache management strategies (Zhu et al., 2004), power-aware prefetching schemes (Son and Kandemir, 2006), and multi-speed settings (Gurumurthi et al., 2003; Helmbold et al., 2000; Krishnan et al., 2005). Many energy saving techniques significantly conserve energy in parallel disk systems at the apparent cost of reliability. For example, in multi-speed disk systems, disks spinning up and down may reduce the reliability of the disk systems; however, it can have the potential effect of saving energy for disk systems. Even worse, a parallel disk system consisting of multiple disks has a higher failure rate compared with a larger single disk system. This is due to the fact that there are more components in the parallel disk system. Due to hardware and software defects, failed components in a parallel disk system can ultimately cause failures in the system at run-time (Bitton, 1998). In addition to the energy efficiency issue, fault tolerance and reliability are major concerns in the design of modern parallel disk systems. These disk systems are expected to be the most stable part of high performance computing systems. Therefore, it is not surprising that academic institutes, industry, and government agencies consider the reliability and energy-efficiency of parallel disks in their computing systems essential and critical to operations.

The long-term goal of this research is to develop novel energy conservation schemes with marginal adverse impacts on the reliability of parallel disk systems. To achieve this long-term objective, in this study we investigate fundamental theories to model the reliability of energy-efficient parallel disk systems using data mirroring.

In this paper we present a quantitative approach to addressing the issue of conserving energy without noticeably degrading reliability of parallel disk systems. We focus on parallel disk systems with data mirroring, which is a technique for maintaining two or more identical disk images on multiple disk devices (see, for example, Bitton, 1998). The data mirroring technique leverages redundant data to significantly enhance reliability. Among the various energy conservation techniques, we focus on a widely adopted technique – dynamic power management, which dynamically changes the power state of disks. The dynamic power management technique turns disks into low power states (e.g., sleep state) when the disk I/O workload is fairly light. In this study, we use a Markov process to develop a quantitative reliability model for energy-efficient parallel disk systems with data mirroring. Next, we evaluate the reliability and energy efficiency of parallel disk systems using the new reliability model. Furthermore, using the Weibull distribution to model disk failure rates, we analyse the reliability of a real world parallel disk system. Finally, we propose a method capable of obtaining a balance between the reliability and energy efficiency of parallel disk systems with data mirroring.

The rest of the paper is organised as follows. Section 2 summarises related work. Section 3 describes the novel reliability model using a Markov process. In Section 4, energy efficiency and reliability of a parallel disk system with data mirroring are evaluated using a tool based on the reliability model. Finally, Section 5 concludes the paper with summary and future directions.

2 Related work

In this section, we will review the main components of this paper: reliability and energy saving for storage systems.

High reliability is one of the key design goals of modern parallel disk systems. In the past decade, a variety of practical and fundamental reliability models have been constructed for storage systems. For example, conventional reliability models include Markov chain models (Baek et al., 2001; Gibson and Patterson, 1993; Hou and Patt, 1997; Geist and Trivedi, 1993), analytic queuing models (Chen and Towsley, 1993, 1996), queuing networks (Drakopoulos and Merges, 1993), and iterative models (Geist, 1986). Although the existing models can accurately measure the reliability of disk systems, all of the models are inadequate to quantify the reliability of parallel disk systems, in which energy saving techniques are implemented.

The issue of energy efficiency in storage systems has received increasing attention. Energy conservation techniques proposed in previous studies include dynamic power management schemes (Douglis et al., 1994; Li et al., 1994), power-aware cache management strategies (Zhu et al., 2004), power-aware prefetching schemes (Son and Kandemir, 2006), software-directed power management techniques (Son et al., 2005), redundancy techniques (Pinheiro et al., 2006), multi-speed settings (Gurumurthi et al., 2003; Helmbold et al., 2000; Krishnan et al., 2005), and dynamic voltage scaling (Krishna and Lee, 2000; Pillai and Shin, 2001; Aydin et al., 2001). However, the previous research did not investigate the impact of energy saving techniques on the reliability of storage systems. Our work represents the first attempt to build reliability models for the design process of fault-tolerant and energy-efficient parallel disk systems.

A number of studies on energy conservation in parallel disk systems showed that energy-saving techniques should not sacrifice system reliability (Carrera et al., 2003; Gurumurthi et al., 2003; Pinheiro et al., 2006; Weddle et al., 2006; Zhu et al., 2005). However, little attention has been paid to the integration of reliability into energy-conservation techniques for disk arrays in general, and for mobile disk arrays in particular. It is imperative to build an energy-dependent reliability model to reveal relationships between energy-saving techniques and the reliability of parallel disk systems. Therefore, in this study we develop a new energy-aware reliability model. This model is accompanied by a reliability analysis tool able to systematically quantify the reliability of a particular type of energy-efficient and fault-tolerant parallel disk system.

Geist and Trivedi (1993) proposed an analytic model to measure the performance and reliability of mirrored disk systems. Their model shows that the disk mirroring technique can substantially improve both reliability and performance of disk systems. Our model is radically different from theirs in the fact that mirrored disk systems considered in our model have high energy-efficiency by the virtue of dynamic power management.

Aydin et al. (2001) developed a technique to leverage slack times in real-time systems to reduce energy consumption while achieving fault tolerance using check pointing policies. Unsal et al. investigated the relationship between fault tolerance techniques and energy consumption in real-time systems. This was achieved by establishing the energy efficiency of application level fault tolerance over other software-based fault tolerance methods (Unsal et al., 2002). Although these two studies integrated fault tolerance and energy-saving techniques in real-time systems, the above techniques are inadequate to address the reliability issues in energy-efficient parallel disk systems. Furthermore, our study utilises the Weibull distribution to model disk failure rates, whereas failure rates were not considered in the previous studies.

3 Modelling reliability of mirroring disks

3.1 Reliability models

In this study, we focus on mirroring disk systems (see Figure 1). Let us suppose disk failure and repair rates are modelled using the exponential distribution, we build a reliability model based on the Markov process to estimate the reliability and energy savings in parallel disk systems with data mirroring. We aim at utilising the reliability model to quantify the mean-time-to-data-loss (MTTDL) of a fault-tolerant parallel disk system with the data mirroring technique. Without loss of generality, we consider disks with two power states, i.e., a high voltage state and a low voltage state. In addition, we focus on disk systems with one primary disk and one backup disk. The Markov reliability models of mirroring disks are diagrammed in Figures 2 and 3. Note that Figure 2 plots the model for mirroring disks with two supply voltage levels; Figure 3 outlines the model for mirroring disks with a single voltage level.

Figure 1 A parallel disk systems with data mirroring (see online version for colours)

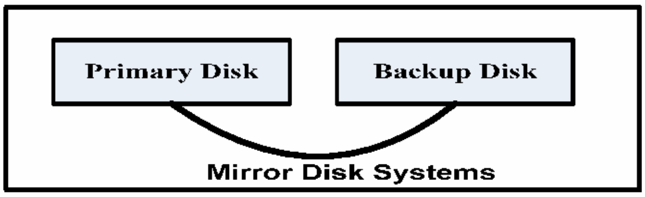


Figure 2 Mirroring disks with two supply voltage levels

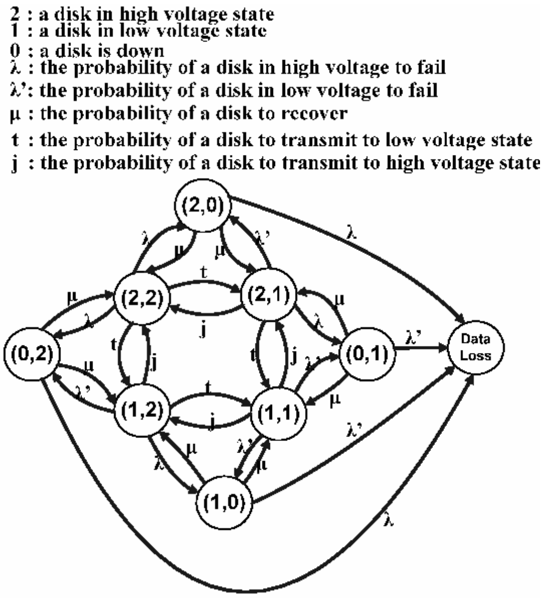
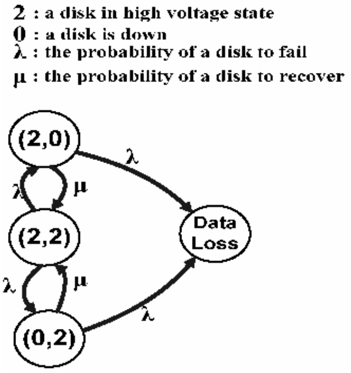

2 : a disk in high voltage state
1 : a disk in low voltage state
0 : a disk is down
λ : the probability of a disk in high voltage to fail
λ': the probability of a disk in low voltage to fail
μ : the probability of a disk to recover
t : the probability of a disk to transmit to low voltage state
j : the probability of a disk to transmit to high voltage state
(2,0)
(2,2)
(2,1)
(0,2)
(0,1)
(1,2)
(1,1)
(1,0)
Data Loss
λ
μ
λ'
t
j

Figure 3 Mirroring disks with one voltage level


2 : a disk in high voltage state
0 : a disk is down
λ : the probability of a disk to fail
μ : the probability of a disk to recover
(2,0)
(2,2)
(0,2)
Data Loss
λ
μ

The reliability of mirroring disks with two voltage states can be quantitatively characterised by parameters summarised in Table 1. The parameters (λ, μ, t, j) capture the reliability of mirroring disks with two supply voltage levels. Let (*Sp*, *Sb*) denote the system state, where *Sp*, *Sb* are the states of primary and backup disks, respectively. *Si* = 2 (*i* = *b* and *p*) signifies that the disk is at the high voltage level; *Si* = 1 (*i* = *b* and *p*) indicates that the disk is at the low voltage level; and *Si* = 0 (*i* = *b* and *p*) means that the disk is in the failure state. For example, an initial state (2, 2) indicates that both primary and backup disks are at the high voltage supply level. The state (0, 0) means that the primary and backup disks have unrecoverable failures, (i.e., data loss). Once the system state transits to one of the four states [(0, 2), (2, 0), (1, 0), (0, 1)], further failure can cause the system state to transition to the data loss state.

We model the failure rate of disks as a function of voltage. Thus, the failure rate function is:

$$\lambda(v) = v \cdot 10^{\frac{d(1-v)}{1-v_{low}}} \tag{1}$$

where v is the normalised supply voltage, v_{low} is the normalised minimum voltage, and d is a constant. A larger d indicates that the failure rate is more sensitive to the discrepancy between the high and low voltages. In our experiments, d is set to 0, 2, 4, 6, respectively. A similar failure rate model for processors was introduced by Zhu et al. (2004).

Table 1 Parameters used to characterise reliability of mirroring disks with two voltage levels

Sp = the state of a primary disk

Sp or *Sb* = 2: disks are in the high voltage state

Sp or *Sb* = 1: disks are in the low voltage state

λ = failure rate of a disk in the high voltage state

t = probability of transition from a high voltage state to a low voltage state

Sb = the state of a backup disk

Sp or *Sb* = 0: disks failed

μ = recovery rate

λ' = failure rate of a disk in the low voltage state

j = probability of transition from a low voltage state to a high voltage state

3.2 Reliability estimation

The MTTDL can be calculated using the fundamental matrix M defined as $M-[m_{ij}]-[I-Q]^{-1}$, where Q is the truncated matrix given below (Krishnan et al., 1995), and matrix I is the identity matrix corresponding to the dimension of matrix Q, N is the number of total disks.

$$Q = \begin{bmatrix} 1-N\lambda & N\lambda \\ \mu & 1-(N-1)\lambda-\mu \end{bmatrix} \tag{2}$$

The truncated stochastic transitional probability matrix for mirroring disks with two voltages is given as follows:

$$Q_1 = \begin{pmatrix} 1-2t-2\lambda & t & t & 0 & \lambda & \lambda & 0 & 0 \\ j & 1-t-j-\lambda-\lambda' & 0 & t & 0 & \lambda' & \lambda & 0 \\ j & 0 & 1-t-j-\lambda-\lambda' & t & \lambda' & 0 & 0 & \lambda \\ 0 & j & j & 1-2j-2\lambda' & 0 & 0 & \lambda' & \lambda' \\ \mu & 0 & \mu & 0 & 1-\lambda-2\mu-t & 0 & t & 0 \\ \mu & \mu & 0 & 0 & 0 & 1-\lambda-2\mu-t & 0 & t \\ 0 & \mu & 0 & \mu & j & 0 & 1-\lambda'-2\mu-j & 0 \\ 0 & 0 & \mu & \mu & 0 & j & 0 & 1-\lambda'-2\mu-j \end{pmatrix} \tag{3}$$

Similarly, we define the transitional probability matrix for mirroring disks with a single voltage as:

$$Q_2 = \begin{pmatrix} 1-\mu-\lambda & \mu & 0 \\ \lambda & 1-2\lambda & \lambda \\ 0 & \mu & 1-\mu-\lambda \end{pmatrix} \tag{4}$$

We assume that the mirroring disk system starts in state (2, 2); thus, MTTDL can be expressed as $MTTDL = \sum_{j=1}^{d+1} m_{ij}$.

3.3 Energy consumption model

The following parameters are introduced to model energy dissipation rates in mirroring disks.

T_{total} The total operation time of a disk system.

E_{total} The total energy consumption (measured in Joule) in a total time of T_{total}.

P_{ij} The probability that a disk system is in state (i, j). Thus, we have $\sum_{j=0}^{2}\sum_{i=0}^{2} P_{ij} = 1$.

V_L The low voltage of a disk; V_H: the high voltage of a disk.

The energy consumption of a mirroring disk system with two voltage levels can be written as:

$$E_{total} = 3600 \cdot \left(\left(\sum_{i=0}^{2} P_{i1} + \sum_{j=0}^{2} P_{1j} \right) \cdot V_L + \left(\left(\sum_{i=0}^{2} P_{i2} + \sum_{j=0}^{2} P_{2j} \right) \cdot V_H \right) \right) \cdot T_{total} \tag{5}$$

The energy consumption of a mirroring disk system with a single high voltage within 24 hours is:

$$E_{total} = 3600 \cdot \left(P_{22} + P_{02} + P_{20}\right) \cdot V_H \cdot T_{total}. \tag{6}$$

3.4 Failure rates with Weibull distribution

We choose to use the Weibull distribution to model disk failure rates. The probability density function of the general Weibull distribution (Krishna and Lee, 2000; Pillai and Shin, 2001; Aydin et al., 2001) is provided using equation (7).

$$f(x) = \frac{\gamma}{\alpha}\left(\frac{x-\mu}{\alpha}\right)^{\gamma-1} \exp\left(-\left((x-\mu)/\alpha\right)^{\gamma}\right) x \geq \mu; \gamma, \alpha > 0 \tag{7}$$

where γ is the shape parameter, μ is the location parameter and α is the scale parameter. The case where $\mu = 0$ is called the two-parameter Weibull distribution. This equation reduces to the standard Weibull distribution:

$$f(x) = \gamma x^{(\gamma-1)} \exp\left(-\left(x^{\gamma}\right) x \geq 0; \gamma > 0,. \tag{8}$$

If the disk failure rate decreases over time, then $\gamma < 1$. When disk failure rate is constant over time, $\gamma = 1$. If the disk failure rate is increasing over time, then $\gamma > 1$. In this study, we set the shape parameter γ to 2 (Pillai and Shin, 2001; Aydin et al., 2001).

3.5 Energy efficiency vs. reliability

Recall that t is the probability of a transition from the high to low voltage. To balance energy efficiency and reliability, we have to determine an optimal value t for a specified reliability requirement α. Thus, a high reliability requirement leads to a small value of t, and vice versa. A small value of t in turn results in high energy consumption.

We first normalise energy consumption and MTTDL denoted as follows.

N_{energy} normalised energy consumption for mirroring disk systems with two voltages

$E_{constant}$ energy consumption when $t = 0.1$

N_{mttdl} normalised MTTDL for mirroring disk systems with two voltages

$MTTDL_{constant}$ MTTDL when $t = 0.1$.

Next, we calculate the energy dissipation as $E_{energy} = \frac{E_{ij}}{E_{contant}}$.

The normalised MTTDL is written as $N_{mttdl} = \frac{MTTDL_{ij}}{MTTDL_{constant}}$.

We can now obtain the following energy/reliability trade-off problem formulation:

$$\begin{aligned} &\textit{Minimise} \quad E_{energy} = \frac{E_{ij}}{E_{contant}} \\ &\text{Subject to} \quad Nmttdl > \alpha \end{aligned} \tag{9}$$

By adjusting the value of α, we can achieve different level of reliability for mirror disk systems.

4 Performance evaluation

To determine the strength of our reliability model, we make use of the model to measure the reliability and energy dissipation of two mirroring disk systems. The first mirroring disk system is energy efficient by the virtue of two supply voltages; the second mirroring disk system is a traditional system without dynamic power management. Table 2 summarises the key configuration parameters of a mirroring disk system. The parameters are chosen to resemble real world disks like the Seagate ST-34501W.

Table 2 Parameters of energy-efficient mirroring disks

Parameters	*Default values*	*Values I*	*Values II*	*Values III*	*Values IV*
MTTF (year)	5–30	5, 10, 15, 30	5, 10, 15, 30	5–30	30
MTTR (day)	1, 3, 6, 7	1–7	1, 3, 7	1	1
T	0.1, 0.5, 0.9	0.1, 0.5, 0.9	0.1, 0.5, 0.9	0.1, 0.5, 0.9	Variable
D	2	2	2	2	2
High voltage (W)	13.5	13.5	13.5	13.5	13.5
Low voltage (W)	10.2	10.2	10.2	10.2	10.2
Time (hour)	24	24	24	24	24
Fmin	0.46	0.46	0.46	0.46	0.46

Note: MTTR – mean-time-to-recover

Figure 4 Impact of disk failure rate λ on MTTDL (see online version for colours)

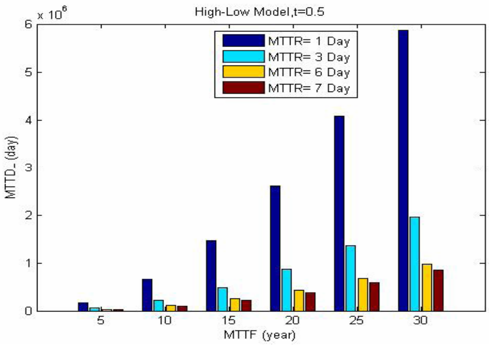
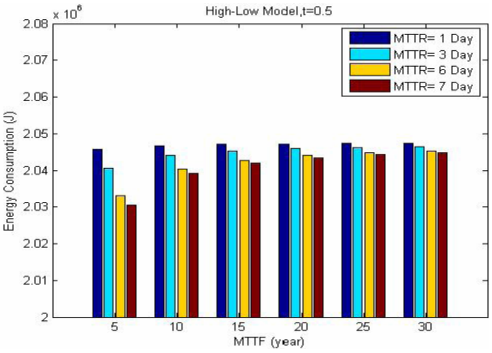


Figure 5 Impact of disk failure rate λ on energy (see online version for colours)



4.1 Impact of disk failure rate

The relationship between disk failure rate λ and MTTDL is shown in Figure 4. Similarly the relationship between λ and energy consumption is shown in Figure 5. The results plotted in Figures 4 and 5 are obtained by using default values taken from Table 2. It is evident from Figure 4 that MTTDL significantly increases with the increase of λ. This is because MTTDL of a system is directly related to MTTF. We observe from Figure 5 that the energy consumption of a disk system increases gradually with the increase of λ. The reason is that more energy is consumed by the disk system when each disk is active for a longer period of time. Figures 4 and 5 suggest that gradually increasing λ allows us to find a compromise between MTTDL and energy consumption.

4.2 Impact of MTTR

Figure 6 reveals the correlation between mean-time-to-recover (MTTR) and MTTDL, whereas Figure 7 shows the relationship between μ and energy consumption. The simulation settings for these figures are using Values I in Table 2. We observe that MTTDL and MTTR are inversely related. This result is expected because a large MTTR leads to a low recover rate μ, which is also inversely proportional to MTTR. As the recovery rate is lowered, the reliability of mirrored disks also decreases. Figure 7 clearly illustrates that the energy consumption for a disk system decreases slowly with the increase of MTTR. This is mainly because a large value of MTTR implies a smaller likelihood of having a failed disk repaired.

Figure 6 Impact of recovery rate on MTTDL (see online version for colours)

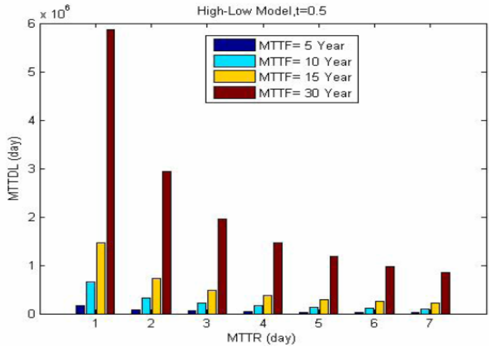
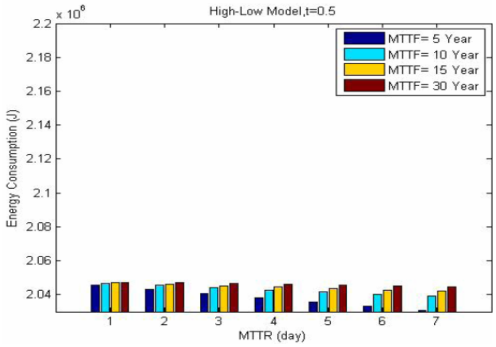


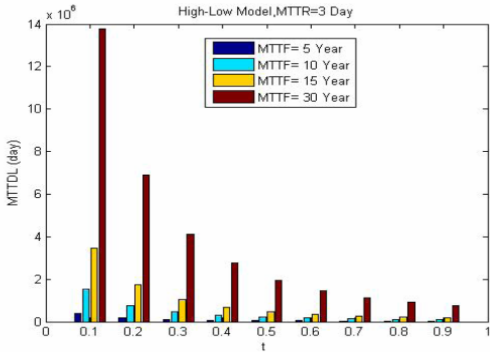
Figure 7 Impact of recovery rate on energy consumption (see online version for colours)



4.3 Probability of power transitions

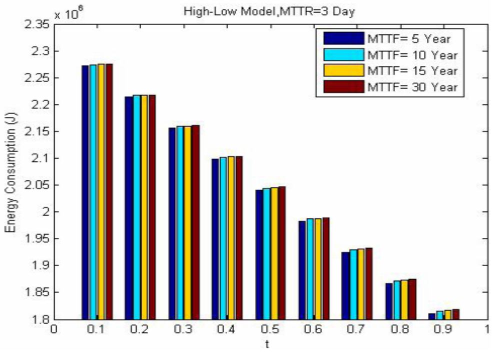
Recall that parameter *t* represent the probability of a transition between the high voltage and low voltage states. Figure 8 shows the relationship between *t* and MTTDL; Figure 9 plots the correlation between *t* and energy consumption. The simulation settings can be found in the third column (Values II) in Table 2. Figure 8 reveals that MTTDL decreases rapidly with the increase of *t*. This is because the MTTDL of a system is related to the amount of time each disk stays in the high voltage state. A smaller *t* means a disk stays longer periods of time in the high voltage state, thereby enhancing the reliability of the disk system. This implies that we can increase the reliability of a disk system by increasing the supply voltage level of each disk in the system. Figure 9 shows that the energy consumption for a disk system gradually decreases with the increase of *t*. This is because less energy is consumed by the disk system when each disk stays in the low voltage state for a longer period of time.

Figure 8 Impact of t on MTTDL (see online version for colours)



Note: *t* – probability of transition from high voltage to low voltage

Figure 9 Impact of t on energy (see online version for colours)



Note: *t* – probability of transition from high voltage to low voltage

4.4 Impact of parameter d

Recall that parameter *d* in equation (1) represents the sensitivity of the failure rate on disk voltage supplies. Figures 10 and 11 plot the impacts of the parameter *d* on MTTDL and energy consumption respectively. The settings of this experiment are identical to the previous experiments except that we now vary d from 0 to 6 by 2. Figure 10 shows that increasing the value of *d* results in a rapid decrease in MTTDL. The rationale behind this result is that MTTDL relies on failure rate, which is directly related to d. We can increase the reliability of a disk system by setting a small value of *d*. In contrast, when *d* is smaller than six, energy consumption is less sensitive to *d*. The energy consumption of the disk system will now increase slowly with the increase of *d* (see Figure 11).

Figure 10 Impact of parameter d on MTTDL (see online version for colours)

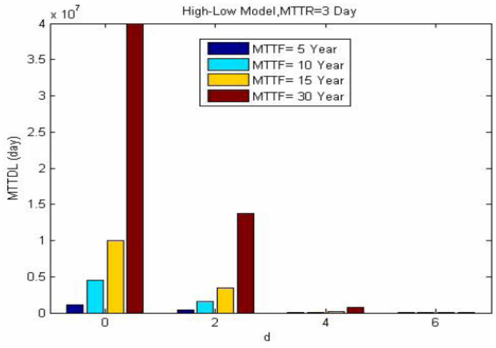
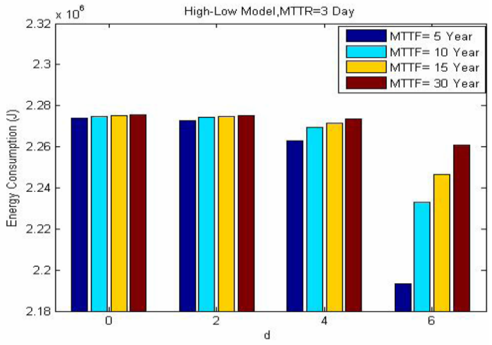


Figure 11 Impact of parameter d on energy consumption (see online version for colours)



4.5 Impact of high voltage level

Figure 12 shows the relationship between high voltage and MTTDL, whereas Figure 13 illustrates the effect high voltage has on energy consumption. Figure 12 shows that MTTDL decreases with an increasing value of the high voltage. This trend can be explained by the fact that MTTDL is likely to decrease if there are a large number of transitions between the high and low voltages. As the gap increases between the high and low voltages, the disk system reliability decreases. We observe from Figure 13 that the energy consumption of the disk system increases with increasing values of high voltage. This is because more energy dissipation is caused by disks when they stay in the high voltage state for longer periods of time.

Figure 12 Impact of high voltage level on MTTDL (see online version for colours)

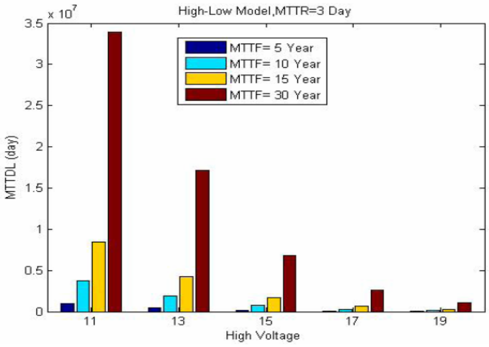
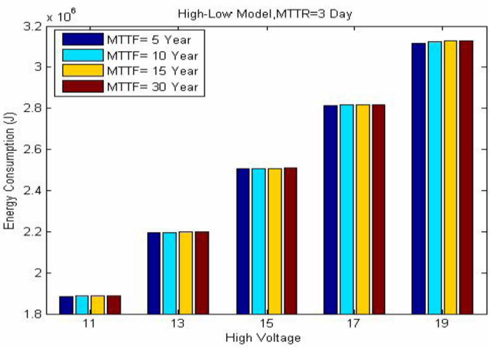


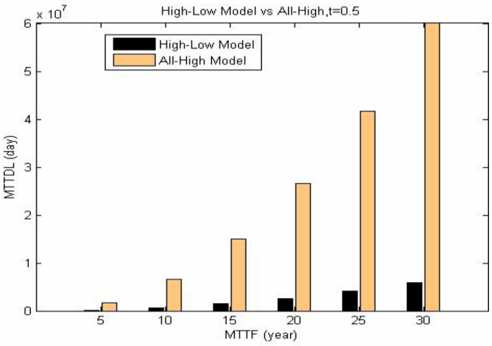
Figure 13 Impact of high voltage level on energy consumption (see online version for colours)



4.6 Real world disks

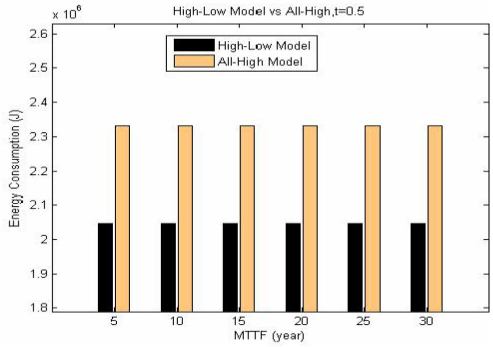
To further validate our reliability model, we conducted experiments using system parameters from two real world disks, the IBM 36Z15 (see Figures 14 and 15) and IBM 73LZX (see Figures 16 and 17). The two disks have different voltage supplies. It is observed from Figures 14–17 that when we increase the probability *t* of a transition from high voltage to low voltage, the MTTDL values of the two disks are quite different. This confirms our previous findings that high reliabilities can be achieved at a marginal cost of energy efficiency. The results validate that our model can apply to the real world disks.

Figure 14 MTTDL of IBM36Z15 (see online version for colours)



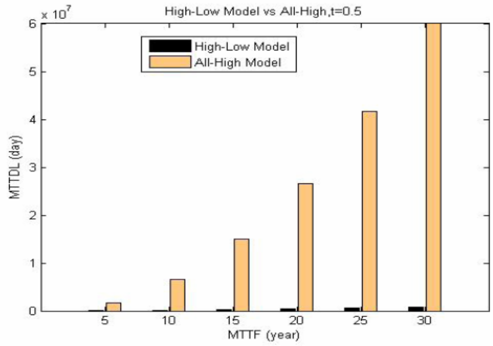
Notes: Value III in Table 2; high voltage = 13.5 W and low voltage = 10.2 W

Figure 15 Energy in IBM36Z15 (see online version for colours)



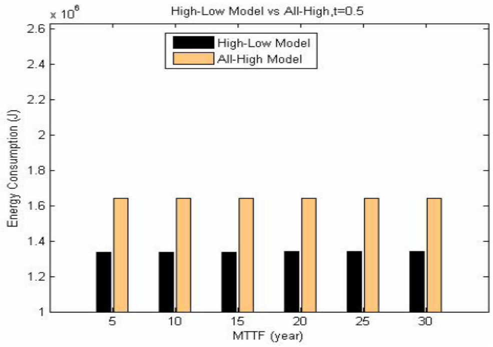
Notes: Value III in Table 2; high voltage = 13.5 W and low voltage = 10.2 W

Figure 16 MTTDL of IBM73LZX (see online version for colours)



Notes: Value III in Table 2; high voltage = 9.5 W and low voltage = 6 W

Figure 17 Energy in IBM73LZX (see online version for colours)



Notes: Value III in Table 2; high voltage = 9.5 W and low voltage = 6 W

4.7 Model fault inter-arrival times using the Weibull distribution

In all the previous experiments, we assumed that inter-arrival times of failures were generated based on the Poisson distribution. A recent study shows that the Weibull distribution fits statistical properties of the time between disk failures closely (Schroeder and Gibson, 2007). Hence, in this experiment we chose to model inter-arrival times of disk failures using the Weibull distribution.

Figure 18 shows the values of MTTDL when the parameter *t* is set to 0.5. Similarly, Figure 19 shows energy dissipation in the mirroring disks when *t* is set to 0.5. Again, we compare the mirroring disk using two voltage levels against one with only a high voltage level. Simulation results plotted in Figures 18 and 19 confirm that both energy consumption and MTTDL decrease with the increasing value of *t*.

Figure 18 Using the Weibull distribution to model disk failures; IBM 36Z15; MTTF vs. MTTDL (see online version for colours)

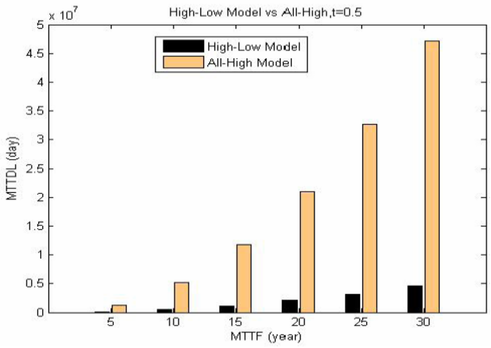
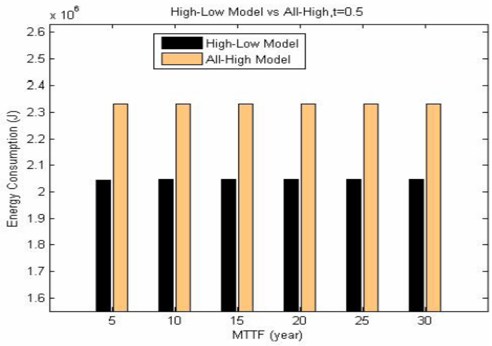


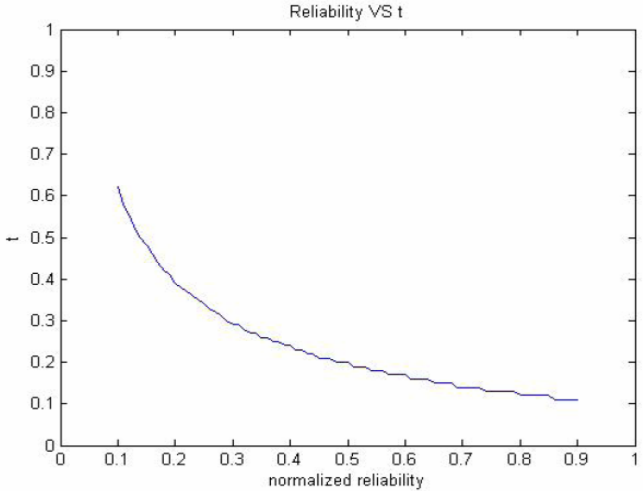
Figure 19 Using the Weibull distribution to model disk failures; IBM 36Z15; MTTF vs. energy (see online version for colours)



4.8 *Reliability vs. energy efficiency*

Now we are able to study the trade-offs between reliability and energy efficiency of parallel disk systems with data mirroring. Figure 20 plots the normalised MTDDL as a function of energy efficiency represented in the form of a parameter *t*, which is the probability of transiting to the low voltage level from the high voltage level. A large value of *t* indicates high energy efficiency in mirroring disks. Figure 20 quantitatively shows that high energy efficiency can be achieved at the cost of reliability. Our model is practical in the sense that storage system designers can apply the model to obtain a balance of reliability and energy efficiency in parallel disk systems. For example, given a minimum reliability requirement, one can leverage this model to aggressively reduce energy consumption subject to the reliability requirement.

Figure 20 Trade-off between normalised MTDDL and probability of transition from the high voltage to low voltage (see online version for colours)



5 Conclusions

Although parallel disk systems are a powerful means of bridging the performance gap between processors and disk I/O, the success of high performance computing centres using parallel disk systems largely depend on the energy efficiency and reliability of storage systems. A wide range of energy conservation techniques have been developed for parallel disk systems. However, most energy saving schemes has significant adverse impacts on the reliability of parallel disk systems. Even worse, a large-scale parallel disk system can have higher failure rates compared with a larger single disk system. Thus, fault tolerance and reliability are major concerns in the design of energy-efficient parallel disk systems.

In this paper we focused on energy-efficient parallel disk systems with data mirroring. We developed a reliability model for mirroring disk systems where the dynamic power management technique is employed to significantly reduce energy dissipation in disks. Our reliability model, which is based on a Markov process, makes it possible to estimate reliability of mirroring disks with two supply voltage levels. With the reliability model in place, we developed a tool to evaluate the reliability and energy efficiency of parallel disk systems with data mirroring. To better quantify the reliability of real world disk systems, we not only used the Weibull distribution to model disk failure rates, but also analysed the reliability of two real world disk systems. In future studies, we first will extend our model by incorporating the effects and cost of state transitions between high voltage and low voltage. Then, we will develop a reliability model for energy efficient RAID systems, which are widely used in data intensive computing systems.

References

Aydin, H., Melhem, R., Moss, D. and Alvarez, P.M. (2001) 'Dynamic and aggressive scheduling techniques for power-aware realtime systems', *Proc. Real-Time Systems Symp.*, pp.95–106.

Baek, S.H., Kim, B.W., Joung, E.J. and Park, C.W. (2001) 'Reliability and performance of hierarchical RAID with multiple controllers', *Proc. ACM Int'l Symp. Principles of Distri. Comp.*, pp.246–254.

Bitton, D. (1998) 'Disk shadowing', *Proc. Int'l Conf. Very Large Data Bases*, pp.331–338.

Carrera, E.V., Pinheiro, E. and Bianchini, R. (2003) 'Conserving disk energy in network servers', *Proc. Int'l Conf. Supercomputing*, pp.86–97.

Chen, S. and Towsley, D. (1993) 'The design and evaluation of RAID 5 and parity striping disk array architectures', *J. Parallel and Distributed Comp.*, Vol. 17, pp.58-74.

Chen, S. and Towsley, D. (1996) 'A performance evaluation of RAID architectures', *IEEE Trans. Computers*, Vol. 45, No. 10, pp.1116–1130.

Douglis, F., Krishnan, P. and Marsh, B. (1994) 'Thwarting the power-hunger disk', *Proc. USENIX Conf.*

Drakopoulos, E. and Merges, M.J. (1993) 'Performance analysis of client-server storage systems', *IEEE Trans. Computers*, Vol. 41, No. 11, pp.1442–1452.

Geist, E., Finkel, D. and Tripathi, S.K. (1986) 'Availability of a distributed computer systems with failures', *Acta Informatica*, Vol. 23, No. 6, pp.643–655.

Geist, R. and Trivedi, K. (1993) 'An analytic treatment of the reliability and performance of mirrored disk subsystems', *Proc. Int'l Symp. Fault-Tolerant Comp.*, pp.442–450.

Gibson, G.A. and Patterson, D.A. (1993) 'Designing disk arrays for high reliability', *J. Parallel and Distri. Computing*, Vol. 17, Nos. 1–2, pp.4–27.

Gurumurthi, S., Sivasubramaniam, A., Kandemir, M. and Fanke, H. (2003) 'DRPM: dynamic speed control for power management in server class disks', *Proc. Int'l Symp. Computer Arch.*, pp. 169–179.

Gurumurthi, S., Zhang, J., Sivasubramaniam, A., Kandemir, M., Fanke, H., Vijaykrishnan, N. and Irwin, M. (2003) 'Interplay of energy and performance for disk arrays running transaction processing workloads', *IEEE Int'l Symp. Perf. Analy. Sys. and Software*, pp.123–132.

Helmbold, D.P., Long, D.E., Sconyers, T.L. and Sherrod, B. (2000) 'Adaptive disk spin-down for mobile computers', *Mobile Networks and Applications*, Vol. 5, No. 4, pp.285–297.

Hou, R.Y. and Patt, Y.N. (1997) 'Using non-volatile storage to improve the reliability of RAID5 disk arrays', *Proc. Int'l Symp. Fault-Tolerant Comp.*, pp.206–215.

Krishna, C.M. and Lee, Y.H. (2000) 'Voltage-clock-scaling techniques for low power in hard real time systems', *IEEE Real-Time Technology and Appl. Symp.*, pp.156–165.

Krishnan, P., Long, P. and Vitter, J. (1995) 'Adaptive disk spindown via optimal rent-to-buy in probabilistic environments', *Proc. Int'l Conf. Machine Learning*, pp.322–330.

Lee, Y.B. and Lui, C.S. (2002) 'Automatic recovery from disk failure in continuous-media servers', *IEEE Trans. Parallel and Distr. Sys.*, Vol. 13, No. 5, pp.499–515.

Li, K., Kumpf, R., Horton P. and Anderson, T.E. (1994) 'A quantitative analysis of disk drive power management in portable computers', *Proc. Winter USENIX Conf.*, pp.279–292.

Pillai, P. and Shin, K.G. (2001) 'Real-time dynamic voltage scaling for low-power embedded operating systems', *Proc. ACM Symp. Operating Sys. Principles*.

Pinheiro, E., Bianchini, R. and Dubnicki, C. (2006) 'Exploiting redundancy to conserve energy in storage systems', *Proc. Sigmetrics and Performance*, pp.15–26.

Pinheiro, E., Bianchini, R. and Dubnicki, C. (2006) 'Exploiting redundancy to conserve energy in storage systems', *Proc. Sigmetrics and Performance*, pp.15–26.

Qin, X. (2007) 'Design and analysis of a load balancing strategy in data grids', *Future Generation Comp. Sys.: The Int'l J. Grid Comp.*, Vol. 23, No.1, pp.132–137.

Schroeder, B. and Gibson, G.A. (2007) 'Disk failures in the real world: what does an MTTF of 1,000,000 hours mean to you?', *5th USENIX Conf. File and Storage Tech.*, pp.1–16.

Son, S.W. and Kandemir, M. (2006) 'Energy-aware data prefetching for multi-speed disks', *Proc. ACM Int'l Conf. Comp. Frontiers*, pp.105–114.

Son, S.W., Kandemir, M. and Choudhary, A. (2005) 'Software directed disk power management for scientific applications', *Proc. Int'l Symp. Parallel and Dist. Processing*, pp.4–14.

Unsal, S., Koren, I. and Krishna, C.M. (2002) 'Towards energy-aware software-based fault tolerance in real time systems', *Proc. Intl' Symp. Low Power Electronics and Design*, pp.124–129.

Weddle, C., Oldham, M., Qian, J. and Wang, A. (2006) 'PARAID: the gear-shifting power-aware RAID', Technical Report 060323, Florida State University, 2006.

Zhu, D-K., Melhem, R. and Mosse, D. (2004) 'The effects of energy management on reliability in real-time embedded systems', *Proc. IEEE/ACM Int'l Conf. Computer-aided Design*, pp.528–534.

Zhu, Q., Chen, Z., Tan, L., Zhou, Y., Keeton, K. and Wilkes, J. (2005) 'Hibernator: helping disk arrays sleep through the winter', *Proc. 12th ACM Symp. Operating Sys. Principles*, pp.177-190.

Zhu, Q., David, F.M., Devaaraj, C.F., Li, Z., Zhou, Y. and Cao, P. (2004) 'Reducing energy consumption of disk storage using power-aware cache management', *Proc. High-Performance Computer Architecture*, pp.118–129.

Zong, Z.L., Briggs, M.E., O'Connor, N.W. and Qin, X. (2007) 'An energy-efficient framework for large-scale parallel storage systems', *Proc. Int'l Parallel and Distributed Processing Symp.*, CA.