MINT: A Reliability Modeling Framework for Energy-Efficient Parallel Disk Systems

Shu Yin, Xiaojun Ruan, Adam Manzanares, Xiao Qin, and Kenli Li

Abstract—The Popular Disk Concentration (PDC) technique and the Massive Array of Idle Disks (MAID) technique are two effective energy conservation schemes for parallel disk systems. The goal of PDC and MAID is to skew I/O load toward a few disks so that other disks can be transitioned to low power states to conserve energy. I/O load skewing techniques like PDC and MAID inherently affect reliability of parallel disks, because disks storing popular data tend to have high failure rates than disks storing cold data. To study reliability impacts of energy-saving techniques on parallel disk systems, we develop a mathematical modeling framework called MINT. We first model the behaviors of parallel disks coupled with power management optimization policies. We make use of data access patterns as input parameters to estimate each disk's utilization and power-state transitions. Then, we derive each disk's reliability in terms of annual failure rate from the disk's utilization, age, operating temperature, and power-state transition frequency. Next, we calculate the reliability of PDC and MAID parallel disk systems in accordance with the annual failure rate of each disk in the systems. Finally, we use real-world trace to validate out MINT model. Validation result shows that the behaviors of PDC and MAID which are modeled by MINT have a similar trend as that in the real-world.

Index Terms—Parallel disk system, energy conservation, reliability, MAID, PDC, load balancing

1 Introduction

Parallel disk systems are of great value to large-scale parallel computers, because parallel disks are capable of providing high I/O performance with large storage capacity [2]. In the past decades, parallel disk systems have increasingly become popular for data-intensive applications running on massively parallel computing platforms [37]. Parallel disk systems comprised of arrays of independent disks are usually cost-effective, since the parallel disk systems can be built from low-cost commodity hardware components.

Recent studies indicate that the energy cost and carbon footprint of parallel disk systems and storage services has become exorbitant [42]. More specifically, storage devices account for approximately 27 percent of the overall energy consumption in a data centre [42]. When it comes to Web proxies, disk energy consumption may account for up to 77 percent [6]. Current utilization and technological trends of parallel disk systems result in unacceptable economical and environmental consequences [20]. To address this problem, a broad spectrum of energy-saving techniques were proposed to achieve high energy efficiency in storage systems. Well-known energy conservation techniques include software-directed power management strategies [33], dynamic power management (DPM) schemes [8], data redundancy techniques [3], [22], [48], [39], [42], workload skew [15], [46], and multi-speed settings [11], [34].

Prior findings show that existing energy conservation techniques in disk drives can deliver significant energy savings in large-scale storage systems. Although few energy-saving schemes such as cache-based energy saving approaches may have marginal impacts on disk reliability, many energy conservation techniques like dynamic power management and workload skew techniques inevitably have adverse impacts on parallel disk systems [3], [43]. For example, the DPM technique reduce energy consumption in disks by the virtue of frequent disk spin-downs and spin-ups, which in turn can shorten disk lifetime [8], [12], [17]. Unlike DPM, workload-skew techniques, for example, MAID [7], PDC [20], BUD [27], and PARAID [40] move frequently accessed data sets to a subset of disks arrays acting as workhorses, thereby keeping other disks in standby mode to save energy. Disks archiving hot data inherently have higher risk of breaking down than those disks storing cold data.

To address the reliability issues of energy-efficient parallel disks, we first proposed a modeling framework—MINT—to perform reliability analysis of parallel disks employing energy-saving techniques. Next, we applied the MINT framework to build two reliability models used to study reliabilities of parallel disk systems coupled with the two well-known energy-saving schemes: the PDC technique [20] and the Massive Array of Idle Disks (MAID) [7]. Finally, we use real-world trace to validate the Access-Pattern-to-Utilizatoin sub model of the MINT model.

This paper builds upon our previous work on modeling reliability of energy-efficient parallel disk systems [47]. We extend our previous work in the following five important ways.

- S. Yin and K. Li are with the School of Information Science and Engineering, Hunan University, Changsha, Hunan 410082, P.R. China. E-mail: {shuyin, jt_lkl}@hnu.edu.cn.
- X. Ruan is with the Computer Science Department, West Chester University of Pennsylvania, West Chester, PA 19383. E-mail: xruan@wcupa.edu.
- A. Manzanares is with the Computer Science Department, California State University-Chico, CA 95929, E-mail: amanzanares@csuchico.edu.
- X. Qin is with the Department of Computer Science and Software Engineering, Auburn University, Shelby Center for Engineering Technology, Suite 3101E, Auburn, AL 36849. E-mail: xqin@auburn.edu.

Manuscript received 2 Jan. 2013; accepted 30 Sept. 2013. Date of publication 3 Nov. 2013; date of current version 16 July 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TDSC.2013.47

- We added a new section that highlights the motivation of this work by discussing the limitations of previous energy-saving techniques and disk reliability models.
- We extended the framework by carefully considering power-state transitions in individual disks. The number of power-state transitions depends on the number of idle time periods that are larger than break-even time. Note that the break-even time is defined as the smallest time interval for which a disk must remain idle to save energy by transitioning to a low-power state.
- We provided details on modeling disk utilization and power-state transition frequency, which are two critical reliability-affecting factors. The disk utilization is used to estimate base annual failure rates (AFRs), whereas the power-state transition frequency is an adder to the base failure rates.
- We conducted stress test on the reliability model of parallel disks equipped with PDC and MAID under high I/O workload conditions. We observed that the utilization trends and annual-failure-rate (AFR) trends of PDC and MAID under heavy workload are different from the utilization and AFR trends of PDC and MAID dealing with light I/O loads.
- We developed a trace-driven model to validate our MINT. After using Poisson Distribution to model the access patterns of PDC and MAID, we further applied Berkeley Web Trace [1] to the Access Pattern-Utilization submodel in MINT. The results indicated that the system utilizations calculated according to our mathematical model tend to have the similar trend as those according to the real-world traces, hence validated the MINT.

The remainder of this paper is organized as follows. Section 2 presents the motivation for the development of a reliability modeling framework for energy-efficient parallel disks. Section 3 outlines the design and implementation of the MINT reliability modeling framework. In Section 4, we leverage MINT to build the reliability models for the Popular Data Concentration scheme (PDC) and the MAID technique. Section 4 also presents reliability impacts of PDC and MAID on parallel disk systems. Section 5 explains means of validating the MINT's submodel-Access-Pattern-to-Utilization. The related work of this study is discussed in Section 6. Section 7 concludes the paper with further discussions.

2 Motivation

2.1 Limitations of Previous Modeling Studies

It is often challenging to improve both reliability and energy efficiency of storage systems, because little attention has been paid to evaluating reliability impacts of power management strategies on storage systems. Many excellent reliability models have been proposed for disk systems (see, for example, [5] and [36]). Shah and Elerath conducted a series of reliability analysis using field failure data of several drive models from various disk disk manufacturers [32]. Hughes and Murray investigated reliability factors of cost-effective serial ATA (SATA) disk drives [13]. They not only studied failure probability of RAID storage systems, but also proposed approaches to improving reliability of storage systems comprised of multiple SATA disks [13]. However, the lack of considerations of disks power transitions and dynamic changing of system utilization makes models mentioned above inaccurate when applied to estimate reliability of energy-efficient disk systems.

Due to the following two reasons, the above disk reliability models are inadequate for evaluating reliability of disk systems equipped with energy-saving mechanisms.

- First, focusing on traditional disk systems, the existing models do not take disk-power transitions into account. Recently studies show that disk reliability is affected by power-state transitions (see Section 3.4, [35] and [43] for details on power-state transitions).
- Second, most existing models assume that disk utilization is unchanged. When I/O workload does not change dramatically, these models can accurately estimate the reliability of disk systems. In many real-world data-intensive processing environments (e.g., Video-On-Demand), I/O load conditions change fairly dynamically. Even under seemingly constant I/O loads, energy conservation techniques tend to make the constant I/O loads changing (see Section 4.1.2 and Section 4.2.2).

In addition to reliability models, simulation studies are another way of investigating the reliability of energy-efficient storage systems (see, for example, [43]). Simulating the reliability of disk systems with energy-conservation techniques is complicated, because accurate simulations of the energy-saving techniques requires storage researchers to seamlessly integrate the energy conservation schemes with conventional disk simulators. It is common sense that reliability of energy-efficient disk systems can be estimated by simulating the behaviors of energy-saving schemes. Unfortunately, there is widespread lack of fast and accurate methodology to evaluate reliabilities of modern parallel disk systems with high energy efficiency. To address this problem, we propose a mathematical modeling framework called MINT to quantitatively investigate the reliabilities of parallel disk systems employing a variety of energy conservation schemes. The MINT framework relies on data access patterns and reliability-affecting factors like disk utilization, temperature, and power-state transition frequency.

We used the MINT model to comprehensively study the reliability of parallel disks equipped with the two s energy-saving schemes, namely, the PDC technique [20] and the MAID [7]. We paid particular attention on PDC and MAID, because our focus is the power optimization strategies that adversely affect reliabilities of parallel disk systems. The MINT model suggests that the reliability of PDC is slightly higher than that of MAID under light workload. We also observe from MINT that MAID is noticeably more reliable than PDC with relatively high data-access rates.

2.2 Outline and Summary of Contributions

The contribution of this paper is 1) to present a new reliability modeling framework for energy-efficient parallel disk systems and 2) to improve energy-efficient parallel disks by alternating disks storing hot data with disks holding cold data. In particular, the main contributions of this study include are summarized as follows:

- We developed a generic mathematical approach—called MINT (Mathematical Reliability Models for Energy-efficient Parallel Disk System) [47]—to modeling reliability of energy-efficient parallel disks coupled with power management optimization policies.
- We built two reliability models for the two well-known energy-saving schemes—Popular Data Concentration scheme (PDC) and MAID.
- We analyzed intriguing the impacts of PDC and MAID on the annual failure rates of parallel disk systems. The analysis is made possible because we first investigated the impact of file access rates on disk utilization of MAID and PDC.
- We applied real-world traces [1] to drive our disk simulator, thereby validating our model that estimates disk utilizations using file access patterns.

3 The MINT Reliability Modeling Framework

3.1 Overview of the Framework

We start the modeling process by capturing the behaviors of parallel disk systems coupled with power management optimization policies. Let us first make use of data access patterns as input parameters, which are used to estimate each disk's utilization and power-state transition frequency. Then, we derive each disk's reliability in terms of annual failure rate from the disk's utilization, operating temperature as well as power-state transition frequency. These three parameters are key reliability-affecting factors in addition to disk ages. Finally, we calculate the reliability of the parallel disk system in accordance with the annual failure rate of each disk in the system.

Fig. 1 depicts the framework of the MINT reliability model for parallel disk systems with energy conservation schemes. MINT is composed of a single-disk reliability modeling module, a system-level reliability modeling module, and four reliability-affecting factors - disk age, temperature, disk state transition frequency (hereinafter referred to as frequency) and utilization. Many energy-saving schemes (e.g., PDC [20] and MAID [7]) inherently affect reliability-related factors like disk utilization and transition frequency. Given an energy optimization mechanism (e.g., PDC [20] and MAID [7]), MINT first transfers data access patterns into frequency and utilization - the two reliability-affecting factors. The single-disk reliability model can derive individual disk's annual failure rate from utilization, power-state transition frequency, age, and temperature. Reliabilities of all the disks in a parallel disk system are used as input to the system-level reliability modeling module that is responsible of estimating the annual failure rate of parallel disk systems.

There are several reliability-related factors, among which we consider four factors in MINT. It does not, however, necessarily imply that disk utilization, age, temperature, and power-state transitions are the only parameters affecting disk reliability. Other factors that may have impacts on disk reliability include handling, humidity, voltage variation, vintage, duty cycle, and altitude [9]. To incorporate a new factor into MINT, one simply needs to extend the single reliability modeling module described in Section 3.5. We pay attention to disks that are more than one year old, because the infant mortality phenomena is out the scope of this study. The reliability models presented in this paper are focused on read-intensive I/O activities, because a wide range of applications are read-intensive in nature. These applications include, but not limited to, web applications (e.g., Gmail and Facebook), video streaming servers (e.g., Youtube, Hulu), and search engines (e.g., Google and Bing).

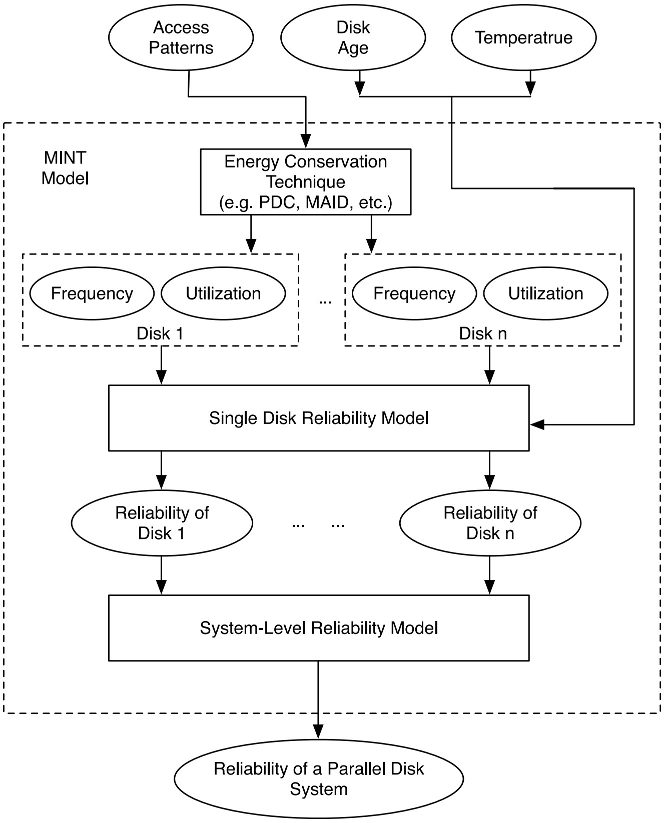

Access Patterns
Disk Age
Temperatrue
MINT Model
Energy Conservation Technique (e.g. PDC, MAID, etc.)
Frequency
Utilization
Disk 1
...
Frequency
Utilization
Disk n
Single Disk Reliability Model
Reliability of Disk 1
... ...
Reliability of Disk n
System-Level Reliability Model
Reliability of a Parallel Disk System

Fig. 1. The framework of the MINT reliability model.

3.2 Impacts of Utilization on Disk Failure Rate

Disk utilization can be characterized as the fraction of active time of a disk drive out of its total powered-on-time [23]. Different from RAID systems in which each disk only handles partial data of a file (see also the data stripping technique), each single disk in our disk array model stores entire files. Therefore, the access patterns of all disks in a disk array are unlikely to be identical. Thus, the disks in a MAID system or a PDC system tend to have various utilizations, thereby leading to different failure rates rather than an identical one. In our single disk reliability model, the impacts of disk utilization on reliability is good way of providing a baseline characterization of disk annual failure rate or AFR.

Using field failure data collected by Google, one can investigate the impact of utilization on AFR across the different age groups. For example, Pinheiro et al. studied AFR value of multiple disk groups with different ages, focusing

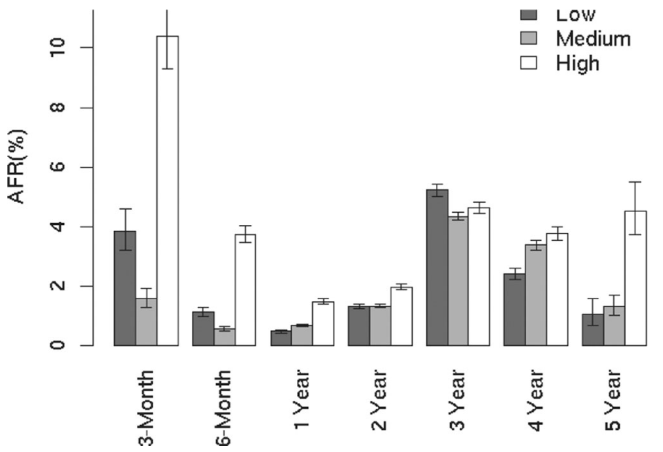


Fig. 2. Utilization affects disk annual failure rate (AFR) (Data by Google [23]).

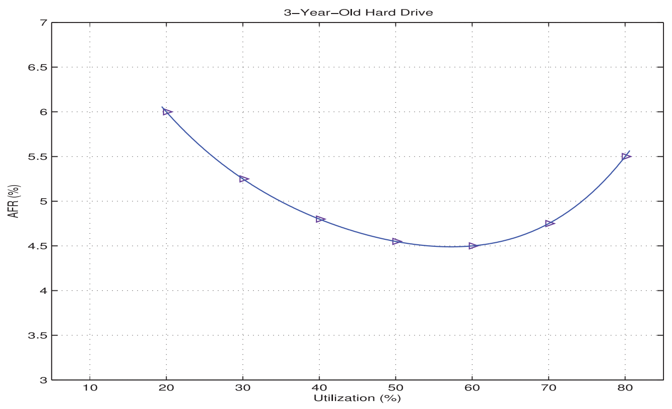


Fig. 3. Impacts of disk utilization on annual failure rate (AFR).

on the impact on disk utilization on AFR. Disk utilization are categorized by Pinheiro et al. in three levels—low, medium, and high. Fig. 2 shows AFRs of seven disk groups, representing disks whose ages are 3 months, 6 months, 1 year, 2 years, 3 years, 4 years, and 5 years under the three utilization levels.

Since the single-disk reliability model in MINT needs a baseline AFR value derived from disk utilization, we make use of the polynomial curve-fitting technique to extrapolate the baseline AFR a single disk from Google's field failure date. Extrapolating AFRs from the field failure data is important, because such an extrapolation approach allows us to estimate failure rate of a disk in accordance to the disk's utilization.

Our extrapolation procedure relies on the dataset (i.e., three data points for each disk group) provided by Google; this approach is reasonable because of the following justifications.

- First, compared with other disk failure datasets, the field failure dataset offered by Google is the most comprehensive assessment. The Google dataset can readily be incorporated into the utilization-to-failure module of our reliability modeling framework. We also obtain the maintenance data from the Los Alamos National Laboratory (LANL); the LANL data represents failure rates of a high-performance computing system. The Google data is more recent than the LANL data in the context of storage systems.
- Second, integrating Google's disk failure data into our model offers a general idea on how disk utilization can affect the lifetime of energy-efficient storage systems.
- Third, the failure dataset from Google represents significant trends of the relationship between failure rates and disk utilizations. For example, a pronounced trend is that disk-failure rate goes up when the disk utilization increases from medium to high.
- Fourth, the failure-rate trend of disks is changing when the disks are aging and; therefore, we incorporate disk ages as an important parameter into the disk failure-rate model.

In what follows, we give an example of a failure-rate model for three-year old disks. Failure-rate models of disks that are not three-year old (e.g., one-year, or five-year old) can be built in the same manner. The baseline value (i.e., R_{base} in (3) of AFR for a disk can be calculated from the disk's utilization. Fig. 2 shows disk utilization impacts on disk annual failure rate. Fig. 3 reveals the annual failure rate of three-year old disks as a function of disk utilization. The data in Fig. 2 is extrapolated from the disk-failure data offered by Google [23]. The curve plotted in Fig. 3 can be expressed as a utilization-reliability function described as (1) below:

$$F(u,3) = 4.167e^{-7}u^4 - 7.5e^{-5}u^3 + 5.968e^{-3}u^2 - 2.575e^{-1}u + 9.3, \tag{1}$$

where $F(u)$ represents the AFR value as a function of a certain disk utilization u. With (1) in place, one can readily derive annual failure rate of a disk if its age and utilization are given. For example, for a 3-year old disk with 50 percent utilization (i.e., $u = 50\%$), we can obtain the AFR value of this disk as $R(u) = 4.8\%$. Fig. 3 suggests that unlike the conclusions drawn in a previous study (see [35]), a low disk utilization does not necessarily lead to low AFR. For instance, given a three-year old disk, the AFR value under 30 percent utilization is even higher than AFR under 80 percent utilization.

3.3 Impacts of Temperature on Disk Annual Failure Rate

Temperature is often considered as the most important environmental factor affecting disk reliability. Field failure data of disks in a Google data center (see Fig. 4) shows that in most cases when temperatures are higher than 35°C, increasing temperatures lead to an increase in disk annual failure rates. On the other hand, Fig. 4 indicates that in the low and middle temperature ranges, the failure rates decreases when temperature increases [23].

Growing evidence shows that disk reliability should reflect disk drives operating under environmental conditions like temperature [9]. Since temperature (e.g., measured 1/2'' from the case) apparently affect disk reliability, the temperature can be considered as a multiplier (hereinafter referred to as temperature factor) to baseline failure rates where environmental factors are integrated [9]. Given a temperature, one must decide the

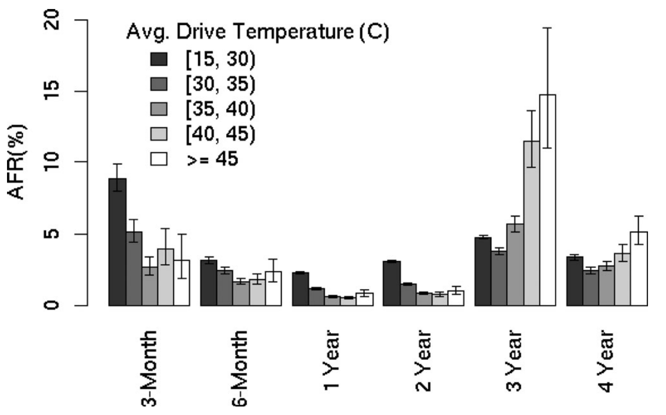


Fig. 4. Average drive temperature affects annual failure rate (AFR) (Data by Google [23]).

corresponding temperature factor (see τ in (3)) that can be multiplied to the base failure rates. Using Google's disk-failure data plotted in Fig. 4, we consider temperature factors under various temperature ranges for disks with different ages. More specifically, Fig. 4 shows annual failure rates of disks whose ages are from three-month to four-year old. For disk drives whose ages fall in each age range, we model the temperature factor as a function of drive temperature. Thus, six temperature-factor functions must be derived.

Now we explain how to determine a temperature factor for each temperature under each age range. Let us choose 25°C as the base temperature value, because room temperatures of data centers in many cases are set as 25°C controlled by cooling systems. Thus, the temperature factor is 1 when temperature is set to the base temperature (25°C). Let R_T denote the average value of AFR at temperature T. For example, R_{25} represents disk's AFR when the temperature is set to 25°C. We model the temperature factor τ as a ratio between R_T and R_{25} (i.e., R_T/R_{25}) under the condition that τ is larger than or equal to 25°C. When τ exceeds 45°C, the temperature factor becomes a constant (i.e., $\tau = 15/5 = 3$, see also Fig. 5). Note that Fig. 5 is derived from the Google's disk failure data. Due to space limit, we only show how temperature affects the temperature factor of a three-year old disk in Fig. 4. Note that the temperature-factor functions for disks in other age ranges can be modeled in a similar way. Fig. 5 shows the temperature-factor function derived from Fig. 4 for three-year old disks. We can observe from Fig. 4 that AFRs increase to 215 percent of the base value when the temperature is between 40 to 45°C.

3.4 Power-State Transition Frequency

To conserve energy in single disks, power management policies turn idle disks from the active state into standby. The disk power-state transition frequency (or frequency for short) is often measured as the number of power-state transitions (i.e., from active to standby or vice versa) per month. The reliability of an individual disk is affected by power-state transitions and; therefore, the increase in failure rate as a function of power-state transition frequency has to be added to a baseline failure rate (see (3) in Section 3.5). We define an increase in AFR due to power-state transitions as power-state transition frequency adder (frequency adder for short). The frequency adder is modeled by combining the disk spindle start/stop failure rate adders described by IDEMA [35] and the PRESS model [43]. Again, we focus on three-year old disk drives. Fig. 6 demonstrates frequency-adder values as a function of power-state transition frequency. Data plotted in Fig. 6 shows that high frequency leads to a high frequency adder to be added into the base AFR value. We used the quadratic curve fitting technique to model the frequency adder function (see (2)) plotted in Fig. 6.

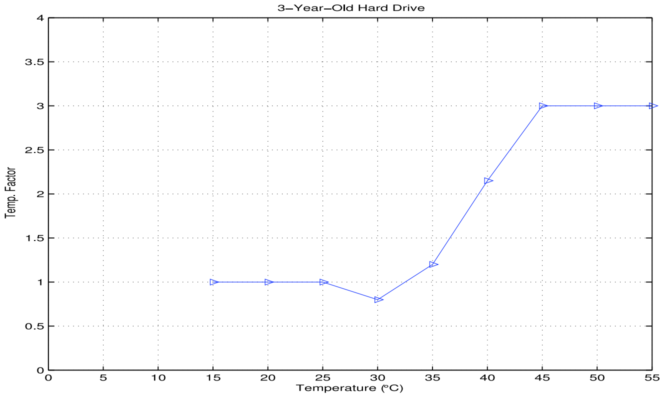


Fig. 5. Correlations between temperature and temperature factor.

$$F(v) = 1.51e^{-6}v^2 - 1.09e^{-5}v + 1.39e^{-2}, v \in [0, 500], \qquad (2)$$

where v is a power-state transition frequency, $F(v)$ represents an adder to the base AFR value. For example, when the transition frequency is 300 per month, the AFR value becomes 0.13 percent (see Fig. 6), which is higher than its corresponding baseline AFR value.

3.5 Single Disk Reliability Rate

Single-disk reliability can not be accurately described by one valued parameter, because the disk drive reliability is affected by multiple factors (see Sections 3.2, 3.3, and 3.4). Though recent studies attempted to consider multiple reliability factors (see, for example, PRESS [43]), few of prior studies investigated the details of combining the multiple reliability factors. We model the single-disk reliability in terms of annual failure rate (AFR) in the

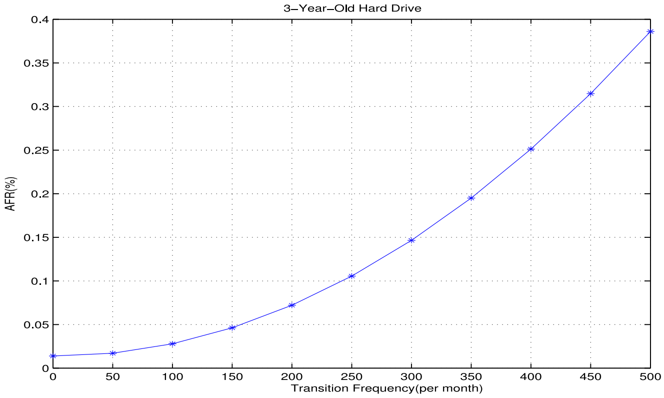


Fig. 6. Impacts of power-state transition frequency on frequency adder for three-year-old hard disks.

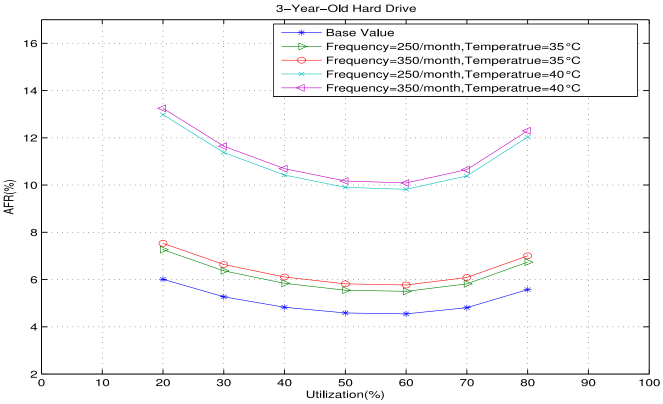


Fig. 7. Annual failure rate of a three-year-old hard disk as a function of disk utilization, power-state transition frequency, and temperature.

following three steps. We first compute a baseline AFR as a function of disk utilization. We then use temperature factor as a multiplier to the baseline AFR. Finally, we add a power-state transition frequency adder to the baseline value of AFR. Hence, the failure rate R of an individual disk can be expressed as

$$F = \alpha \times F_{base} \times \tau + \beta \times F_{freq}, \tag{3}$$

where F_{base} is the baseline failure rate derived from disk utilization (see Section 3.2), τ is the temperature factor (or temperature multiplier described in Section 3.3), F_{freq} is the power-state transition frequency adder to the base AFR (see Section 3.4), and α and β are two coefficients for the value of reliability F. If reliability F is more sensitive to frequency than to utilization and temperature, then β must be greater than α. Otherwise, β is smaller than α. In either case, α and β can be readily set in accordance with F's sensitivities to utilization, temperature, and frequency. In our experiments, we assume that all the three reliability-related factors are equally important (i.e., $\alpha = \beta = 1$). Ideally, extensive field tests allow us to analyze and test the two coefficients.

Calculating the correlation coefficients α and β is out the scope of this study. Nevertheless, we may apply the linear regression analysis to compute the two coefficients from field test samples under various conditions. Equation (3) represents an appropriate reliability model because of the following four rationales. First, existing studies show that both power-state transition frequency and temperatures have a negative influence on the reliability of storage systems, indicating that α and β are greater than or equal to 1. Second, although the values of α and β do affects the accuracy of the reliability model, α and β will not change general trends described by (3). Thus, setting α and β to 1 makes (3) demonstrate the general trends of the reliability model. Third, (3) illustrates a consistent relationship between annual failure rate and the three major factors (i.e., disk utilization, power-state transition frequency and temperature presented in Section 4.1). Last, our approach is a heuristic way to model reliability of energy-efficient storage systems. Depending upon the availability of field test data offered by disk vendors, in near future we will adjust α and β to improve the quality of the proposed reliability model.

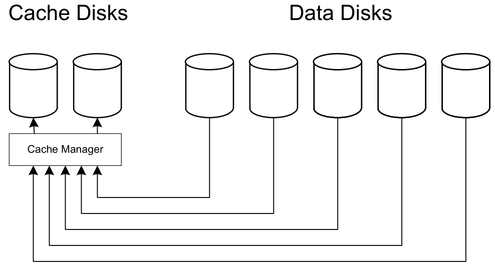


Fig. 8. The MAID system structure.

Equation (3) enables us to analyze a disk's reliability in turns of annual failure rate (AFR). Fig. 7 shows the AFR values of a three-year-old disk when its utilization is in the range between 20 and 80 percent, the power-state transition frequency is 250 and 350 no./month, and temperature is 35 and 40°C. We observe from Fig. 7 that unlike temperature, the frequency adder in (3) has marginal impact on disk failure rate, because the transition frequency is relatively low (e.g., 300 per month; about 10 times per day). It is expected that when transition frequency is extremely high, AFR will become more sensitive to frequency than to temperature. A second reason is that the β value in (3) equals to the α value; however, in real-world scenarios, β might be greater than that α. Furthermore, a data center's cooling systems can keep the space at a desired low temperature (e.g., 25 to 35°C); hence, the value τ is kept less than 1.5 or even less than 1. In our future study, we will further investigate temperature impacts [14] on AFR of storage systems.

4 Reliability Models for MAID and PDC

In this section, we develop two reliability models for MAID and PDC, which are two well-designed energy-efficient storage systems. System reliability is derived from disk utilization and power-state transition frequency. We show how to estimate the disk utilization (see Sections 4.1.2 and 4.2.2) and the power-state transition frequency (see Sections 4.1.3 and 4.2.3). Finally, we show how to use these two models to evaluate the impact of file access rates on annual failure rate of MAID and PDC (see Section 4.4).

4.1 Modeling Reliability of MAID

4.1.1 Overview of MAID

The Massive Arrays of Idle Disks technique or MAID developed by Colarelli and Grunwald aims to reduce energy consumption of large disk arrays while maintaining acceptable I/O performance [7]. MAID relies on data temporal locality to place replicas of active files on a subset of cache disks, thereby allowing other disks to spin down. Fig. 8 shows that MAID maintains two types of disks—cache disks and data disks. Popular files are copied from data disks into cache disks, where the LRU policy is implemented to manage data replacement in cache disks. Replaced data is discarded by a cache disk if the data is clean; dirty data has to be written back to the corresponding data disk. Note that the overhead of writing dirty data back to corresponding data disks has impacts on both energy efficiency and reliability of cache and data disks. For simplicity, we ignore the impact of LRU on failure rates of disks in the MAID system.

In other words, we focus on reliability of the MAID system dealing with read-only requests issued from read-intensive applications. We emphasize the impact of reads on the reliability of MAID, because there exists a wide range of read-intensive applications such as decision support systems, data mining, and web servers [1], [25], [38]. In our future work (see Section 7, we plan to extend our models to study the impact of writes on the reliability of MAID.

To prevent cache disk from being overly loaded, MAID avoids copying data to cache disks that have reached their maximum bandwidth. Three components integrated in the MAID model include: 1) power management policy, by using which drives that have not seen any requests for a specified period are spun down to sleep, or an adaptive spin-down to active; 2) data layout, which is either linear, with successive blocks being placed on the same drive, or striped across multiple drives; and 3) cache, which indicates the number of drives of the array which will be used for cache [7].

4.1.2 Modeling Disk Utilization in MAID

Recall that the annual failure rate of each disk can be calculated using disk age, utilization, operating temperature as well as power-state transition frequency. To model reliability of a disk array equipped with MAID, we have to first address the issue of modeling disk utilization used to calculate base annual failure rates. In this subsection, we develop a utilization model capturing behaviors of a MAID-based disk array. The utilization model takes file access patterns as an input and calculates the utilization of each disk in the disk array.

Disk utilization is computed as the fraction of active time of a disk drive out of its total powered-on-time. Now we describe a generic way of modeling the utilization of a disk drive. Let us consider a sequence of I/O accesses with N I/O phases. We denote T_i as the length or duration of the ith I/O phase. Without loss of generality, we assume that a file access pattern in an I/O phase remains unchanged. The file access pattern, however, may vary in different phases. The relative length or weight of the ith phase is expressed as $W_i = T_i/T$ where $T = \sum_{i=1}^{N} T_i$ is the total length of all the I/O phases. Suppose the utilization of a disk in the ith phase is ρ_i, we can write the overall utilization ρ of the disk as the weighted sum of the utilization in all the I/O phases. Thus, we have

$$\rho = \sum_{i=1}^{N} (W_i \times \rho_i) = \sum_{i=1}^{N} \left(\frac{T_i}{T} \times \rho_i\right). \tag{4}$$

Let $F_i = (f_{i1}, f_{i2}, \dots, f_{in_i})$ be a set of n_i files residing in the disk in the ith phase. The utilization ρ_i (see (4)) of the disk in phase i is contributed by I/O accesses to each file in set F_i. Thus, ρ_i in (4) can be written as

$$\rho_i = \sum_{j=1}^{n_i} (\lambda_{ij} \times s_{ij}), \tag{5}$$

where λ_{ij} is the file access rate of file f_{ij} in F_i and s_{ij} is the mean service time of file f_{ij}. Note that I/O accesses to each file in set F_i are modeled as a Poisson process; file access rate and service time in each phase i are given a priori. We assume that there are n hard drives with k phases. In the lth phase, let f_{ijl} be the jth file on the ith disk, where $i \in (1, 2, \dots, n)$, $j \in (1, 2, \dots, m_i)$, $l \in (1, 2, \dots k)$. We have:

$$\begin{aligned}
\overline{f}_{1l} &= \left\{f_{11l}, f_{12l}, \dots, f_{1m_1l}\right\},\\
\overline{f}_{2l} &= \left\{f_{21l}, f_{22l}, \dots, f_{2m_2l}\right\},\\
&\vdots\\
\overline{f}_{nl} &= \{f_{n1l}, f_{n2l}, \dots f_{nm_nl}\},
\end{aligned} \tag{6}$$

where m_i is the number of files on the ith disk and $\overline{f}_{il}$ is the total files on the same disk. Since frequently accessed files are duplicated to cache disks, we model below an updated file placement after copying the frequently accessed files.

$$\begin{aligned}
\overline{f}'_{1l} &= \{f'_{11l}, f'_{12l}, \dots, f'_{1m'_1l}\},\\
\overline{f}'_{2l} &= \left\{f'_{21l}, f'_{22l}, \dots, f'_{2m'_2l}\right\},\\
&\vdots\\
\overline{f}'_{nl} &= \left\{f'_{n1l}, f'_{n2l}, \dots, f'_{nm'_nl}\right\},
\end{aligned} \tag{7}$$

where m'_i is the number of the files on the ith disk, f'_{ijl} is the jth file at the lth phase and $\overline{f}'_{il}$ is the set of files on the same disk after the files are copied. We can calculate the utilization for jth file in the lth phase on the ith disk as

$$\rho_{ijl} = \lambda_{ijl} \times s_{ijl}. \tag{8}$$

We assume that $\rho_{i1l} \geq \rho_{i2l} \geq \cdots \geq \rho_{im_1l}$, meaning that files are placed in a descending order of utilization. After the frequently accessed files are copied to the cache disks, we denote the updated utilization contributed by files including copied ones as $\rho'_{i1l} \geq \rho'_{i2l} \geq \cdots \geq \rho'_{im_1l}$. It is intuitive that the utilization of disk i should be smaller than 1. When a disk reaches its maximum utilization, the disk also reaches its maximum bandwidth denoted as B_i. Regardless of cache disks or data disks, we express the utilization for ith disk in phase l as

$$\begin{aligned}
\rho'_{il} &= \frac{T_{IO} + T_{copy}}{T}\\
&= \sum_{j \in F'_{il}} \lambda_{ijl} \cdot s_{ijl} + \frac{T_{copy}}{T}\\
&= \sum_{j=1}^{m'_i} \rho'_{ijl} + \frac{T_{copy}}{T},
\end{aligned} \tag{9}$$

where T is the time interval of the lth I/O phase and T_{copy} is the time spent in moving popular data from data disks into cache disks. The first and second items on the bottom-line on the right-hand side of (9) are the utilizations caused by accessing files and duplicating files from data disks to cache disks, respectively.

Since files on cache disks are duplicated from data disks, frequently accessed files must be copied from data disks and written down to cache disks. As such, we must consider disk utilization incurred by the data duplication process. To quantify utilization overhead caused by data replicas, we define a set $F_{il}^{M\_out}$ of files copied from the ith data disk to cache disks in phase l. Similarly, we define a set $F_{il}^{M\_in}$ of files copied to the ith cache disk from data disks in phase l.

At the lth phase, the file f_{ijl}, which is the jth file on the ith disk will be copied out of the data disk if its utilization ρ_{ijl} is higher than ρ_{i1l}, which is the highest utilization held by file f_{i1l}. For all files to be copied out on the ith disk can be expressed as

$$f_{il}^{M\_out} = \{\forall 1 \leq j \leq m_i', \exists \rho_{ijl}' > \rho_{i1l}\}. \tag{10}$$

With respect to the ith data disk, the utilization $\rho_{il-data}'$ in phase l is the sum of utilization caused by accessing files on the data disk and reading files to be duplicated to cache disks. Thus, $\rho_{il-data}'$ can be written as

$$\rho_{il-data}' = \sum_{j=1}^{m_i'} \rho_{ijl}' + \frac{\sum_{j \in F_{il}^{M\_out}} t_{ijl}}{T}, \tag{11}$$

where the first and second items on the right-hand side of (11) are the utilizations of accessing files and reading files from the data disk to make replicas on cache disks, respectively.

When it comes to the ith cache disk, the utilization $\rho_{il-cache}'$ in phase l is the sum of utilization contributed by accessed files and written file replicas to cache disks. Thus, $\rho_{il-data}'$ can be written as

$$\rho_{il-cache}' = \sum_{j=1}^{m_i'} \rho_{ijl}' + \frac{\sum_{j \in F_{il}^{M\_in}} t_{ijl}}{T}, \tag{12}$$

where the first and second items on the right-hand side of (12) are the utilizations of accessing files and writing files to the cache disk to make replicas, respectively.

4.1.3 Modeling Power-State Transition Frequency for MAID

Equation 3 in Section 3.5 shows that the power-state transition frequency adder is an important factor to model disk annual failure rate. The number of power-state transitions largely depends on I/O workload conditions in addition to the behaviors of MAID. In this subsection, we derive the number of power-state transitions from file access patterns.

We define the T_{BE} as the disk break-even time - the minimum idle time required to compensate the cost of entering the disk standby mode (T_{BE} values are usually anywhere between 10 to 15 seconds). Given file access patterns of the ith phase for a disk, we need to calculate the number v_i of idle periods that are larger than the break-even time T_{BE}. The number of power-state transitions during phase i is $2v_i$, because there is a spin-down at the beginning of each large idle time and a spin-up by the end of the idle time. For an access pattern with N I/O phases, the total number of power-state transitions v (see (2)) is expressed as

$$v = 2 \times \sum_{i=1}^{N} v_i. \tag{13}$$

We model a workload condition where I/O burstiness can be leveraged by the dynamic power management policy to turn idle disks into the standby mode to save energy. To model I/O burstiness, we assume the first I/O requests of files within an access phase are arriving in a short period of time, within which disks are too busy to be switched into standby. After the period of high I/O load, there is an increasing number of opportunities to place disks into the standby mode. This workload model allows MAID to achieve high energy efficiency at the cost of disk reliability, because the workload model leads to a large number of power-state transitions.

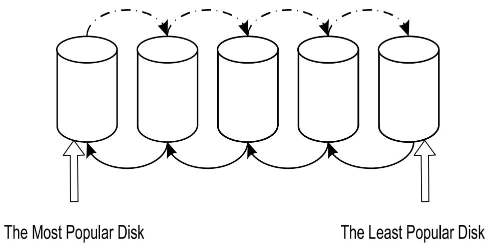


Fig. 9. The PDC system structure.

To conduct a stress test on reliability of MAID, we assume that the first requests of files on a disk arrive at the same time. For the first few time units, the workloads are so high that no data disks can be turned into standby. As the I/O load is descreasing, some data disks may be switched to standby when idle time intervals are larger than T_BE. In this workload model, MAID can achieve the best energy efficiency with the worst reliability in terms of the number of power-state transitions.

Let us model power-state transition frequency for a single disk. Assume that the h_ith file (h_i is placed in an descending order from m_i, $h_i \in (m_i, m_{i-1}, \ldots, 2, 1)$) on the ith disk at the lth phase is the dividing point of the request access rate. Given a file whose index is larger than h_i, the file's access rate λ_{ijl} enables the disk to be turned in standby because idle times are larger than break-even time T_BE. Hence, we have

$$\begin{cases} \dfrac{1}{\lambda_{i1}} - \displaystyle\sum_{j=1}^{h_i} \dfrac{1}{\lambda_{ijl}} > T_{BE} \\ \dfrac{1}{\lambda_{i1}} - \displaystyle\sum_{j=1}^{h_i-1} \dfrac{1}{\lambda_{ijl}} < T_{BE}, \end{cases} \tag{14}$$

where $\frac{1}{\lambda_{i1}}$ is the interval time of the first file on the ith disk. The number of the spin ups/downs is $2(T \cdot \lambda_{i1l} - T \cdot \lambda_{ih_il})$. Cache disks in MAID are very unlikely to be switched into the standby mode. Since cache disks store replicas, any failure in the cache disks has no impact on data loss. Only failures in data disks affect the reliability of disk arrays.

4.2 Modeling Reliability of PDC

4.2.1 Overview of PDC

The Popular Data Concentration technique or PDC proposed by Pinheiro and Bianchini migrates frequently accessed data to a subset of disks in a disk array [20]. Fig. 9 demonstrates the basic idea behind PDC: the most popular files are stored in the far left disk, while the least popular files are stored in the far right disk. PDC can rely on file popularity and migration to conserve energy in disk arrays, because several network servers exhibit I/O loads with highly skewed data access patterns. The migrations of popular files to a subset of disks can skew disk I/O load towards this subset, offering other disk more opportunities to be switched to standby to conserve energy. To void performance degradation of disks storing popular data, PDC aims at migrating data onto a disk until its load is approaching the maximum bandwidth of the disk.

The main difference between MAID and PDC is that MAID makes data replicas on cache disks, whereas PDC lays data out across disk arrays without generating any replicas. If one of the cache disks fails in MAID, files residing in the failed cache disks can be found in the corresponding data disks. In contrast, any failed disk in PDC can inevitably lead to data loss. Although PDC tends to have lower reliability than MAID, PDC does not need to trade disk capacity for improved energy efficiency and I/O performance.

4.2.2 Modeling Disk Utilization in PDC

Since frequently accessed files are periodically migrated to a subset of disks in a disk array, we have to take into account disk utilization incurred by file migrations. Hence, the ith disk's utilization ρ'_{il} during phase l is computed as the sum of the utilization contributed by accessing files residing in disk i and the utilization introduced by migrating files to/from disk i. Thus, we can express utilization ρ'_{il} as:

$$\begin{aligned}\rho'_{il} &= \frac{T_{IO} + T_{migration}}{T} \\ &= \sum_{j\in F'_{il}} \lambda_{ijl} \cdot s_{ijl} + \frac{T_{migration}}{T} \\ &= \sum_{j=1}^{m'_i} \rho'_{ijl} + \frac{T_{migration}}{T}\end{aligned}$$

where T is the time interval of I/O phase l and $T_{migration}$ is the time spent in migrating popular data to hot disks and migrating non-popular data to cold disk. The first and second items on the bottom-line on the right-hand side of (15) are the utilizations caused by accessing files and duplicating files from data disks to cache disks, respectively.

To quantify utilization introduced by the file migration process (see the second item on the bottom-line on the right-hand side of (15)), we define two set of files for the ith disk in the lth I/O phase. The first set $F_{il}^{M\_out}$ contains all the files migrated from disk i to other disks during the lth phase. Similarly, the second set $F_{il}^{M\_in}$ consists of files migrated from other disks to disk i in phase l.

Now we can formally express the utilization of disk i in phase l using the two file sets $F_{il}^{M\_out}$ and $F_{il}^{M\_in}$. Thus, we have:

$$\rho'_{il} = \sum_{j=1}^{m'_i} \rho'_{ijl} + \frac{\sum_{j\in F_{il}^{M}} s_{ijl}}{T_l} \tag{16}$$

where the second item on the right-hand side of (16) is the utilization incurred by 1) migrating files in set $f_{il}^{M\_out}$ from disk i to other disks and 2) migrating files in set $f_{il}^{M\_in}$ from other disks to disk i during phase l. We agree with the reviewer. The reworded Theorem has been included to the revised paper.

TABLE 1
Configurations of the MAID and PDC Systems

Energy-saving Scheme	Number of Disks	File Access Rate (No. per month)	File Size (KB)
PDC	20 data (20 in total)	0˜10^6	300
MAID-1	15 data+5 cache (20 in total)	0˜10^6	300
MAID-2	20 data+5 cache (25 in total)	0˜10^6	300

4.2.3 Modeling Power-State Transition Frequency for PDC

We used the same way described in Section 4.1.3 to model power-state transition frequency for PDC. The transition frequency plays a critical role in modeling reliability, because the frequency adder to base AFRs is derived from the transition frequency measured in terms of the number of spin-ups and spin-downs per month. Unlike MAID, PDC allows each disk to receive migrated data from other disks. In light of PDC, disks storing the most popular files are most likely to be kept in the active mode.

4.3 Reliability of PDC and MAID

Let $R_i(t)$ be the reliability of a single disk system. $R_i(t)$ is the probability that disk i functions at time $t > 0$. Statistical analysis of failed disks shows that reliability $R_i(t)$ can be approximated by the following expression:

$$R_i(t) = e^{-F_i * t}, \tag{17}$$

where F_i is the failure rate of disk i. Thus, the mean time to disk failure is $1/F_i$ [37]. Note that F_i can be derived from (3).

Given reliability $R_i(t)$ of the ith disk in the PDC system, we can derive the reliability $R_{PDC}(t)$ of PDC as a product of the reliability of all the N disks in the system:

$$R_{PDC}(t) = \prod_{t=1}^{N} R_i(t). \tag{18}$$

Let C be the number of cache disks in the MAID system. The reliability $R_{PDC}(t)$ of MAID can be expressed as a product of the reliability of all the $N - C$ data disks. Thus, we have:

$$R_{MAID}(t) = \prod_{t=1}^{N-C} R_i(t). \tag{19}$$

4.4 Reliability Evaluation

4.4.1 Experimental Setup

We developed a simulator in which the two reliability models for MAID and PDC were implemented. Table 1 shows the configuration parameters of MAID and PDC. Let us evaluate the reliability of the MAID system with two different configurations. In the first configuration, existing data disks are used as cache disks. For example, in the first MAID system (see MAID-1 in Table 1), there are five cache disks and 20 data disks. In the second configuration, extra cache disks are added to a disk array to cache frequently

TABLE 2
West Digital HDD Property for Simulator

Capacity	500 GB
Cach Size	SATA 16MB
Type	WD 5000AAKS-00V1A0
Buffer to Host Transfer Rate	300MB/s) (MAX)
Host to Host Transfer Rate (tested manually)	83MB/s

accessed data. For example, in the second MAID system, (see MAID-2 in Table 1), there are five cache disks and 15 data disks. For the case of PDC, we set the number of disks to 20. Thus, the MAID-1 and PDC systems contains the 20 disks; whereas the MAID-2 system consists of five extra cache disks in addition to 20 data disks (i.e., 25 disks in total). The file access rate is in the range from 0 to $5 * 10^5$ no./month, which represent a wide range of read-intensive applications. The operating temperature is set to 35°C. The simulated disks are West Digital hard drives, the parameters of which can be found in Table 2.

4.4.2 Disk Utilization

We first investigate the impacts of file access rate on utilization of MAID and PDC. Fig. 10 shows that when the average file access rate increases, the utilizations of PDC, MAID-1, and MAID-2 increase accordingly. Compared with the utilizations of MAID-1 and MAID-2, the utilization of PDC is a whole lot more sensitive to the file access rate.

The utilization of PDC is significantly higher than those of MAID-1 and MAID-2. For example, when the average file access rate is $5 * 10^5$ no./month, the utilizations of PDC, MAID-1, and MAID-2 are approaching to 90, 48, and 40 percent, respectively. PDC has high utilization, because disks in PDC spend noticeable amount of time in migrating data among the disks. Increasing the file access rate leads to an increase in the number of migrated files among the disks, thereby giving rise to an increased utilization due to file migrations. Unlike PDC, MAID simply needs to make replicas on cache disks without migrating the replicas back from the cache disks to data disks.

Under low I/O load levels, the utilizations of MAID-1 and MAID-2 are very close to each other. When I/O load becomes relatively high, the utilization of MAID-1 is slightly higher than that of MAID-2. This is mainly because the capacity of MAID-2 is larger than that of MAID-1.

4.4.3 Annual Failure Rate

Fig. 11 illustrates the AFRs) of MAID-1, MAID-2, and PDC. Results plotted in Fig. 11 show that AFR of PDC keeps increasing from 5.6 to 8.3 percent when the file access rate is larger than 150. We attribute this trend to high disk utilization due to data migrations. More interestingly, if the file access rate is lower than $15 * 10^4$, AFR of PDC slightly reduces from 5.9 to 5.6 percent when the access rate is increased from $5 * 10^4$ to $15 * 10^4$ no./month. This result can be explained by the nature of the utilization function that is concave rather than linear. The concave nature of the utilization function is consistent with the empirical results reported in [23]. When the file access rate is $15 * 10^4$, the disk utilization is approximately 50 percent, which is the turning point of the utilization function.

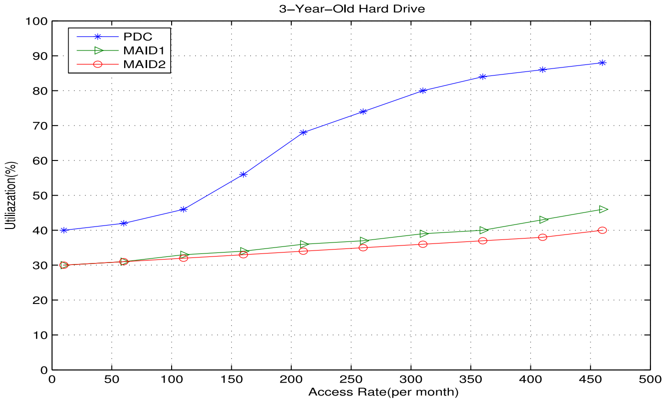


Fig. 10. Impacts of file access rate on disk utilization. Access rate varies from $1 * 10^4$ to $5 * 10^5$ no./month.

Unlike PDC's AFR, the AFRs of MAID-1 and MAID-2 continue decreasing from 6.3 to 5.8 percent with the increasing file access rate. This declining trend might be explained by two reasons. First, increasing the file access rates reduces the number of power-state transitions. Second, the range of the disk utilization is close to 40 percent, which is in the declining part of the curve.

4.4.4 Heavy I/O Load

Now we conduct a stress test by considering heavy I/O load levels. Let us further increase the average file access rate from $5 * 10^5$ to $1 * 10^6$ no./month. Interestingly, we observe from experimental results that the utilization trends and AFR trends of PDC and MAID plotted in Figs. 12 and 13 are different from the trends of utilization and AFR plotted in Figs. 10 and 11. More specifically, Fig. 12 shows that the utilization of PDC is saturated at the level of 90 percent after the average file access rate exceeds $5 * 10^5$ no./month. In contrast, the utilizations of both MAID-1 and MAID2 continue increasing with an increase in the average file access rate. Regardless of the value of access rate, the growing trends in utilization are consistantly true for MAID-1 and MAID-2 even when the access rate is larger than $5 * 10^5$ no./month. The PDC's utilization almost becomes a constant when the

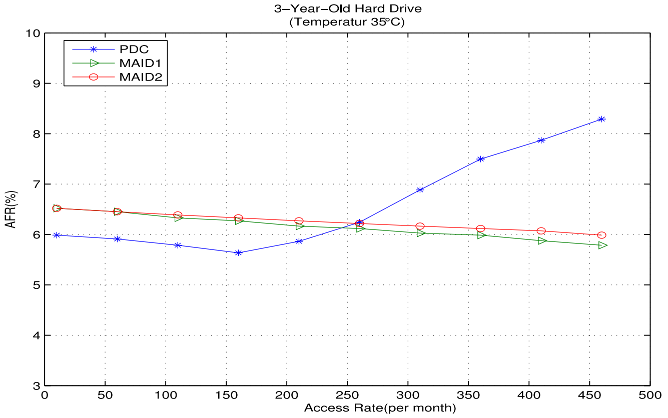


Fig. 11. Impacts of file access rate on annual failure rate (AFR) of PDC, MAID-1, and MAID-2. Access rate varies from $1 * 10^4$ to $50 * 10^5$ no./month.

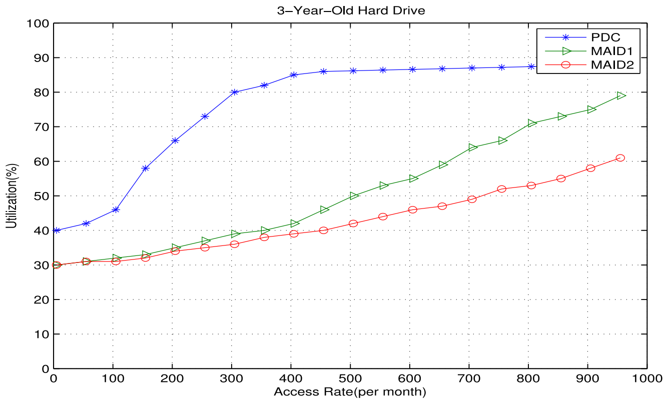


Fig. 12. Impacts of file access rate on disk utilization under heavy I/O loads. Access rate varies from $1 * 10^4$ to $1 * 10^6$ no./month.

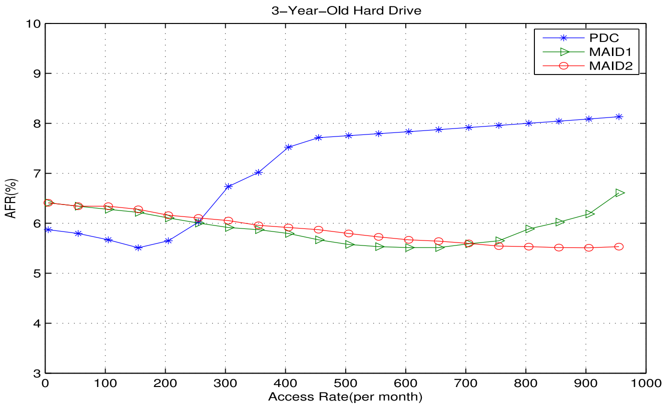


Fig. 13. Annual failure rate (AFR) of PDC, MAID-1/MAID-2 under heavy I/O loads. Access rate varies from 10 to $1 * 10^6$ no./month.

access rate is larger than $5 * 10^5$ no./month, because the high access rate can easily saturate the bandwidth of disks storing popular files. After the hot disks are saturated, the utilization caused by file migrations is significantly decreasing as the number of file migrations drops. When the access rate approaches $1 * 10^6$, the utilization of MAID-1 becomes significantly higher than that of MAID-2. The utilization of MAID-1 grows faster than that of MAID-2, because MAID-2 has more data disks than MAID-1.

Figs. 13 shows the AFRs of PDC and MAID-1/MAID-2 as functions of file access rate when temperature is set to 35°C. The first observation drawn drawn from Figs. 13 is that the AFR value of PDC only slightly increases after the average file access rate exceeds 500 per month. PDC's AFR grows very slowly under high I/O load because the utilizations of disks in PDC are almost saturated at the level of 90 percent when the access rate is higher than 500. More importantly, the second observation concluded from Figs. 13 is that when the access rate is higher than $7 * 10^5$ no./month, AFR of MAID-1 is higher than that of MAID-2. The main reason is that MAID-2 has five more data disks compare to MAID-1 which ensures that the utilization MAID-2 will be lower than MAID-1. The utiliaztion of MAID-1 keeps rising up over 60 percent (see Fig. 12) when access rate is higher than $7 * 10^5$ no./month while at the same time the utilization of MAID-2 is around 50 percent which lies in the lowest point of the AFR-Utilization curve shown in Fig. 3.

5 Model Validation

5.1 The Validation Techniques

It is reasonable to use MINT to compare the reliability performance of different energy-efficient storage systems, because the reliability models of the MAID and PDC storage systems use the same experimental data. It is challenging to validate the accuracy of the MINT modeling framework, since we are unable to watch MAID and PDC running for a couple of decades. One way to address this problem is to maintain and monitor a large number of MAID and PDC systems for a short period of time (e.g., 5 to 10 years). If one can watch the MAID and PDC systems over their entire service life, failure-rate data will be collected to validate reliability models. Even if we can test MAID and PDC with 100 disks for five years, the sample size is still considered small.

To address this validation problem, we verify MINT using the combination of the following two validation techniques [28], which are practical approaches to verification and validation of models.

- *Event Validity.* Events of occurrences of the model are compared to those of the real storage system to determine if they are similar. For example, in our validation process, we compared the file access rates in a real-world file system.
- *Historical Data Validation.* We first used part of the historical file access data (i.e., file I/O traces) for building our models. Then, we relied on the remaining data to test the models.

Recall that MINT consists of two major components—the utilization model (see Sections 4.1.2 and 4.2.2) and the failure-rate model. The utilization model estimates disk utilization of the MAID and PDC systems based on I/O access rates. The failure-rate model relies on real world failure data (see [23]) to predict the failure rate of a disk from its utilization.

To validate MINT, we need to validate the utilization model and the failure-rate model. The failure-rate model is derived from Google's dataset [23]; the justifications of this approach can be found in Sections 3.2 and 3.5. We used the existing failure-rate data obtained from a recent study [23]; the data have been validated and are accurate. The failure-rate model is a plugin of the MINT framework. Any future improved failure-rate model can be readily incorporated into MINT to boost accuracy of MINT. In this section, we pay particular attention to the validation of the utilization model.

We performed the following six steps repeatedly to validate the utilization model described in Sections 4.1.2 and 4.2.2.

- *Step 1.* We made use of the real-world I/O trace (i.e., Berkeley web trace) to derive file access rates.
- *Step 2.* The file access rates are applied to our utilization model to estimate disk utilizations of the MAID and PDC storage systems.

TABLE 3
File Access Rates of the One-Month Web Trace

File Access Rate Interval (No./Month)	The number of files
0 ∼10	185383
10 ∼10^2	112203
10^2 ∼10^3	4539
10^3 ∼10^4	244
10^4 ∼10^5	113
10^5 ∼10^6	33
10^6 ∼10^7	4

- *Step 3.* We implemented a trace replay tool, which captures the rapid evolution of web server workloads.
- *Step 4.* We developed the simple MAID and PDC systems that handle I/O requests created by the trace reply tool.
- *Step 5.* The utilizations of disks in the MAID and PDC storage systems are measured.
- *Step 6.* We compare the measured disk utilizations from the two real storage systems (see Step 5) with the disk utilizations derived from our models (see Step 2).

5.2 Berkeley Web Trace Replay

The Berkeley Web Trace [1] used in the model validation procedure was collected from a web server for an online library project from 22 January to 23 February, 1997. The Berkeley Web Trace data represents intensive I/O activities of a real-world system, for which MAID and PDC can conserve energy. Because I/O access rates in this study are measured in term of number I/O per/month or no./month, we decided to replay a one-month trace containing 33 trace files and 25,205,132 I/O requests. Among all the requests, 24,481,520 are file accesses requesting 302,519 web files. The trace replay period is 1,631,753 seconds or 453.3 hours.

Before applying file access rates into the utilization models presented in Sections 4.1.2 and 4.2.2, we performed an analysis on file access rates of the web traces. The goal of this analysis is to determine the access rate of each web file accessed over the one month period. Table 3 summarizes the distribution of file access rates of the 12,304,467 web files recorded in the 33 traces. Table 3 indicates that a vast majority (i.e., more than 61 percent) of web files were accessed less than ten times within a month. However, there are a few web files that were accessed for more than 1,000 times over a one-month period. The analysis result shows that the highest file access rate is 3180697 no./month.

Fig. 14 shows the files accesses distribution pattern using a bar chart. The distribution pattern suggests that when the access rate increases, the number of files that have such access rate decreases dramatically.

5.3 Experimental Result

Since the Utilization-AFR model, which transfers the utilization of systems to reliability, is employing the same date from the validated Google report, we only show the validation of Access Rate-Utilization model in this subsection.

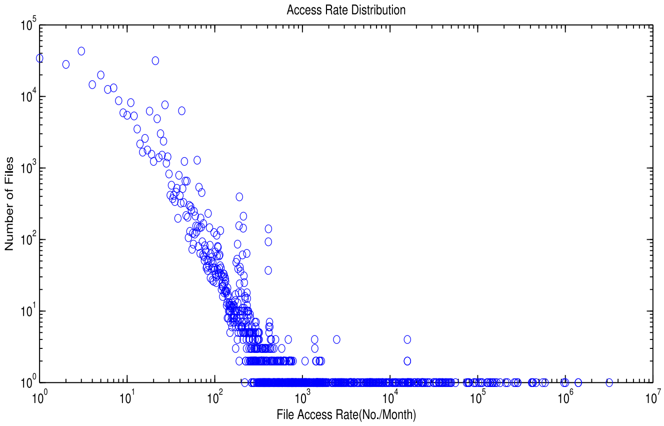


Fig. 14. The file access rate distribution of the one-month Berkeley web trace. Access Rate ranges from 1 to $4.5 * 10^4$ no./month

Fig. 15 indicates the utilization comparison between the MINT model and Berkeley Web Trace-driven simulation. In order to make a clearer comparison between the MINT model and the trace-driven simulation, we divided the utilization comparison of PDC, MAID-1 and MAID-2 separately (as shown in Figs. 16, 17 and 18). The results plotted in the figures confirm that the utilization curves obtained from the MINT model and the simulator follow a similar trend. The difference between the simulation results and modeling data is around 10 percent. The simulated utilization is slightly higher than that generated by the model, because the simulator doesn't take into account disk rotational delay and seek time.

After validating the Access Rate-Utilization sub-model, we further present the comparison results of Access Rate-AFR between the MINT model and the simulation. We are able to build up a Utilization-AFR sub-model of our own and insert it to our MINT model. However, due to the lack of maintenance date recently, how to validate the sub-model becomes a hard issue to deal with. Instead, we are using the validated data published by Google [23] in this part. Once we get more updated data in the future, such sub-model could be re-modified.

Figs. 19, 20, and 21 show the impacts of file access rate on AFR. Even thought the trends of Access Rate-Utilization sub-model appeared similar between the model and the simulation (as shown in Figs. 16, 17 and 18), there are noticeable differences between them when we discussed

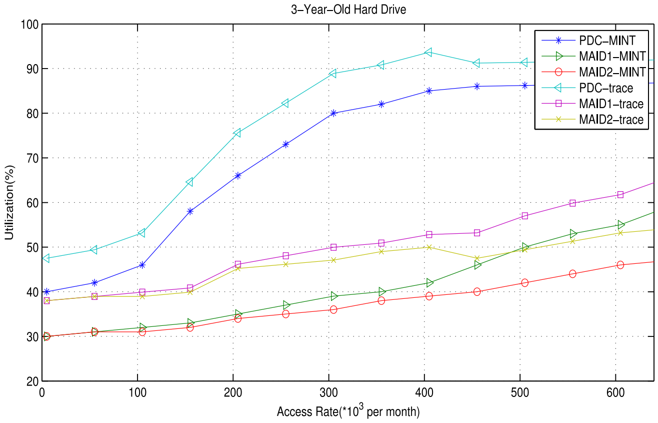


Fig. 15. Impacts of file access rate on disk utilization. Access rate varies from 10 to $64 * 10^4$ no./month.

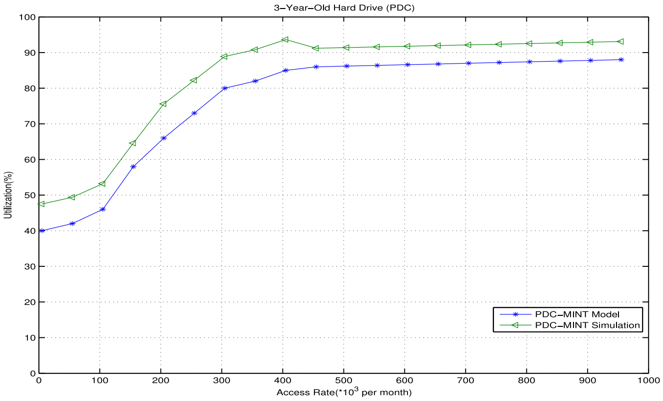


Fig. 16. Impacts of file access rate on disk utilization (PDC). Access rate varies from 10 to $64 * 10^4$ no./month.

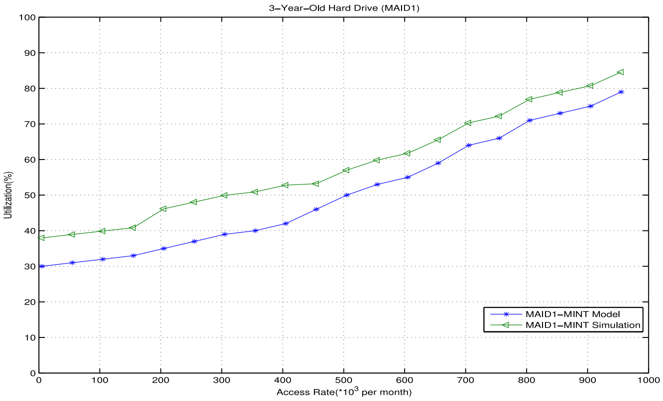


Fig. 17. Impacts of file access rate on disk utilization (MAID1). Access rate varies from 10 to $64 * 10^4$ no./month.

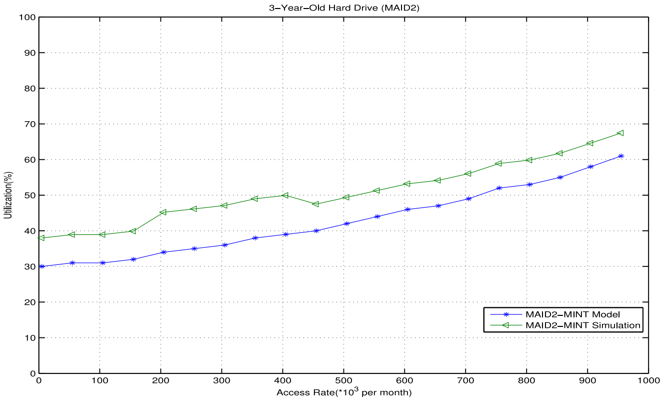


Fig. 18. Impacts of file access rate on disk utilization (MAID2). Access rate varies from 10 to $64 * 10^4$ no./month.

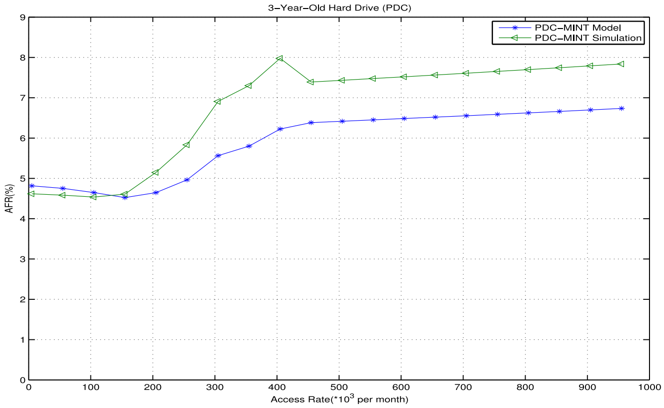


Fig. 19. Impacts of file access rate on AFR (PDC). Access rate varies from 10 to $64 * 10^4$ no./month.

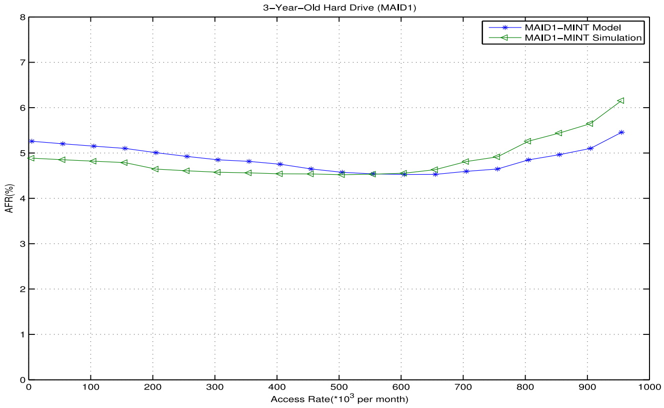


Fig. 20. Impacts of file access rate on AFR (MAID1). Access rate varies from 10 to $64 * 10^4$ no./month.

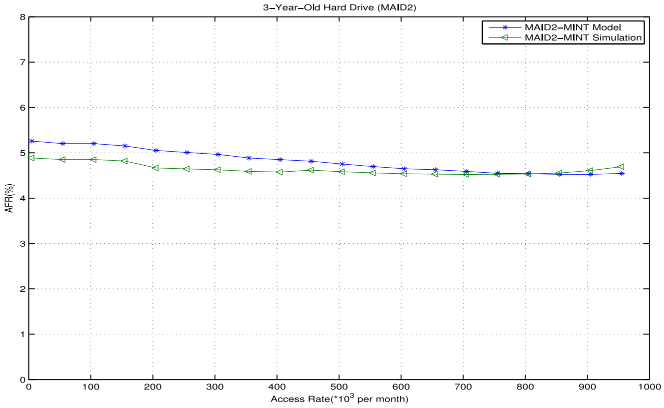


Fig. 21. Impacts of file access rate on AFR (MAID2). Access rate varies from 10 to $64 * 10^4$ no./month.

the AFR issue. Recall that the average difference between the simulation results and the modeling results in terms of utilization is below 10 percent. The model is 10 percent less sensitive to utilization than the simulation does. According to Fig. 3, the average difference between the simulated reliability and the analytical reliability is also below 10 percent.

6 Related Work

Energy-efficient parallel disk systems. Hard disk drives (HDD) are made up of various electrical, electronic, and mechanical components [45]. An array of techniques were developed to save energy in single HDDs. Energy dissipation in disk drives can be reduced at the I/O level (e.g., dynamic power management [8], [17], [38] and multi-speed disks [12]), the operating system level (e.g., power-aware caching/prefetching [48], [34]), and the application level (e.g., software DMP [33] and cooperative I/O [41]).

Existing energy-saving techniques for parallel disk systems often rely on one of the two basic ideas - power management and workload skew. Power management schemes conserve energy by turning disks into standby after a period of idle time. Although multi-speed disks are not widely adopted in storage systems, power management has been successfully extended to address the energy-saving issues in multi-speed disks [12], [11], [16]. The basic idea of workload skew is to concentrate I/O workloads from a large number of parallel disks into a small subset of disks allowing other disks to be placed in the standby mode [20], [7], [27], [21].

Reliability impacts of power management on disks. Recent studies show that both power management and workload skew schemes inherently impose adverse reliability impacts on disk systems [3], [43]. For example, the power management schemes are likely to result in a huge number of disk spin-downs and spin-ups that can significantly reduce the lifespan of hard disks.

The workload skew techniques dynamically migrates frequently accessed data to a subset of disks [26], [18], which inherently have higher risk of breaking down than other disks usually being kept standby. Disks storing popular data tend to have high failure rates due to extremely unbalanced workload. Thus, the popular data disks have a strong likelihood to become reliability bottleneck. The design of our MINT is orthogonal to the aforementioned energy saving studies, because MINT is focused on reliability impacts of the power management and workload skew schemes in parallel disks.

Reliability models of disk systems. A malfunction of any components in a hard disk drive could lead to a failure of the disk. Reliability—one of the key characteristics of disks—can be measured in terms of mean-time-between-failure (MTBF). Disk manufacturers usually investigate MTBFs of disks either by laboratory testing or mathematical moedling. Although disk drive manufacturers claim that MTBF of most disks is more than 1 million hours [31], users have experienced a much lower MTBF from their field data [9]. More importantly, it is challenging to measure MTBF because of a wide range of contributing factors including disk age, utilization, temperature, and power-state transition frequency [9].

A handful of reliability models have been successfully developed for storage systems. For example, Pâris et al. investigated an approach to computing both average failure rate and mean time to failure in distributed storage systems [19]; Elerath and Pecht proposed a flexible model for estimating reliability of RAID storage [10]; Xin et al. developed a model to study disk infant mortality [44]; and Rao et al. used Markov models to determine MTTDL [24]. Unlike these reliability models tailored for conventional parallel and distributed disk systems, our MINT model pays special attention to reliability of parallel disk systems coupled with energy-saving mechanisms.

Model validations. Model validation is a process of improving levels of confidence [30]. Major approaches to validating models include historical methods and extreme condition tests. For example, Brown and Ochoa validated their reliability models of distributions systems using historical results [4]. In the second approach, model structures and outputs should be plausible for any extreme and unlikely combination of levels of factors in a system as well as comparison to other models [29]. We developed a trace-driven simulator using the Berkeley Web Trace [1] as a reference model with which our MINT model is compared. The Web trace is used to drive the simulator, because we focus on read-intensive applications [25].

7 Conclusion

In recognition that there is a lack of models designed to evaluate reliability of energy-efficient disk systems, we propose a new modeling framework called MINT to measure the reliability of parallel disk systems equipped with reliability-affecting energy conservation techniques. We first developed models to study the impacts of disk utilization and power-state transition frequency on reliability of each disk in a parallel disk system. We then derived the reliability of an individual disk from its utilization, age, temperature, and power-state transitions. Finally, we applied the MINT framework to investigate the reliability of parallel disks coupled with the MAID technique and the PDC technique. Compared our MINT reliability framework with other existing models, MINT has the following advantages:

- MINT captures the behaviors of PDC and MAID in terms of data movement and migration as well as power-state transitions.
- MINT seamlessly integrates multiple reliability-affecting factors into a coherent form.
- MINT can be used to evaluate the system-level reliability of an energy-efficient parallel disk system.

To address the issue of model validation, we developed a trace-drive simulator as a control group. The results obtained from the simulator driven by the Berkeley Web trace are compared to those provided by our MINT model. The validation results indicate that MINT exhibits a similar trend as that of the simulator driven by a real-world trace.

We have identified the following future directions of this research. First, the reliability models presented in this paper are focused on read-intensive I/O activities. We will extend the MINT modeling framework to investigate mixed read/write workloads. Second, we will investigate a fundamental way of making tradeoffs between reliability and energy-efficiency in the context of energy-efficient parallel disks. A tradeoff curve will be used as a unified framework to justify circumstances under which it is worth trading reliability for high energy efficiency in parallel disks. Third, we will develop a Weibull-Distribution based model to enhance our MINT's Utilizaiton-AFR sub-model and to make our MINT represent more general situation.

Last, we will collaborate with disk vendors to collect field test data, which will be applied to tune the α and β parameters to further improve the accuracy of the MINT model.

Acknowledgments

Qin's research reported in this paper was supported by the US National Science Foundation under Grants CCF-0845257 (CAREER), CNS-0757778 (CSR), CCF-0742187 (CPA), CNS-0917137 (CSR), CNS-0831502 (CyberTrust), CNS-0855251 (CRI), OCI-0753305 (CI-TEAM), DUE-0837341 (CCLI), and DUE-0830831 (SFS), as well as Auburn University under a startup grant and a gift (Number 2005-04-070) from the Intel Corporation. Shu Yin's research was supported by the Hunan Province Natural Science Foundation under

Grant 14JJ4025, the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of the People's Republic of China, and Hunan University under a start-up grant.

REFERENCES

[1] "The now Trace Collection Project," http://tracehost.cs.berkeley.edu/web/.

[2] "The Distributed-Parallel Storage System (Dpss) Home Pages," http://www-didc.lbl.gov/DPSS/, June 2004.

[3] K. Bellam, A. Manzanares, X. Ruan, X. Qin, and Y.-M. Yang, "Improving Reliability and Energy Efficiency of Disk Systems via Utilization Control," *Proc. IEEE Symp. Computers and Comm.*, 2008.

[4] R.E. Brown and J.R. Ochoa, "Distribution System Reliability: Default Data and Model Validation," *IEEE Trans. Power Systems*, vol. 13, no. 2, pp. 704-709, May 1998.

[5] W.A. Burkhard and J. Menon, "Disk Array Storage System Reliability," *Proc. 23rd Int'l Symp. Fault-Tolerant Computing*, pp. 432-441, 1993.

[6] Enrique V. Carrera, Eduardo Pinheiro, and Ricardo Bianchini, "Conserving Disk Energy in Network Servers," *Proc. 17th Ann. Int'l Conf. Supercomputing (ICS '03)*, pp. 86-97, 2003.

[7] D. Colarelli and D. Grunwald, "Massive Arrays of Idle Disks for Storage Archives," *Proc. ACM/IEEE Conf. Supercomputing*, pp. 1-11, 2002.

[8] F. Douglis, P. Krishnan, and B. Marsh, "Thwarting the Power-Hungry Disk," *Proc. USENIX Winter Technical Conf.*, pp. 23-23, 1994.

[9] J.G. Elerath, "Specifying Reliability in the Disk Drive Industry: No More MTBF's," *Proc. Ann. Reliability and Maintainability Symp.*, pp. 194-199, 2000.

[10] J.G. Elerath and M. Pecht, "Enhanced Reliability Modeling of Raid Storage Systems," *Proc. IEEE/IFIP Int'l Conf. Dependable Systems and Networks*, 2007.

[11] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, "DRPM: Dynamic Speed Control for Power Management in Server Class Disks," *Proc. 30th Ann. Int'l Symp. Computer Architecture*, pp. 169-179, June 2003.

[12] D.P. Helmbold, D.E. Long, T.L. Sconyers, and B. Sherrod, "Adaptive Disk Spin—Down for Mobile Computers," *Mobile Networks and Applications*, vol. 5, no. 4, pp. 285-297, 2000.

[13] G.F. Hughes and J.F. Murray, "Reliability and Security of Raid Storage Systems and d2d Archives using Sata Disk Drives," *ACM Trans. Storage*, vol. 1, no. 1, pp. 95-107, Feb. 2005.

[14] X.-F. Jiang, M.I. Alghamdi, J. Zhang, M. Al Assaf, X.-J. Ruan, T. Muzaffar, and X. Qin, "Thermal Modeling and Analysis of Storage Systems," *Proc. IEEE 31st Int'l Performance Computing and Comm. Conf.*, 2012.

[15] S. Jin and A. Bestavros, "Gismo: A Generator of Internet Streaming Media Objects and Workloads," *ACM SIGMETRICS Performance Evaluation Rev.*, vol. 29, Nov. 2001.

[16] P. Krishnan, M.P. Long, and Scott J. Vitter, "Adaptive Disk Spin-down via Optimal Rent-to-Buy in Probabilistic Environments," Technical report, Duke Univ. , 1995.

[17] K. Li, R. Kumpf, P. Horton, and T. Anderson, "A Quantitative Analysis of Disk Drive Power Management in Portable Computers," *Proc. USENIX Winter Technical Conf.*, pp. 22-22, 1994.

[18] A. Manzanares, X. Ruan, S. Yin, and M. Nijim, "Energy-Aware Prefetching for Parallel Disk Systems: Algorithms, Models, and Evaluation," *Proc. IEEE Int'l Symp. Network Computing and Applications*, 2009.

[19] J.-F. Pâris, T.J. Schwarz, and D.D.E. Long, "Evaluating the Reliability of Storage Systems," *Proc. IEEE Int'l Symp. Reliable and Distributed Systems*, 2006.

[20] E. Pinheiro and R. Bianchini, "Energy Conservation Techniques for Disk Array-Based Servers," *Proc. 18th Int'l Conf. Supercomputing*, 2004.

[21] E. Pinheiro, R. Bianchini, E. Carrera, and T. Heath, "Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems," *Proc. Workshop Compilers and Operating Sys. for Low Power*, Sept. 2001.

[22] E. Pinheiro, R. Bianchini, and C. Dubnicki, "Exploiting Redundancy to Conserve Energy in Storage Systems," *Proc. Joint Int'l Conf. Measurement and Modeling of Computer Systems*, 2006.

[23] E. Pinheiro, W.-D. Weber, and L.A. Barroso, "Failure Trends in a Large Disk Drive Population," *Proc. USENIX Conf. File and Storage Technologies*, February 2007.

[24] K.K. Rao, J.L. Hafner, and R.A. Golding, "Reliability for Networked Storage Nodes," *IEEE Trans. Dependable and Secure Computing*, vol. 8, no. 3, pp. 404-418, May 2011.

[25] D. Roselli, J.R. Lorch, and T.E. Anderson, "A Comparison of File System Workloads," *Proc. Ann. Conf. USENIX Ann. Technical Conf. (ATEC '00)*, pp. 4-4, 2000.

[26] X.J. Ruan, A. Manzanares, K. Bellam, Z.L. Zong, and X. Qin, "DARAW: A New Write Buffer to Improve Parallel I/O Energy-Efficiency," *Proc. ACM Symp. Applied Computing*, 2009.

[27] X.-J. Ruan Run, A. Manzanares, S. Yin, Z.-L. Zong, and X. Qin, "Performance Evaluation of Energy-Efficient Parallel I/O Systems with Write Buffer Disks," *Proc. 38th Int'l Conf. Parallel Processing*, Sept. 2009.

[28] R.G. Sargent, "Verification and Validation of Simulation Models," *Proc. 37th Conf. Winter Simulation (WSC '05)*, pp. 130-143, 2005.

[29] R.G. Sargent, "Verification and Validation of Simulation Models," *Proc. 37th conf. Winter simulation Conf. (WSC '05)*, pp. 130-143, 2005.

[30] S. Schlesinger, R.E. Crosbie, R.E Gagne, and I. GSd, "Terminology for Model Credibility," *Simulation*, vol. 32, pp. 103-104, 1979.

[31] B. Schroeder and G.A. Gibson, "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?" *Proc. USENIX Conf. File and Storage Technologies*, 2007.

[32] S. Shah and J.G. Elerath, "Reliability Analysis of Disk Drive Failure Mechanisms," *Proc. Ann. Reliability and Maintainability Symp.*, pp. 226-231, 2005.

[33] S.W. Son, M. Kandemir, and A. Choudhary, "Software-Directed Disk Power Management for Scientific Applications," *Proc. IEEE Int'l Parallel and Distributed Processing Symp.*, 2005.

[34] S.W. Son and M. Kandemir, "Energy-Aware Data Prefetching for Multi-Speed Disks," *Proc. Int'l Conf. Computing Frontiers*, 2006.

[35] *Specification of Hard Disk Drive Reliability*, IDEMA Standards PagesDocument Number R298, 1998.

[36] A. Thomasian and M. Blaum, "Mirrored Disk Organization Reliability Analysis," *IEEE Trans. Computers*, vol. 55, no. 12, pp. 1640-1644, Dec. 2006.

[37] P.J. Varman and R.M. Verma, "Tight Bounds for Prefetching and Buffer Management Algorithms for Parallel I/O Systems," *IEEE Trans. Parallel and Distributed Systems*, vol. 10, no. 12, pp. 1262-1275, Dec. 1999.

[38] A. Verma, R. Koller, L. Useche, and R. Rangaswami, "SRCMap: Energy Proportional Storage using Dynamic Consolidation," *Proc. Eighth USENIX Conf. File and Storage Technologies (FAST '10)*, 2010.

[39] J. Wang, H.-J. Zhu, and D. Li, "eRAID: Conserving Energy in Conventional Disk-Based Raid System," *IEEE Trans. Computers*, vol. 57, no. 3, pp. 359-374, Mar. 2008.

[40] C. Weddle, M. Oldham, J. Qian, A.-I.A. Wang, P. Reiher, and G. Kuenning, "PARAID: A Gear-Shifting Power-Aware Raid," *IEEE Trans. Storage*, vol. 3, article 13, Oct. 2007.

[41] A. Weissel, B. Beutel, and F. Bellosa, "Cooperative I/O: A Novel I/O Semantics for Energy-Aware Applications," *Proc. Fifth Symp. Operating Systems Design and Implementation*, pp. 117-129, 2002.

[42] T. Xie, "SEA: A Striping-Based Energy-Aware Strategy for Data Placement in Raid-Structured Storage Systems," *IEEE Trans. Computers*, vol. 57, no. 6, pp. 748-761, June 2008.

[43] T. Xie and Y. Sun, "Sacrificing Reliability for Energy Saving: Is it Worthwhile for Disk Arrays?" *Proc. IEEE Symp. Parallel and Distributed Processing*, pp. 1-12, Apr. 2008.

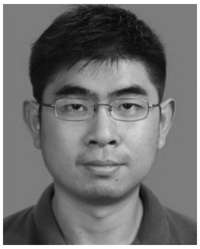
[44] Q. Xin, J.E. Thomas, S.J. Schwarz, and E.L. Miller, "Disk Infant Mortality in Large Storage Systems," *Proc. IEEE Int'l Symp. Modeling, Analysis, and Simulation of Computer and Telecomm. Systems*, 2005.

[45] J. Yang and F.-B. Sun, "A Comprehensive Review of Hard-Disk Drive Reliability," *Proc. Ann. Reliability and Maintainability Symp.*, 1999.

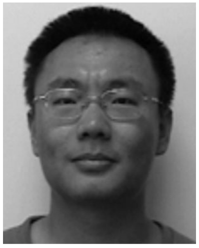
[46] Q. Yang and Y.-M. Hu, "DCD - Disk Caching Disk: A New Approach for Boosting I/O Performance," *Proc. Int'l Symp. Computer Architecture*, pp. 169-169, May 1996.

[47] S. Yin, X. Ruan, A. Manzanares, and X. Qin, "How Reliable are Parallel Disk Systems When Energy-Saving Schemes are Involved?" *Proc. IEEE Int'l Conf. Cluster Computing (CLUSTER)*, 2009.

[48] Q.-B. Zhu, F.M. David, C.F. Devaraj, Z.-M. Li, Y.-Y. Zhou, and P. Cao, "Reducing Energy Consumption of Disk Storage using Power-Aware Cache Management," *Proc. Int'l Symp. High Performance Computer Architecture*, 2004.



Shu Yin (S'09) received the BS degree in communication engineering from Wuhan University of Technology (WUT), China, in 2006, the MS degree in signal and information processing from WUT in 2008, and the PhD degree in computer science from the Auburn University, Alabama, in 2012. Currently, he is an assistant professor in the School of Information Science and Engineering at Hunan University, China. During July-December 2011, he worked as an intern at the Los Alamos National Laboratory, New Mexico. His research interests include storage systems, reliability modeling, fault tolerance, energy-efficient computing, high performance computing and wireless communications.



Xiaojun Ruan (S'07) received the BS degree in computer science from Shandong University, China, in 2005. He received the PhD degree in Computer Science from Auburn University, Alabama, in 2011. He is presently an assistant professor with the Department of Computer Science, West Chester University of Pennsylvania. His research interests include parallel and distributed systems, storage systems, real-time computing, performance evaluation, and fault-tolerance, with a focus on highperformance parallel cluster computing, storage system, and distributed system.

Adam Manzanares (S'06) received the BS degree in computer science from the New Mexico Institute of Mining and Technology, Socorro, in 2006. He received the PhD degree in Computer Science from Auburn University, Alabama, in May 2011. He was an assistant professor with the Computer Science Department, California State University-Chico. Currently he is a research staff member at HGST, a Western Digital company. During the summers of 2002-2007, he worked as a student intern at the Los Alamos National Laboratory, New Mexico. His research interests include energy efficient computing, modeling and simulation, and high performance computing.

Xiao Qin (S'00-'04M-SM'09) received the BS and MS degrees in computer science from Huazhong University of Science and Technology in 1992 and 1999, respectively. He received the PhD degree in computer science from the University of Nebraska-Lincoln in 2004. He is currently an associate professor in the Department of Computer Science and Software Engineering at Auburn University. Prior to joining Auburn University in 2007, he had been an assistant professor with New Mexico Institute of Mining and Technology (New Mexico Tech) for three years. He won an NSF CAREER award in 2009. His research is supported by the US National Science Foundation (NSF), Auburn University, and Intel Corporation. He has been on the program committees of various international conferences, including IEEE Cluster, IEEE MSST, IEEE CCGrid, IEEE IPCCC, and ICPP. His research interests include parallel and distributed systems, storage systems, fault tolerance, real-time systems, and performance evaluation. He is a member of the ACM and a senior member of the IEEE.

Kenli Li received the PhD in computer science from Huazhong University of Science and Technology, China, in 2003. He was a visiting scholar at University of Illinois at Champaign and Urbana from 2004 to 2005. Now He is a professor of Computer science and Technology at Hunan University, associate director of National Supercomputing Center in Changsha. His major research includes parallel computing, Grid and Cloud computing, and DNA computer. He has published more than 70 papers in international conferences and journals, such as IEEE TC, IEEE TPDS, JPDC, ICPP, CCGrid. He is an outstanding member of CCF.