How Reliable are Parallel Disk Systems When Energy-Saving Schemes are Involved?

Shu Yin, Xiaojun Ruan, Adam Manzanares, and Xiao Qin
Department of Computer Science and Software Engineering
Auburn University, Auburn, AL 36849
Email: {szy0004, xzr0001, acm0008, xqin}@auburn.edu

Abstract—**Many energy conservation techniques have been proposed to achieve high energy efficiency in disk systems. Unfortunately, growing evidence shows that energy-saving schemes in disk drives usually have negative impacts on storage systems. Existing reliability models are inadequate to estimate reliability of parallel disk systems equipped with energy conservation techniques. To solve this problem, we propose a mathematical model - called MINT - to evaluate the reliability of a parallel disk system where energy-saving mechanisms are implemented. In this paper, we focus on modeling the reliability impacts of two well-known energy-saving techniques - the Popular Disk Concentration technique (PDC) and the Massive Array of Idle Disks (MAID). We started this research by investigating how PDC and MAID affect the utilization and power-state transition frequency of each disk in a parallel disk system. We then model the annual failure rate of each disk as a function of the disk's utilization, power-state transition frequency as well as operating temperature, because these parameters are key reliability-affecting factors in addition to disk ages. Next, the reliability of a parallel disk system can be derived from the annual failure rate of each disk in the parallel disk system. Finally, we used MINT to study the reliability of a parallel disk system equipped with the PDC and MAID techniques. Experimental results show that PDC is more reliable than MAID when disk workload is low. In contrast, the reliability of MAID is higher than that of PDC under relatively high I/O load.**

Keywords-**MAID; PDC; reliability; enery-saving; cluster;**

I. 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 storge capacity. In the past decades, parallel disk systems have increasingly become popular for data-intensive applications running on massively parallel computing platforms [23]. 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. More specifically, storage devices account for approximately 27 percent of the overall energy consumption in a data centre. Current utilization and technological trends of parallel disk systems result in unacceptable economical and environmental consequences [14]. To address this issue, 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 [20], dynamic power management schemes [4], data redundancy techniques [1], workload skew [10][26], and multi-speed settings [7]

Prior findings show that existing energy conservation techniques in disk drives can deliever 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 [1][24]. For example, the dynamic power management (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 [4][8][11]. Unlike DPM, workload-skew techniques such as MAID [3], PDC [14], and BUD [17] 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.

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, [2] and [22]). However, existing disk reliability models are inadequate for evaluating reliability of disk systems epuipped with energy-saving mechanisms. For example, Shah and Elerath conducted a series of reliability analyses using field failure data of several drive models from various disk drive manufacturers [19]. Hughes and Murray investigated SATA disk drive reliability factors that bear on storage system performance [9]. They not only studied SATA drive operating failure rates, but also proposed approaches to improving reliability of storage systems comprised of multiple SATA disks [9]. Reliability models that do not consider energy-saving mechanisms are quite inaccurate when it comes to the estimation of reliability of energy-efficient disk systems.

In addition to reliabilty model, simulation is another key technique to study the reliaiblity of energy-efficient storage

978-1-4244-5012-1/09/$25.00 ©2009 IEEE

systems (see, for example, [24]. Simulating reliability of disk systems with energy-conservation techniques is often complicated, because accurate simulation 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 lack of fast and accurate methodology to evaluate reliability of modern parallel disk systems with high energy efficiency. To address this problem, we propose a mathematical model called MINT to quantitatively investigate the reliability of parallel disk systems employing a variety energy conservation schemes.

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.

We used the MINT model to comprehensively study the reliability of parallel disk systems equipped with two well-known energy-saving schemes, namely, the Popular Disk Concentration technique (PDC) [14] and the Massive Array of Idle Disks (MAID) [3]. We paid particular attention on PDC and MAID, because our focus is the power optimization strategies that adversely affect reliability 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. In particular, the main contributions of this paper include:

1) A generic mathematical approach to modeling reliability of energy-efficient parallel disks coupled with power management optimization policies.
2) Two reliability models for the two well-known energy-saving schemes - Popular Data Concentration scheme (PDC) and Massive Array of Idle Disks (MAID).
3) Intriguing impacts of PDC and MAID on the reliability of parallel disk systems.

The remainder of this paper is organized as follows. Section II introduces the related work of this study. Section III outlines the design and implementation of the MINT reliability modeling framework, which relies on disk utilization, temperature, and power-state transition frequency. In Section IV, we apply MINT to propose a reliability model for the Popular Data Concentration scheme (PDC). Section V presents a reliability model for the Massive Array of Idle Disks technique (MAID). In Section VI, we discuss results of experimental studies by comparing the reliability impacts of PDC and MAID. Section VII contains concluding remarks.

II. Related Work

A hard disk drive (HDD) is a complex dynamic system made up of various electrical, electronic, and mechanical components. 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 [4] and multi-speed disks [8]), the operating system level (e.g., power-aware caching/prefetching [12]and the application level (e.g. software DMP [20]). 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 [8][7] 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 [14][3][17].

Recent studies show that both power management and workload skew schemes inherently impose adverse impacts on disk systems [1][24]. 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 hard disk lifetime. The workload skew techniques dynamically migrates frequently accessed data to a subset of disks [16] [12], 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.

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 [18], users have experienced a much lower MTBF from their field data [5]. 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 [5].

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 [13]; Elerath and Pecht proposed a flexible model for estimating reliability of RAID storage [6]; and Xin *et. al* developed a model to study disk infant mortality [25]. 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.

Xie and Sun developed an empirical reliability model called PRESS (Predictor of Reliability for Energy Saving Schemes) [24]. The PRESS model can be used to estimate reliability of an entire disk array [24]. To fully leverage PRESS to study the reliability of disk arrays, one has to properly simulate the disk arrays. Our MINT approach differs itself from PRESS in the sense that the goal of MINT is to evaluate reliability of disk systems by modeling the behavior of parallel disks where energy conservation mechanisms are integrated.

III. The MINT Reliability Model

A. Framework

MINT is composed of a single disk reliability model, a system-level reliability model, and three reliability-affecting factors - temperature, disk state transition frequency (hereinafter referred to as frequency) and utilization. Many energy-saving schemes (e.g., PDC [14] and MAID [3]) inherently affect reliability-related factors like disk utilization and transition frequency. Given an energy optimization mechanism, MINT first transfers data access patterns into the two reliability-affecting factors - frequency and utilization. The single-disk reliability model can derive individual disk's annual failure rate from utilization, power-state transition frequency, age, and temperature. Each disk's reliability is used as input to the system-level reliability model that estimates the annual failure rate of parallel disk systems.

For simplicity without losing generality, we consider four reliability-related factors in MINT. This assumption does not necessarily indicate that disk utilization, age, temperature, and power-state transitions are the only parameters affecting disk reliability. Other factors having impacts on reliability include handling, humidity, voltage variation, vintage, duty cycle, and altitude [5]. If a new factor has to be integrated into MINT, we can extend the single reliability model described in Section III-E. Since the infant mortality phenomena is out the scope of this study, we pay attention to disks that are more than one year old.

B. Impacts of Utilization on Disk Annual Failure Rate

Disk utilization can be characterized as the fraction of active time of a disk drive out of its total powered-on-time [15]. 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 (AFR). Using field failure data collected by Google, Pinheiro *et al.* shows the impact of utilization on AFR across the different age groups. Pinheiro *et al.* studied the impact of utilization on AFR accross different disk age groups. They categorized disk utilization in three levels - low, medium, and high. Pinheiro *et al.* illustrated AFRs of 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 needs a baseline AFR derived from a numerical value of utilization, we make use of the polynomial curve-fitting technique to model the baseline value of a single disk's AFR as a function of utilization. Thus, the baseline value (i.e., $BaseValue$ in Eq. 3) of AFR for a disk can be calculated from the disk's utilization. For example, Fig. 1 shows the AFR value of a 3-year old disk as a function of its utilization. The curve plotted in Fig. 1 can be modeled as a utilization-reliability function described as Eq. 1 below:

$$\begin{aligned} R(u) =& 4.167e^{-7}u^4 - 7.5e^{-5}u^3 + 5.968e^{-3}u^2 - \\ &- 2.575e^{-1}u + 9.3, \quad \text{for all } u \in [0, 100] \end{aligned} \tag{1}$$

where $R(u)$ represents the AFR value as a function of a certain disk utilization u. With Eq. 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% utilization (i.e., $u = 50\%$), we can obtain the AFR value of this disk as $R(u) = 4.8\%$. Fig. 1 suggests that unlike the conclusions drawn in a previous study (see [21]), a low disk utilization does not necessarily lead to low AFR. For instance, given a 3-year old disk, the AFR value under 30% utilization is even higher than AFR under 80% utilization.

C. 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 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, in the low and middle temperature ranges, the failure rates decreases when temperature increases [15].

Growing evidence shows that disk reliability should reflect disk drives operating under environmental conditions like temperature [5]. Since temperature (e.g., meaured 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 integrted [5]. Given a temperature,

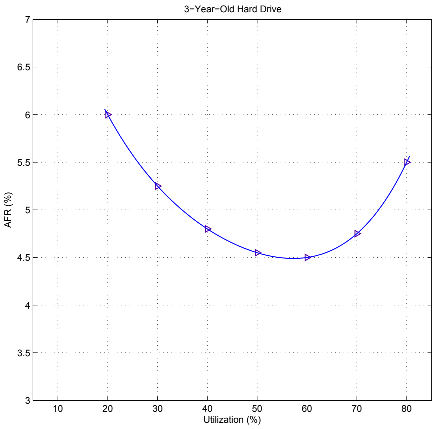

Figure 1: 3-Year-Old HDD Utilization Impacts on AFR

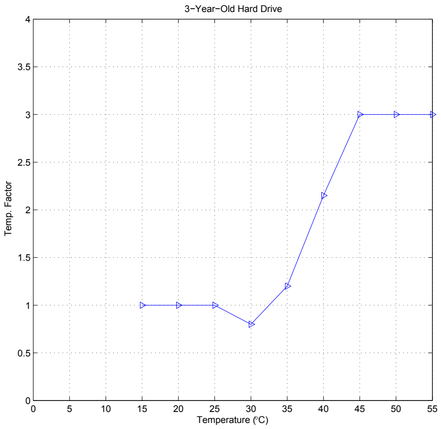

Figure 2: Temperature-Factor Function of 3-Year-Old HDDs

one must decide the corresponding temperature factor (see $TemperatureFactor$ in Eq. 3) that can be multiplied to the base failure rates. Using Google's field failure data, we attempted to calculate temperature factors under various temperatures ranges for disks with different ages.

Now we explain how to determine a temperature facotr 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 T denote the average temperature, we define the temperature factor for temperature T as $T/25$ if T is larger than 25°C. When T exceeds 45°C, the temperature factor becomes a constant (i.e., $1.8 = 45/25$). Due to space limit, we only show how temperature affects the temperature factor of a 3-year old disk. Note that the temperature-factor functions for disks in other age ranges can be modeled in a similar way. Fig. 2 shows the temperature-factor function derived from the Google's field failure data for 3-year old disks.

D. 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 Eq. 3 in Section III-E). 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 [21] and the PRESS model [24]. Again, we focus on 3-year old disk drives. Fig. 3 demonstrates frequency adder values as a function of power-state transition frequency. Fig. 3 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 Eq. (2)) plotted in Fig. 3.

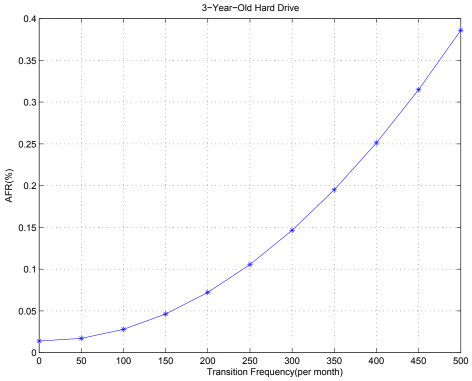

Figure 3: Impacts of Transition Frequency on Frequency adder of 3-Year-Old HDDs

$$R(f) = 1.51e^{-6}f^2 - 1.09e^{-5}f + 1.39e^{-2}, f \in [0, 100] \quad (2)$$

where f is a power-state transition frequency, $R(f)$ represents an adder to the base AFR value. For example, suppose the transition frequency is 300 per month, the base AFR value needs to be increase by 1.33%.

E. Single Disk Reliability Model

Single-disk reliability can not be accurately described by one valued parameter, because the disk drive reliability is affected by multiple factors (see Sections III-B,III-C, and III-D). Though recent studies attempted to consider multiple reliability factors (see, for example, PRESS [24]), 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 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:

$$\begin{aligned} R =& \alpha \times BaseValue \times TemperatureFactor + \\ &+ \beta \times FrequencyAdder \end{aligned} \tag{3}$$

where $BaseValue$ is the baseline failure rate derived from disk utilization (see Section III-B), $TemperatureFactor$ is the temperature factor (or temperature multiplier; see Section III-C), $FrequencyAdder$ is the power-state transition frequency adder to the base AFR (see Section III-D), and α and β are two coefficients to reliability R. If reliability R 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 cases, α and β can be set in accordance with R's sensitivities to utilization, temperature, and frequency. In our experiments, we assume that all the three reliability-related factors are equally important (i.e., α=β=1). Ideally, extensive field tests allow us to analyze and test the two coefficients. Although α and β are not fully evaluated by field testing, reliability results are valid because of the following two reasons: first, we have used the same values of α and β to evaluate impacts of the two energy-saving schemes on disk reliability (see Section IV); second, the failure-rate trend of a disk when α and β are set to 1 are very similar to those of the same disk when the values of α and β do not equal to 1.

With Eq. 3 in place, we can analyze a disk's reliability in turns of annual failure rate (AFR). Fig. 4 shows AFR of a three-year-old disk when its utilization is in the range between 20% and 80%. We observe from Fig. 4 that increasing temperature from 35°C to 40°C gives rise to a significant increase in AFR. Unlike temperature, power-state transition frequency in the range of a few hundreds per month has marginal impact on AFR. It is expected that when transition frequency is extremely high, AFR becomes more sensitive to frequency than to temperature.

IV. Reliability Model for MAID

A. Background

The MAID (Massive Arrays of Idle Disks) technique - developed by Colarelli and Grunwald - aims to reduce energy consumption of large disk arrays while maintaining acceptable I/O performance [3]. 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. 5 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. 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) a power management policy that switches idle disks into the standby mode if the disks are sitting idle for a certain period of time; (2) a data placement module that either linearly places successive blocks on a disk drive or uniformly distributes data blocks across multiple drives; (3) a cache disk controller that determines the number of disks performing as cache disks [3].

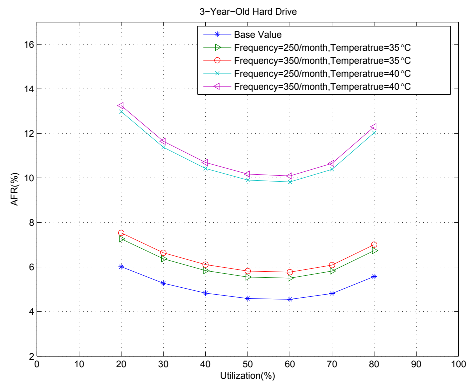

Figure 4: 3-Year-Old HDD Combined Factors Impacts on AFR
(Single Disk Reliability Model)

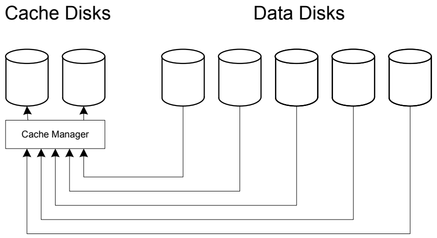

Figure 5: MAID System Structure

B. Modeling Utilization of Disks 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

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

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

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

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 l-th phase, let f_{ijl} be the j-th file on the i-th disk, where $i \in (1, 2, \cdots, n)$, $j \in (1, 2, \cdots, m_i)$, $l \in (1, 2, \cdots k)$. We have:

$$\begin{aligned} F_{1l} &= \{ f_{11l}, f_{12l}, \cdots, f_{1m_1l} \} \\ &\vdots \\ F_{nl} &= \{ f_{n1l}, f_{n2l}, \cdots f_{nm_nl} \} \end{aligned} \qquad (6)$$

where m_i is the number of files on the ith disk and 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.

$$\begin{aligned} F'_{1l} &= \left\{ f'_{11l}, f'_{12l}, \cdots, f'_{1m'_1l} \right\} \\ &\vdots \\ F'_{nl} &= \left\{ f'_{n1l}, f'_{n2l}, \cdots, f'_{nm'_nl} \right\} \end{aligned} \qquad (7)$$

where $m'_i$is the number of the files on the i-th disk, f'_{ijl} is the j-th file at the l-th phase and 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 $\rho_{ijl} = \lambda_{ijl} \times t$. 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. For both cache and data disks, we express the utilization for ith disk in phase l as:

$$\begin{aligned} \rho'_{il} &= \frac{{}^{I}\!/\!_{O} time + Copying\ time}{T} \\ &= \sum_{j=1}^{m'_i} \rho'_{ijl} + \frac{Copying\ time}{T} \end{aligned} \qquad (8)$$

where T is the time interval of the lth I/O phase. The first and second items on the bottom-line on the right-hand side of Eq. 8 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.

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:

$$\rho'_{il-data} = \sum_{j=1}^{m'_i} \rho'_{ijl} + \frac{\sum\limits_{j \in F_{il}^{M\_out}} t_{ijl}}{T} \qquad (9)$$

where the first and second items on the right-hand side of Eq. 9 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:

$$\rho'_{il-cache} = \sum_{j=1}^{m'_i} \rho'_{ijl} + \frac{\sum\limits_{j \in F_{il}^{M\_in}} t_{ijl}}{T} \qquad (10)$$

where the first and second items on the right-hand side of Eq. 10 are the utilizations of accessing files and writing files to the cache disk to make replicas, respectively.

C. Modeling Power-State Transition Frequency for MAID

Eq. 3 in Section III-E 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 τ_i of idle periods that are larger than the break-even time T_{BE}. The number of power-state transitions during phase i is $2\tau_i$, 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 τ can be expressed as: $\tau = 2 \times \sum_{i=1}^{N} \tau_i$.

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.

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.

V. Reliability Model for PDC

A. Background

The PDC (Popular Data Concentration) technique proposed by Pinheiro and Bianchini migrates frequently accessed data to a subset of disks in a disk array [14]. Fig. 6 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 to migrate data onto a disk until its load is approaching the maximum bandwidth.

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.

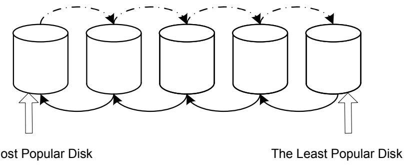

Figure 6: PDC System Structure

B. Modeling Utilization of Disks 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:

$$\rho'_{il} = \sum_{j=1}^{m'_i} \rho'_{ijl} + \frac{Migration\ time}{T} \tag{11}$$

where T is the time interval of I/O phase l. The first and second items on the right-hand side of Eq. 11 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 Eq. 11), 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,

$$\rho'_{il} = \sum_{j=1}^{m'_i} \rho'_{ijl} + \frac{\sum\limits_{j \in F_{il}^{M}} t_{ijl}}{T_l} \tag{12}$$

where the second item on the right-hand side of Eq. 12 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.

C. Modeling Power-State Transition Frequency for PDC

We used the same way described in Section IV-C to model power-state transition frequency for PDC. Unlike MAID, PDC allows each disk to receive migrated data from other

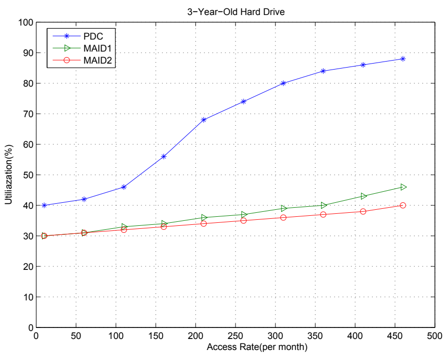

Figure 7: Impacts of File Access Rate on Utilization Access Rate is increased from 10 to 500 No./month

disks. In light of PDC, disks storing the most popular files are most likely to be kept in the active mode.

VI. Experimtnal Study

We considered two approaches to configuring MAID. In the first approach, extra cache disks are added to a disk array to cache frequently accessed data. In the second one,we make use of existing data disks as cache disks. We analyze the reliability of the two different MAID configurations. In the first configuration or MAID-1, there are 5 cache disks and 20 data disks. In the second configuration or MAID-2, there are 5 cache disks and 15 data disks. For the case of PDC, we set the number of disks to 20. The file access rate is in the range from 0 to 500 No./month. The temperature is set to 35°C.

We first investigate the impacts of file access rate on utilization of MAID and PDC. Fig. 7 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. Moreover, the utilization of PDC is significantly higher than those of MAID-1 and MAID-2. For example, when the average file access rate is 500 No./month, the utilizations of PDC, MAID-1, and MAID-2 are approaching to 90%, 48%, and 40%, 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-1 and MAID-2 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, 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.

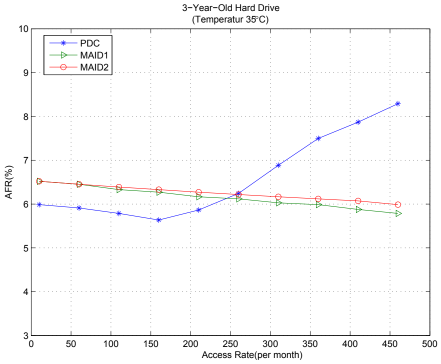

Figure 8: Impacts of File Access Rate on Annual Failure Rate (AFR) of PDC, MAID-1, and MAID-2 (Temperature=35°C)

Fig. 8 shows the annual failure rates or AFR of MAID-1, MAID-2, and PDC. We observe from Fig. 8 that the AFR value of PDC keeps increasing from 5.6% to 8.3% 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 150, AFR of PDC slightly reduces from 5.9% to 5.6% when the access rate is increased from 5 to 150. 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 [15]. When the file access rate 150, the disk utilization is approximately 50%, which is the turing point of the utilization function.

Unlike the AFR of PDC, the AFRs of MAID-1 and MAID-2 continue decreasing from 6.3% to 5.8% 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%, which is in the declining part of the curve.

VII. Conclusion

In recognition that existing disk reliability models cannot be used to evaluate reliability of energy-efficient disk systems, we propose a new model called MINT to evaluate the reliability of a disk array equipped with reliability-affecting energy conservation techniques. We first model the impacts of disk utilization and power-state transition frequency on reliability of each disk in a disk array. We then derive the reliability of an individual disk from its utilization, age, temperature, and power-state transition frequency. Finally, we use MINT to study the reliability of disk arrays coupled with the MAID (Massive Array of Idle Disks) technique and the PDC (Popular Disk Concentration technique) technique.

Acknowledgment

The work reported in this paper was supported by the U.S. National Science Foundation under Grants CCF-0845257(CAREER), CNS-0757778, CCF-0742187, CNS-

0917137, CNS-0831502, CNS-0855251, OCI-0753305, DUE-0837341, and DUE-0830831.

REFERENCES

[1] K. Bellam, A. Manzanares, X. Ruan, X. Qin, and Y.-M. Yang. Improving reliability and energy efficiency of disk systems via utilization control. In *Proc. IEEE Symp. Computers and Comm.*, 2008.

[2] W.A. Burkhard and J. Menon. Disk array storage system reliability. In *Proc. 23rd Int'l Symp. Fault-Tolerant Comp.*, pages 432–441, 1993.

[3] D. Colarelli and D. Grunwald. Massive arrays of idle disks for storage archives. In *Proc. ACM/IEEE Conf. Supercomputing*, pages 1–11, 2002.

[4] F. Douglis, P. Krishnan, and B. Marsh. Thwarting the power-hungry disk. In *Proc. USENIX Winter 1994 Technical Conf.*, pages 23–23, 1994.

[5] J.G. Elerath. Specifying reliability in the disk drive industry: No more mtbf's. pages 194–199, 2000.

[6] J.G. Elerath and M. Pecht. Enhanced reliability modeling of raid storage systems. In *Proc. Int'l Conf. Dependable Sys. and Networks*, 2007.

[7] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke. Drpm: dynamic speed control for power management in server class disks. In *Proc. Int'l Symp. Comp. Arch.*, pages 169–179, June 2003.

[8] D.P. Helmbold, D.E. Long, T.L. Sconyers, and B. Sherrod. Adaptive disk spin—down for mobile computers. *Mob. Netw. Appl.*, 5(4):285–297, 2000.

[9] G.F. Hughes and J.F. Murray. Reliability and security of RAID storage systems and D2D archives using SATA disk drives. *ACM Trans. Storage*, 1(1):95–107, Dec. 2004.

[10] S. Jin and A. Bestavros. Gismo: A generator of internet streaming media objects and workloads. *ACM SIGMETRICS Performance Evaluation Review*, November 2001.

[11] K. Li, R. Kumpf, P. Horton, and T. Anderson. A quantitative analysis of disk drive power management in portable computers. In *Proc. USENIX Winter Technical Conf.*, pages 22–22, 1994.

[12] A. Manzanares, X.-J. Ruan, S. Yin, M. Nijim, and X. Qin. Energy-aware prefetching for parallel disk systems: Algorithms, models, and evaluation. *IEEE Int'l Symp. Network Comp. and Appl.*, 2009.

[13] J.-F. Pâris, T.J. Schwarz, and D.D.E. Long. Evaluating the reliability of storage systems. In *Proc. IEEE Int'l Symp. Reliable and Distr. Sys.*, 2006.

[14] E. Pinheiro and R. Bianchini. Energy conservation techniques for disk array-based servers. In *Proc. 18th Int'l Conf. Supercomputing*, 2004.

[15] E. Pinheiro, W.-D. Weber, and L.A. Barroso. Failure trends in a large disk drive population. In *Proc. USENIX Conf. File and Storage Tech.*, February 2007.

[16] 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. In *Proc. ACM Symp. Applied Computing*, 2009.

[17] 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. In *Proc. Int'l Conf. Parallel Processing*, Sept. 2009.

[18] B. Schroeder and G.A. Gibson. Disk failures in the real world: what does an MTTF of 1,000,000 hours mean to you? In *Proc. USENIX Conf. File and Storage Tech.*, page 1, 2007.

[19] S. Shah and J.G. Elerath. Reliability analysis of disk drive failure mechanisms. In *Proc. Annual Reliability and Maintainability Symp.*, pages 226–231, 2005.

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

[21] IDEMA Standards. Specification of hard disk drive reliability.

[22] A. Thomasian and M. Blaum. Mirrored disk organization reliability analysis. *IEEE Trans. Computers*, 55(12):1640–1644, 2006.

[23] P.J. Varman and R.M. Verma. Tight bounds for prefetching and buffer management algorithms for parallel I/O systems. *IEEE Trans. Parallel Distr. Syst.*, 10(12):1262–1275, 1999.

[24] T. Xie and Y. Sun. Sacrificing reliability for energy saving: Is it worthwhile for disk arrays? In *Proc. IEEE Symp. Parallel and Distr. Processing*, pages 1–12, April 2008.

[25] Q. Xin, J.E. Thomas, S.J. Schwarz, and E.L. Miller. Disk infant mortality in large storage systems. In *Proc. IEEE Int'l Symp. Modeling, Analysis, and Simulation of Computer and Telecomm. Sys.*, 2005.

[26] Q. Yang and Y.-M. Hu. DCD - Disk Caching Disk: A new approach for boosting I/O performance. In *Proc. Int'l Symp. Computer Architecture*, pages 169–169, May 1996.