2013 IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing

Reliability Analysis of Energy-Efficient Parallel Storage Systems

Shu Yin, and Kenli Li∗
School of Info. Sci. and Eng.
Hunan University
Changsha, China 410002
Email:{shuyin, jt_lkl}@hnu.edu.cn
∗Kenli Li is the corresponding author

Xiao Qin
Dept. of Comp. Sci. and Software Eng.
Auburn University
Auburn, AL 36849
Email: xqin@auburn.edu

Xiaojun Ruan
Dept. of Computer Science
West Chester University of Pennsylvania
West Chester, PA 19383
Email: xruan@wcupa.edu

Haiquan Chen
Dept of Math. and Computer Sci.
Valdosta State University
Valdosta, GA 31698
Email: hachen@valdosta.edu

Jiong Xie
Info. and Comm. Center
Inner Mongolia Electric Power
Hohhot, China 010020
Email: jzx0009@auburn.edu

Xiaomin Zhu
Sci. and Tech. on Info. Systems Eng. Laboratory
National University of Defense Technology
Changsha, China 410073
Email: xmzhu@nudt.edu.cn

***Abstract*—Existing reliability models evaluate the lifetime of storage systems well. However, few models aim to analysis parallel storage systems, especially systems equipped with energy-efficient solutions. MREED model provides an inspiring reliability analysis idea for energy-efficient parallel storage systems, especially RAIDs. Unfortunately, the MREED has its disadvantage that it can evaluate RAID-0 only, in which a basic method of data-stripping is applied. In this paper, we extend the MREED model to cover the most energy-efficient RAID setups including RAID-1, which introduces disk-mirroring techniques, and RAID-5, which applies a data-parity techniques based on data-stripping schemes for a data recovery purpose. The MREED is then applied to evaluate an existing well-known power-aware RAID system–PARAID. Experimental results indicate the data locality affection and data parity impaction on the reliability of PARAID. Furthermore, we investigate a fundamental trade-off between reliability and energy-efficiency in terms of energy-saving RAID systems. A tradeoff curve is provided to justify whether it worth to sacrifice reliability for better energy efficiency.**

***Keywords*-Parallel storage system, RAID, energy-efficient, reliability**

I. Introduction

Existing reliability models for conventional parallel and distributed disk systems do not consider energy-saving issues or data-stripping mechanisms. The introduction of **M**athematical **R**eliability model for **E**nergy- **E**fficient RAI**D** system called MREED is rather heuristic. As a mathematical model, MREED shows its advantage of presenting the reliability trend of energy-aware storage systems [1]. However, MREED evaluates RAID-0 system only, in which a basic mechanism of data stripping is applied. To address the generality issue of MREED, we extend the model to study the reliability of two parallel disk systems– RAID-1 and RAID-5– equipped with energy-aware techniques. Secondly, we study impacts of the I/O load skewing technique –gear shifting – in particular on the reliability of PARAID [2], a well known energy-aware data stripping storage system. Finally, we study trade-offs between reliability and energy-saving in the context of PARAID. With the tradeoff curve provided by our experiments, we can offer a helpful suggestion on whether it is necessary to introduce energy-aware mechanism to a storage system.

Existing energy conservation techniques can yield significant energy savings in disks. While several energy conservation schemes like cache-based energy-saving approaches normally have marginal impact on disk reliability, many energy-saving schemes (e.g., dynamic power management and workload skew techniques) inevitably have noticeable adverse impacts on storage systems [3][4]. For example, dynamic power management (DPM) techniques save energy by using frequent disk spin-downs and spin-ups, which in turn can shorten disk lifetime [5][6][7], redundancy techniques [8][9][10][11], workload skew [12][13][14], and multi-speed settings [15][16]. We pay attention on the reliability issue of RAID systems, existing energy conservation techniques can not be applied for RAID systems for the following reasons:

- Conventional RAIDs balance I/O load across all disks in the array for maximized disk parallelisms and performance, meaning that all disks are spinning even under a light load. No opportunity is offered to spin down any of disks;
- Server class disks are not designed for frequent power cycles, which significantly reduce life expectancy;
- Server systems cannot rely on caching and dynamic power management because the servers are too busy to have long idle time.

In this paper, our contributions are summaries as follows:

1) We extend the reliability model MREED for different levels of energy-aware RAID (i.e., RAID-1, which applied disk-mirroring, and RAID-5, which utilized data-parity based on data-stripping techniques.);
2) We introduce Weibull distribution analysis to MREED. Using the utilization of a storage system as an input, we can estimate the annual failure rate (a.k.a, AFR) of this system;
3) We apply MREED to evaluate PARAID, which is a specific energy-aware data-stripping parallel storage system.
4) We evaluate trade-offs between reliability and energy-efficiency in the context of energy-aware RAID systems to justify if it is beneficial to sacrifice reliability for energy efficiency.

We study impacts of the I/O load skewing technique especially on PARAID-1, and PARAID-5 which are energy-aware RAID-1 and RAID-5 systems. Experimental results shows that gear-shifting affects reliability of parallel disks due to two reasons: First, disks working at all gears tend to have high I/O utilization than disks that only works at high gears. Second, disks with high utilization are likely to have high risk of breaking down.

The remainder of this paper is organized as follows. Section II presents the overview of the extended MREED model. In Section III,

978-0-7695-5088-6/13 $31.00 © 2013 IEEE
DOI 10.1109/HPCC.and.EUC.2013.205

IEEE computer society

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:56:00 UTC from IEEE Xplore. Restrictions apply.

we apply MREED model to quantitatively estimate the reliability of PARAID. Section **??** studies tradeoff between reliability and energy-efficiency. Section IV presents experimental results and performance evaluation. In Section V, the related work is discussed. Finally, Section VI concludes the paper with discussions.

II. The MREED Modeling Framework

A. Overview

MREED is a framework developed to model reliability of parallel disk systems employing energy conservation techniques. One critical module in MREED is to model the impact of energy-efficient schemes on the utilization and power-state transition frequency of each disk in a parallel disk system. Another important module developed in MREED is to calculate the annual failure rate of each disk as a function of the disk's utilization, power-state transition frequency. Given the annual failure rate of each disk in the parallel disk system, MREED is able to derive the reliability of an energy-efficient parallel disk system.

Fig. 1 outlines the MREED reliability modeling framework. MREED is composed of a Weibull-based disk reliability model, a system-level reliability model, and three reliability-affecting factors—temperature, power state transition frequency (hereinafter referred to as transition frequency or frequency) and utilization. Many energy-saving schemes inherently affect reliability-related factors like disk utilization and transition frequency. Given an energy optimization mechanism, MREED first converts data access patterns into the two reliability-affecting factors—frequency and utilization. The Weibull-based disk reliability model can derive individual disk's possibility of failure from utilization and power-on hours per year because these parameters are key reliability-affecting factors. Each disk's reliability is used as input to the system-level reliability model that evaluates the annual failure rate of parallel disk systems.

For simplicity without losing generality, we considered in MREED three reliability-related factors, namely: disk utilization, temperature, and power-state transitions. This assumption does not necessarily indicate by any means that there are only three parameters affecting disk reliability. Other factors having impacts on reliability include: handling, humidity, voltage variation, vintage, duty cycle, and altitude [17]. If a new factor has to be taken into account, one can extend the single reliability model by integrating the new factor with other reliability-affecting factors in MREED. Since the infant mortality phenomenon is out the scope of this study, we pay attention to disks that are no less than one year old.

The single-disk reliability can not be accurately described by one valued parameter because the disk drive reliability is affected by multiple factors. There are three major factors that affect disk reliability.

1) Disk Utilization can be characterized as the fraction of active time of a disk drive out of its total powered-on-time. The baseline value (i.e. R_{base} in Eq. 1) of AFR for a disk, which is derived from the Weibull distribution analysis, can be calculated from the disk's utilization. The details will be discussed in the subsection II-B;
2) Temperature, which acts as a multiplier to base failure rates in the MREED model. The temperature factor shown in the Table I was reported by Seagate Storage Group in Longmont, Colorado [18]. From the Table I, we observe that as the temperature rises, the derating factor and the MTBF show clear decreasing. In our research, we will use the Derating Factor (DF) as the Temperature Factor (i.e. τ in Eq. 1) of AFR. For example, at 30°C, the DF value is 0.78, which indicates that the AFR at this temperature is 22% higher than the AFR at 25°C. The main reason that we only use partial data from the report (25°C $\sim$ 46°C) is that we believe the cooling systems will prevent the temperature keeping higher than 46°C for long.
3) Power-State Transition Frequency, which is measured as the number of power-state transition (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. 1 in the next subsection).

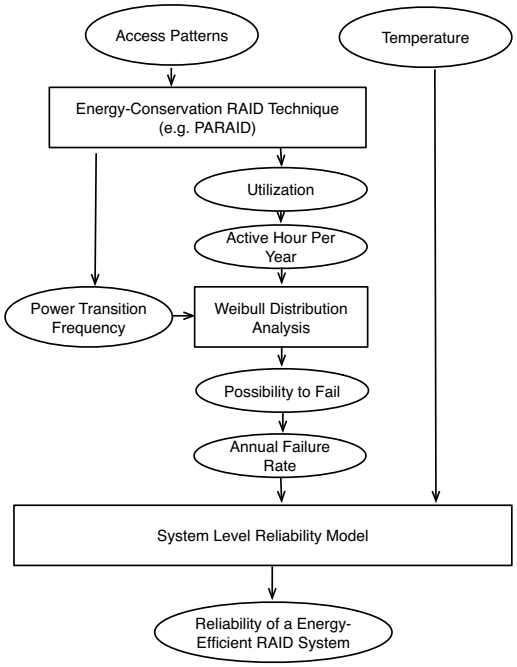


Fig. 1: Overview of the MREED reliability modeling methodology

TABLE I: Temperature Factor

Temperature (°C)	Acceleration Factor	Derating Factor	Adjusted MTBF
25	1.0000	1.00	232,140
26	1.0507	0.95	220,553
30	1.2763	0.78	181,069
34	1.5425	0.65	150,891
38	1.8552	0.54	125,356
42	2.2208	0.45	104,463
46	2.6465	0.38	88,123

Hence, the failure rate R of an individual disk can be expressed as:

$$R = R_{util} * \tau + \alpha * R_{freq} \quad (1)$$

where R_{util} is the baseline failure rate derived from disk utilization, which will be discussed in the next section, τ is the temperature factor, α is a coefficient to reliability R, and R_{freq} is the power-state transition frequency adder to the baseline failure rate, which can be calculated by Eq. 2 [19].

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

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:56:00 UTC from IEEE Xplore. Restrictions apply.

where f is a power-state transition frequency, $R_{freq}(f)$ 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%.

B. Weibull Distribution Analysis for Disk Systems

In order to reveal the relationship between the disk utilization and disk reliability, we introduce a widely referred reliability evaluation tool– Weibull distribution– to exam the R_{util}, which is the baseline failure rate in Eq. 1. Weibull distribution analysis is a leading method in the world for fitting life date. The primary advantage of Weibull analysis is the ability to provide accurate failure analysis and failure forecasts with extremely small samples [20]. It is now widely used reliability engineering and failure analysis including mechanical, electronic, materials, and human failures [21]. The Weibull reliability function describes the probability of survival as a function of time, and is described as follows in Eq. 3:

$$\begin{aligned} R(t) &= \int_t^{\infty} \frac{\beta(x)^{(\beta-1)}}{\theta^{\beta}} \exp[-(\frac{x}{\theta})^{\beta}]dx \\ &= \exp[-(\frac{t-\beta}{\theta})^{\beta}] \end{aligned} \tag{3}$$

where β is the shape parameter or slope parameter ($0 < \beta < \infty$), and θ is the scale parameter or characteristic life ($0 < \theta < \infty$). The values of these two parameters depend on types of hard disk drives, which can be obtained either by disk manufactures or by mathematical means such as parameter estimation. Given a disk drive's total power-on hours per year, and the disks utilization according to the power management which is taken into use, we can calculate its total active hours during one year by Eq. 4

$$T_{active} = T_{power_on} * \rho \tag{4}$$

where ρ is a disk utilization. With active hours as an input along with β and θ, we can use Eq. 3 to estimate its annual failure rate and MTBF (which serves as R_{util} in Eq. 1).

However, it is hard to determine accurate value of active time of a disk drive due to the following reasons:

- It is hard to monitor the active time in the system. Even though there are some system functions, for instances *nmon* and *iotop* in linux, can show a display of disk activity per device by average, values presented by these functions can not represent the real-time situation in the systems;
- A hard disk drive's I/O access time consist of rotational delay and seek time in addition to the transfer time, among which only the transfer time should be counted as the active time, while the rest of two need to be treated as portion in total time. Extraction of transfer time from the total access time increasing the complexity of the monitoring;
- As power management schemes are involved in the parallel storage system, disks status will be switched between stop and active more frequently than usual, which makes it even harder to determine the active time of a disk.

In order to obtain the disk utilization, we consider size of files, number of files transferred in a time unit, and the bandwidth of a disk instead. The formula is as follows:

$$\rho = A \times X \tag{5}$$

where A is the transfer time, X is the number of data transfer per time unit. For example, a disk with 0.03 second access time and an average of 10 transfer requests per second, the utilization of this disk in a second is $\rho = 0.03 * 10 = 30\%$. As $A = \frac{File\ Size}{Disk\ Bandwidth \times Time\ Unit}$ and $X = \frac{Number\ of\ Data\ Transfer}{Time\ Unit}$, Eq. 5 is written as:

$$\rho = \frac{S_{file} * N_{data}}{B * T^2} \tag{6}$$

where S_{file} denotes the size of data (usually in MB), N_{data} is the number of data to be transferred, B is the disk I/O bandwidth (usually in MB per second, or MB/s) and T denotes the time unit (usually in seconds). In a parallel storage system consists of n disks, the system bandwidth is n times of single disk I/O bandwidth.

In order to examine the system utilization and its reliability accordingly in a more detailed way, we study a particular RAID system which applied energy-efficient mechanisms–PARAID [2]–as an example in the next section.

III. Reliability Model for PARAID

A. Background

Different from traditional disk array systems, RAID balances the load across all disks in the array for maximized disk parallelism and performance [22]. In a RAID system, all disks are spinning even under a light load. Instead of spinning down inactive disks under a light load as MAID [23] or PDC [24] behave, PARAID exploits unused storage to replicate and stripe data blocks in a skewed fashion, so that disks can be organized into hierarchical overlapping sets of RAIDs. Each set contains a different number of disks, and can serve all requests via either its data blocks or replicated blocks. PARAID introduces a skewed striping pattern that allows RAID devices to use just enough disks to meet the system load. Each set is analogous to a gear in automobiles as PARAID has aggregated disk bandwidth. PARAID varies the number of powered-on disks via *gear-shifting* among sets of disks to reduce power consumption [2]. The authors confirmed that PARAID system can save up to 34% energy compared to the conventional 5-disk RAID system. However, such energy-efficient technique may have adverse impacts on the reliability of the storage system. The system has to spend extra disks utilization on copying data from disks that are about to be spun down, which leads to higher risk of system failures. Furthermore, after a gear-shifting down, less number of disks will provide the same amount of service as it is before the gear-shifting, which pushes the power-on disks into higher utilization range and thus makes the system even less reliable. Thirdly, due to the data stripping technique, each single disk in the PARAID system only holds part of files. PARAID may face absolute data lose if the number of failure disks exceeds the system's failure tolerance. The reliability issue of PARAID counts much more than conventional disk array systems. Fig. 2 is a PARAID system consists

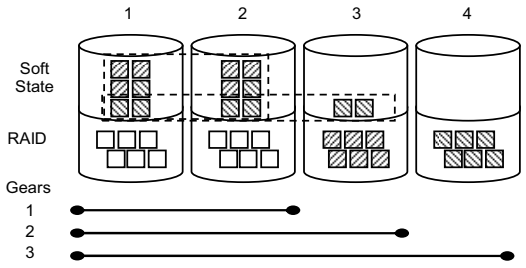


Fig. 2: Framework of PARAID: skewed striping of replicated blocks in soft state, creating 3 RAID gears over 4 disks [2]

of four disks. Fig. 2 shows that each disk in PARAID has two separate states– the Soft State and RAID State. When operating in gear 3, with

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:56:00 UTC from IEEE Xplore. Restrictions apply.

all four disks powered, PARAID works as the way of conventional RAID system offering maximized disk parallelism and performance accordingly. As I/O load decreases, PARAID down-shifts into gear 2 by spinning down the fourth disk. Before the down-shifting, the blocks stored in the RAID states on disk 4 are copied to disk 1∼3 one by one. In this case, disk 1 holds the $1st$ and the $4th$ block of disk 4, disk 2 keeps the $2nd$ and the $5th$ block of disk 4, and disk 3 will store the $3rd$ and the $6th$ block of disk 4. If the load keeps decreasing, PARAID will further down-shift into gear 1 by powering down the third disk.

B. Modeling Utilization of Disks in PARAID

Recall that the annual failure rate of each disk can be calculated using utilization, operating temperature as well as power-state transition frequency. To model reliability of a disk array equipped with PARAID, we have to first address the issue of modeling disk utilization used to calculate base annual failure rates (R_{util} in Eq.1 shown in Section II). In this subsection, we develop a utilization model capturing behaviors of a RAID-based energy efficient disk array. The utilization model takes file access patterns as an input and calculates the utilization of each disk in the disk array.

Since a PARAID system requires minimum number of disks to achieve a fundamental I/O parallelism, the PARAID system consists of n disks has $(n-i)$ gears to shift, where the value of i depends on the RAID setups. For example, RAID-0 requires at least two disks to meet the data stripping; RAID-1 requires at least two disks same as RAID-0 does, however, due to the mirroring mechanism, the number of disks in RAID-1 should be even; and RAID-5 requires minimum three disks to achieve data parity mechanism for data recovery purpose. Hence, $i = 1$ for PARAID-0, $i = n/2$ for PARAID-1, and $i = 2$ for PARAID-5. From Eq. 6 we observe that the calculation of utilization is fully rely on three values: the number of data to be transferred (N_{data}), the size of data (S_{file}), and the time lapse to transfer the data (T). As the size of data and the time period for data transfer are determined by file access patterns which act as inputs to our model, the N_{data}, value of which is affected by the number of active disks and energy saving techniques, is the major concern in this section. Since the PARAID-0 was carefully examined by the original MREED [1], we extend MREED for PARAID-1 and PARAID-5 evaluation only. The details of modeling utilization of disks in PARAID-1 and PARAID-2 will be covered by the following two sub-sections.

1) PARAID-1: Power-Aware RAID with Mirroring: Recall a classic RAID-1 system contains two disks over a single disk. Since each member contains a complete copy and can be addressed independently, ordinary wear-and-tear reliability is raised by the power of the number of self-contained copies. In order to keep this property, the number of disks in a RAID-1 system must be even, half of which server as data storage while the rest half are mirroring disks. Assume that at the $G_{\frac{N}{2}}$ gear, in which case all n disks of the system are kept spinning in order to offer the maximum reliability, each single disk stores m blocks. When a pair of disks N_{data} and N_{mirror} are spun down, all m data blocks along with their m replica blocks will be separated into $\frac{n-2}{2}$ sets in a way that each of the rest $n-2$ disks will handle making replicas for the two spun down disks. Since the all $2m$ blocks contains one copy of replica, there are two ways to handle data copying. The first way is that all m data blocks are copied from the disk N_{data} to the rest of data disks while all m mirror data blocks are copied from disk N_{mirror} to the rest of active mirror disks. The second solution is to copy the m data blocks on disk N_{data} to the rest of data disks and the rest of mirror disks while spinning down disk N_{mirror} without any data transfer. We consider the former solution in this section for the reason that such method achieves high data migration I/O performance. Furthermore, as the capacity of single disk keeps increasing much sharper than transfer rate, the second solution may waste too much time on data migration. Thus, we have:

$$F_{out_G_{(\frac{N}{2})(\frac{N-2}{2})}} = 2m \tag{7}$$

and

$$F_{in_G_{(\frac{N}{2})(\frac{N-2}{2})}} = \frac{F_{out_G_{(\frac{N}{2})(\frac{N-2}{2})}}}{n-2} \tag{8}$$

where $F_{out_G_{(\frac{N}{2})(\frac{N-2}{2})}}$ represents replicas of the blocks moved out from the disk pair N_{data} and N_{mirror} when PARAID shifts down the gear from $G_{\frac{N}{2}}$ to $G_{\frac{N-2}{2}}$ due to the decreasing workload. $F_{in_G_{(\frac{N}{2})(\frac{N-2}{2})}}$ represents the set of replicated blocks that moved into each of the $n-2$ disks.

Similarly, when PARAID shifts down from gear $G_{\frac{N-2}{2}}$ to $G_{\frac{N-4}{2}}$, we have:

$$F_{out_G_{(\frac{N-2}{2})(\frac{N-4}{2})}} = 2m + F_{in_G_{(\frac{N}{2})(\frac{N-2}{2})}} \tag{9}$$

and

$$F_{in_G_{(\frac{N-2}{2})(\frac{N-4}{2})}} = \frac{F_{out_G_{(\frac{N-2}{2})(\frac{N-4}{2})}}}{n-4} \tag{10}$$

It is noticed that the disk to be powered off needs to duplicate blocks, which were copied during the first downshifting period of time, apart from its own m blocks. The rest $n-4$ disks move in more replicated blocks accordingly.

In general, when PARAID shifts down from gear G_j to G_{j-1}, where $j \in (2, ..., \frac{n}{2})$, the number of blocks that the disk to be powered off must handle the following number of reads copy out is

$$\begin{aligned} F_{out_G_{(j)(j-1)}} =& 2m + F_{in_G_{(\frac{N}{2})(\frac{N-2}{2})}} + F_{in_G_{(\frac{N-2}{2})(\frac{N-4}{2})}} + \\ &+ ... + F_{in_G_{(j+1)(j)}} \end{aligned} \tag{11}$$

while the number of blocks that must be written to the rest $2(j-1)$ disks is expressed as:

$$F_{in_G_{(j)(j-1)}} = \frac{F_{out_G_{(j)(j-1)}}}{2(j-1)} \tag{12}$$

where j represents the current gear number while $(j-1)$ indicated the gear number that the PARAID system is about to be shifted to. We assume that every single file has the same number of blocks, each of which has the same size. Hence, the I/O time for accessing each single block is the same. Now we can formally express the utilization of disk i in time phase l as follows:

For the disk to be power-off, we have:

$$\rho_{power-off} = \rho_{I/O} + \rho_{read} \tag{13}$$

, while for the rest of disks, we have:

$$\rho_{power-on} = \rho_{I/O} + \rho_{read} \tag{14}$$

Combination with Eq. 6, the above two equations can be expressed as:

$$\begin{aligned} \rho_{power-off} &= \frac{S_{file} \times (N_{data} + F_{out_G_{(j)(j-1)}})}{B \times l^2} \\ \rho_{power-on} &= \frac{S_{file} \times (N_{data} + F_{in_G_{(j)(j-1)}})}{B \times l^2} \end{aligned} \tag{15}$$

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:56:00 UTC from IEEE Xplore. Restrictions apply.

2) PARAID-5: Power Aware RAID with Data Parity: Compared to RAID-0, RAID-1 benefits high system reliability from data replica techniques. However, such benefit is not a free lunch, RAID-1 sacrificed half of storage capacity and the other side effects including additional storage energy-consumption. RAID-5 configuration is built to balance the reliability, I/O performance and the data capacity. RAID 5 incorporates striping of data just like in a RAID 0 array, however, in a RAID 5 there are redundant pieces of the data that are also distributed across the drives and are referred to as parity. Having the parity blocks staggered across each drive allows any single drive in the RAID 5 array to fail without any data loss. The minimum number of disks in a RAID-5 system is three.

Assume that at the G_{N-2} gear, in which case all n disks of the system are kept spinning in order to offer the maximum reliability, each single disk stores m blocks plus one parity block for data recovery. When disk N is spun down, all m data blocks will be separated into $n-1$ sets in a way that each of the rest $n-1$ disks will handle making replicas for the two spun down disks. Furthermore, all parity data that resides on active disks will be updated after the m blocks are distributively copied. This procedure consists of three major steps: data is read from disks to processors, processors compute new parity data vis XOR algorithm, and the updated parity blocks are written back to disks. We only consider the I/O portion of the procedure in this paper. Thus, we have:

$$F_{out_G_{(N-2)(N-3)}} = m \tag{16}$$

and

$$F_{in_G_{(N-2)(N-3)}} = \frac{F_{out_G_{(N-2)(N-3)}}}{n-1} \tag{17}$$

where $F_{out_G_{(N-2)(N-3)}}$ represents replicas of the blocks moved out from the disk N when PARAID shifts down the gear from G_{N-2} to G_{N-3} due to the decreasing workload. $F_{in_G_{(N-2)(N-3)}}$ represents the set of replicated blocks that moved into each of the $n-1$ disks. Furthermore, different from what RAID-0 or RAID-1 does, RAID-5 have to update parity data for recovery purpose every time when information on any disk changed. During the update process, all changed blocks will be read for processors to compute new parity data. Since each disk in a RAID-5 system contains a parity data, when one of the disk needs to be powered off for the as workload is relatively low, parity data that distributed on all disks should be updated. The total number of parity data depends on the number of blocks a disk stores–each data has one corresponding parity data. In our case, there should be in total m parities and each disk stores $\frac{m}{n}$ of them when all disks are active. In the case of one disk is being spun down, we have following formulas calculating I/O workload for parity data updating in addition to blocks movement:

$$F_{pa_G_{(N-2)(N-3)}} = \frac{m}{n} \tag{18}$$

where $F_{pa_G_{(N-2)(N-3)}}$ represents the set of parity blocks that are moved into each of the $n-1$ disks. After the shifting, each of the active $n-1$ disks holds original m blocks plus $\frac{m}{n-1}$ blocks that are copied from the power-off disk N, in addition to $\frac{m/n}{(n-1)}$ parity blocks. Hence, at this step there should be $m + \frac{m}{n-1} + \frac{m}{n(n-1)}$ parity blocks accordingly.

Similarly, when PARAID shifts down from gear G_{N-3} to G_{N-4}, we have:

$$F_{out_G_{(N-3)(N-4)}} = m + F_{in_G_{(N-3)(N-4)}} \tag{19}$$

and

$$F_{in_G_{(N-3)(N-4)}} = \frac{F_{out_G_{(N-3)(N-4)}}}{n-2} \tag{20}$$

As aforementioned, each disk holds $m + \frac{m}{n-1} + \frac{m}{n(n-1)}$ parity blocks, we have:

$$F_{pa_G_{(N-3)(N-4)}} = (m + \frac{m}{n-1} + \frac{m}{n(n-1)})/(n-1) \tag{21}$$

In general, when PARAID shifts down from gear G_j to G_{j-1}, where $j \in (2, ..., \frac{n}{2})$, the number of blocks that the disk to be powered off must handle the following number of reads copy out is

$$\begin{aligned} F_{out_G_{(j)(j-1)}} =& m + F_{in_G_{(N-2)(N-3))}} + F_{in_G_{((N-3)(N-4))}} + \\ &+ ... + F_{in_G_{(j)(j-1)}} \end{aligned} \tag{22}$$

while the number of blocks that must be written to the rest $(j-1)$ disks along with the number of parities are expressed as:

$$F_{in_G_{(j)(j-1)}} = \frac{F_{out_G_{(j)(j-1)}}}{(j-1)} \tag{23}$$

$$F_{pa_G_{(N-3)(N-4)}} = (m + F_{in_G_{(j)(j-1)}} + \frac{m}{n(n-1)})/(j) \tag{24}$$

where j represents the current gear number while $(j-1)$ indicated the gear number that the PARAID system is about to be shifted to. We assume that every single file has the same number of blocks, each of which has the same size. Hence, the I/O time for accessing each single block is the same. Now we can formally express the utilization of disk i in time phase l as follows:

For the disk to be power-off, we have:

$$\rho_{power-off} = \rho_{I/O} + \rho_{read} \tag{25}$$

, while for the rest of disks, we have:

$$\rho_{power-on} = \rho_{I/O} + \rho_{read} \tag{26}$$

Combination with Eq. 6, the above two equations can be expressed as:

$$\begin{aligned} \rho_{power-off} &= \frac{S_{file} \times (N_{data} + F_{out_G_{(j)(j-1)}} + F_{pa_G_{(j)(j-1)}})}{B \times l^2} \\ \rho_{power-on} &= \frac{S_{file} \times (N_{data} + F_{in_G_{(j)(j-1)}} + F_{pa_G_{(j)(j-1)}})}{B \times l^2} \end{aligned} \tag{27}$$

IV. Reliability Evaluation

A. Experimental Setup

We developed a simulator in which the PARAID-10 and PARAID-5 systems (a.k.a Power-Aware RAID-10 and RAID-5) are implemented. Table II shows the parameters of configurations for PARAID-10 and PARAID-5. We evaluate the reliability of a six-disk PARAID-5 system, in which the highest gear of the system is 4 and the capacity of the system is the capacity summary of five disks as RAID-5 sacrifices one disk for data recovery. As the RAID-10 requires duplicates, the number of disks in RAID-10 should be even. For comparison purpose, we evaluate the reliability of a ten-disk PARAID-10 system so that it retains the same capacity as a six-disk PARAID-5 system. As a RAID-5 system requires parity data for information recovery purpose, the system needs at least three disks. Hence, in order to keep the RAID-10 and RAID-5 configurations, the number of disks kept active at the lowest gear one are set to two for RAID-10 and three for RAID-5.The number of file accessed within time unit, or hit rate is generated by Poisson distribution. The operating temperature is set to 38°C. Furthermore, we are using properties of Seagate hard disk drive in our simulator. The properties are also shown in Table II. Since Seagate's disks properties are introduced to our experimental

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:56:00 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Experiment Parameter Setup

Disk Type	SEAGATE ST3146855FC
Capacity	146 GB
Cach Size	SATA 16MB
Buffer to Host Transfer Rate	4Gb/s) (MAX)
Total Number of Disks	6(PARAID-5), 10(PARAID-10)
File Size	100 MB
Number of Files	1000
Synthetic Trace	Poisson Distribution
Time Period	24 hours
Interval Time (Time Phase)	1 hour
Power On Hour Per Year	8760

setups, we set $\beta = 0.55$, $\theta = 8410332$ in the Weibull analysis model, and 0.54 as the τ, which is the temperature factor in Eq. 3 [18].

The PARAID-10 is investigated firstly. Fig. 3 represents the utilization comparison of PARAID-10 and RAID-10 under low average access rate through a whole day. The curves provides a similar trend as the RAID-0. The major reason is that a RAID-10 is a mirrored RAID-0, hence shares the similar system workload with RAID-0.

We then investigate the impacts of hit rate (λ in Poisson distribution) on utilization of PARAID-5. We set values of utilization to trigger gear-shifting are set to 60% for gear up while 30% for gear down. The PARAID-5 is assumed to be started at the top gear– all five disks are working . Fig. 4 plots the utilization comparison of PARAID-5 and RAID-5 within 24 hours. The average hit rate is set to 20 per hour ($\lambda = 20$), which is relatively low. We observe from Fig. 4 that as time goes, the utilization of RAID-5 stays stable around 18%, while that of PARAID-5 increases three times and decreases once then stays stable around 52%. Those three increasing points are caused by the gear-down shiftings hence the decreasing the number of active disks. We pay much attention to the gear shifts down to the least gear, in which system has minimum three active disks. The system utilization is sharply increased to over 80% due to a large amount of parity data updating. Such large amount of parity data updating can not be handled by three active disks within next couple of hours and triggers a gear up-shifting. Even though the utilization of PARAID-5 is more than 60% higher than that of RAID-5 at the end hour 24, the energy consumption of PARAID-5 is more than 30% lower than that of RAID-5 since there are only four active disks by then.

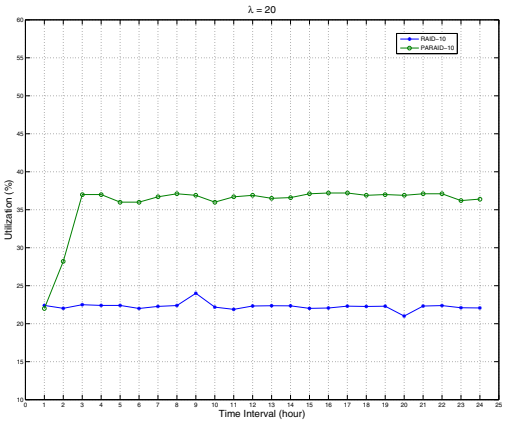


Fig. 3: Disks Utilization Comparison Between PARAID-10 And RAID-10 at A Low Access Rate(20 times per hour).

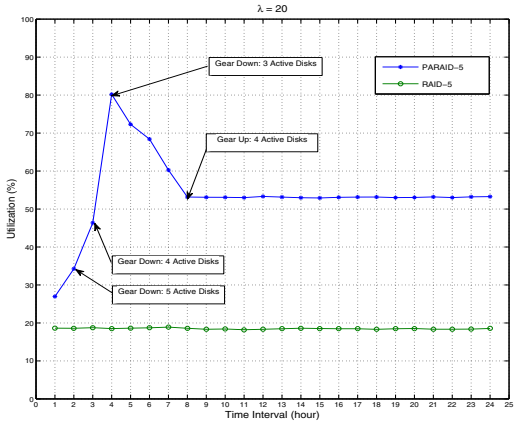


Fig. 4: Disks Utilization Comparison Between PARAID-5 And RAID-5 at A Low Access Rate(20 times per hour).

Fig. 5 presents the AFR of PARAID-10 and RAID-10. Based on the vague difference of utilization that is derived from Fig. 3, each half of the RAID-10 shares similar AFR trend with RAID-0, however, the entire system benefits much better AFR due to its mirroring mechanism. In order to gain such high reliability, RAID-10 sacrifices storage capacity and energy efficiency. In general, RAID-10 consumes twice as much energy as RAID-0 does to store the same capacity of data as what RAID-0 does.

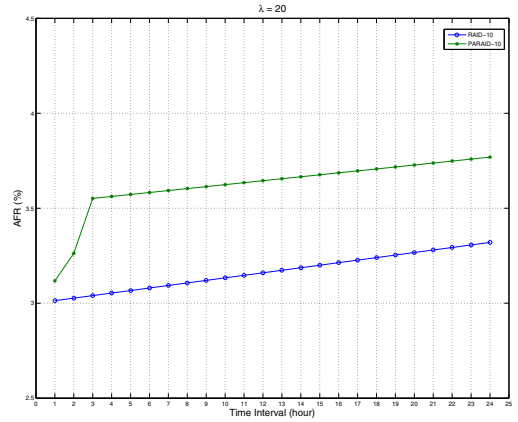


Fig. 5: AFR Comparison Between PARAID-10 And RAID-10 at A Low Access Rate(20 times per hour).

Fig. 6 illustrates the annual failure rates (AFR) of PARAID-5 and RAID-5 based on their utilization which is derived from Fig. 4. Results plotted in Fig. 6 show that AFR values of RAID-5 keeps increasing from 0.89% to 5.1% when hour lapses, while AFR of PARAID-5 increases much faster than that of RAID-5. We attribute this trend to the decreasing of the number of active disks due to gear-down shiftings. We also observe that from the hour 1 to hour 4, the AFR of PARAID-5 increases sharply, which is mainly because of the additional utilization overhead by parity data updating. Furthermore, the system's continuously shifting down gears causes more parity data needs to be updated one less number of active disks, which leads to utilization surge (shown in Fig. 4) hence affects AFR accordingly.

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:56:00 UTC from IEEE Xplore. Restrictions apply.

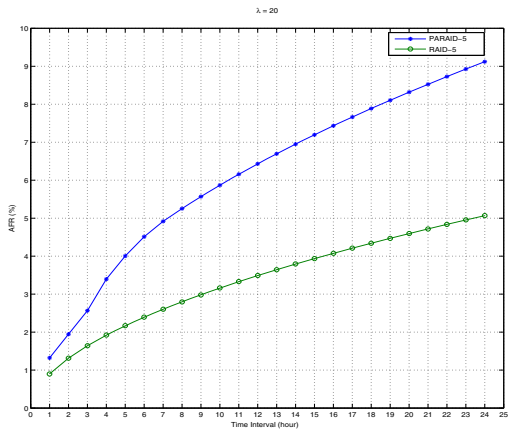


Fig. 6: AFR Comparison Between PARAID-5 And RAID-5 at A Low Access Rate(20 times per hour).

If the average hit rate is set to 80 per hours, ($\lambda = 80$), the utilization difference between PARAID-5 and RAID-5 is very vague. It is mainly because that when the hit rate is relatively high, the high utilization of PARAID-5 will not trigger any gear-shifting mechanism. Hence the AFR difference between PARAID-5 and RAID-5 is ignored. We observe that there is not point to apply any energy-saving mechanism to a system which is busy with services.

V. Related Work

A hard disk drive (HDD) is a complex dynamic system made up of various electrical, electronic, and mechanical components [25]. 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 [26][7] and multi-speed disks [6]), the operating system level (e.g., power-aware caching/prefetching [9][16]), and the application level (e.g. software DMP [27] and cooperative I/O [28]). 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 [6][15][29]. 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 [24][23][30][2].

Recent studies show that both power management and workload skew schemes inherently impose adverse impacts on disk systems [3][4]. 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 [31], which inherently have higher risk of breaking down than other disks usually being kept on standby. Disks that store 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 MREED is orthogonal to the aforementioned energy saving studies, because MREED 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 modeling. Although disk drive manufacturers claim that MTBF of most disks is more than 1 million hours [32], users have experienced a much lower MTBF from their field data [17]. 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 [17].

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 [33]; Elerath and Pecht proposed a flexible model for estimating reliability of RAID storage [34]; and Xin *et. al* developed a model to study disk infant mortality [35]. Unlike these reliability models tailored for conventional parallel and distributed disk systems, our MREED model pays special attention to reliability of parallel disk systems coupled with energy-saving mechanisms.

Very recently,Sivathanu [36] proposed a framework for evaluating techniques for power optimization in storage and used it to evaluate PARAID. Our MREED approach differs itself from theirs in the sense that the goal of MREED is to evaluate reliability of disk systems by modeling the behavior of parallel disks where energy conservation mechanisms are integrated.

VI. Conclusion

This paper extends a reliability model named MREED to study the reliability of energy-efficient parallel disk systems on different levels. The former MREED only evaluates energy-efficient RAID-0 which achieves maximum parallelism while retaining the worst system reliability– any of the disk failure leads to the entire system dysfunction. To strengthen the generosity of MREED, this model is extended so that it can evaluate most common used levels of RAID systems with energy saving mechanisms, systems of which includes RAID-1 which achieves higher reliability by the mirroring mechanism and RAID-5 which meets higher disks I/O performance while retaining reliability by the parity technique. In particularly, MREED is applied to study systems reliability with the PARAID technique, a.k.a. power-aware RAID. Experimental results reveal the affection of data locality and data recovery techniques on the reliability of power-aware parallel disk systems. Furthermore, we investigate a fundamental trade-off between reliability and energy-efficiency in the context of energy-efficient RAID systems. A tradeoff curve is presented as a unified framework to justify whether or not it is wise to trade reliability for high energy efficiency.

Future directions of this research can be performed in the following. First of all, we will evaluate and compare an array of energy saving techniques with respect to specific application domains, as we noticed during experiments that lifetime of storage systems defers towards different applications being implemented on the systems. Second. we will provide a system reliability forecast schemes. A storage system healthy curve will be used as a unified framework to justify if it is necessary to exchange some almost-wear-out disks drives from the system. Last, we will provide a energy-efficient storage solution that aims at retaining better storage reliability.

Acknowledgment

The work reported in this paper was supported by the US National Science Foundation under Grants CCF-0845257 (CAREER),

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:56:00 UTC from IEEE Xplore. Restrictions apply.

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 work was supported by the Hunan University under Start-up Grants, Hunan Science Foundation and National Science Foundation China under Grants 61370095.

REFERENCES

[1] S. Yin, Y. Tian, J. Xie, X. Qin, M. Alghamdi, X. Ruan, and M. Qiu, "Reliability analysis of an energy- aware raid system," in *Performance Computing and Communications Conference (IPCCC), 2011 IEEE 30th International*, nov. 2011, pp. 1 –8.

[2] C. Weddle, M. Oldham, J. Qian, A.-I. A. Wang, P. Reiher, and G. Kuenning, "Paraid: a gear-shifting power-aware raid," in *FAST '07: Proceedings of the 5th USENIX conference on File and Storage Technologies*. Berkeley, CA, USA: USENIX Association, 2007, pp. 30–30.

[3] 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.

[4] T. Xie and Y. Sun, "Sacrificing reliability for energy saving: Is it worthwhile for disk arrays?" April 2008, pp. 1–12.

[5] F. Douglis, P. Krishnan, and B. Marsh, "Thwarting the power-hungry disk," in *Proc. USENIX Winter 1994 Technical Conf.*, 1994, pp. 23–23.

[6] D. P. Helmbold, D. D. E. Long, T. L. Sconyers, and B. Sherrod, "Adaptive disk spin—down for mobile computers," *Mob. Netw. Appl.*, vol. 5, no. 4, pp. 285–297, 2000.

[7] K. Li, R. Kumpf, P. Horton, and T. Anderson, "A quantitative analysis of disk drive power management in portable computers," in *WTEC'94: Proceedings of the USENIX Winter 1994 Technical Conference on USENIX Winter 1994 Technical Conference*. Berkeley, CA, USA: USENIX Association, 1994, pp. 22–22.

[8] E. Pinheiro, R. Bianchini, and C. Dubnicki, "Exploiting redundancy to conserve energy in storage systems," in *Proc. Joint Int'l Conf. Measurement and Modeling of Computer Systems*, 2006.

[9] Q. Zhu, F. M. David, C. F. Devaraj, Z. Li, Y. Zhou, and P. Cao, "Reducing energy consumption of disk storage using power-aware cache management," in *HPCA '04: Proceedings of the 10th International Symposium on High Performance Computer Architecture*. Washington, DC, USA: IEEE Computer Society, 2004, p. 118.

[10] J. Wang, H. Zhu, and D. Li, "eraid: Conserving energy in conventional disk-based raid system," *IEEE Transactions on Computers*, vol. 57, no. 3, pp. 359–374, 2008.

[11] 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.

[12] A. E. Papathanasiou and M. L. Scott, "Power-efficient server-class performance from arrays of laptop disks," 2004. [Online]. Available: http://hdl.handle.net/1802/314

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

[14] Q. Yang and Y. Hu, "Dcd — disk caching disk: A new approach for boosting i/o performance," May 1996, pp. 169–169.

[15] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, "Drpm: dynamic speed control for power management in server class disks," June 2003, pp. 169–179.

[16] S. W. Son and M. Kandemir, "Energy-aware data prefetching for multi-speed disks," in *CF '06: Proceedings of the 3rd conference on Computing frontiers*. New York, NY, USA: ACM, 2006, pp. 105–114.

[17] J. Elerath, "Specifying reliability in the disk drive industry: No more mtbf's," 2000, pp. 194–199.

[18] G. Cole, "Estimating drive reliability in desktop computers and consumer electronics systems." Seagate Personal Storage Group, 2000.

[19] S. Yin, X. Ruan, A. Manzanares, and X. Qin, "How reliable are parallel disk systems when energy-saving schemes are involved?" in *Proc. IEEE International Conference on Cluster Computing (CLUSTER)*, 2009.

[20] R. B. Abernethy, *The New Weibull Handbook 5th edition*. Humble, TX, USA: Barringer & Associates, Inc, 2010.

[21] B. Dodson, *Weibull Analysis*. Milwaukee, WI,USA: ASQC Quality Press, 1994.

[22] D. A. Patterson, G. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (raid)," in *SIGMOD '88: Proceedings of the 1988 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1988, pp. 109–116.

[23] D. Colarelli and D. Grunwald, "Massive arrays of idle disks for storage archives," in *Proc. ACM/IEEE Conf. Supercomputing*, 2002, pp. 1–11.

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

[25] J. Yang and F.-B. Sun, "A comprehensive review of hard-disk drive reliability," in *Proc. Annual Reliability and Maintainability Symp.*, 1999.

[26] F. Douglis, P. Krishnan, and B. Marsh, "Thwarting the power-hungry disk," in *WTEC'94: Proceedings of the USENIX Winter 1994 Technical Conference on USENIX Winter 1994 Technical Conference*. Berkeley, CA, USA: USENIX Association, 1994, pp. 23–23.

[27] 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.

[28] A. Weissel, B. Beutel, and F. Bellosa, "Cooperative i/o: a novel i/o semantics for energy-aware applications," in *OSDI '02: Proceedings of the 5th symposium on Operating systems design and implementation Due to copyright restrictions we are not able to make the PDFs for this conference available for downloading*. New York, NY, USA: ACM, 2002, pp. 117–129.

[29] P. Krishnan, M. P. Long, and S. J. Vitter, "Adaptive disk spindown via optimal rent-to-buy in probabilistic environments," Durham, NC, USA, Tech. Rep., 1995.

[30] E. Pinheiro, R. Bianchini, E. Carrera, and T. Heath, "Load balancing and unbalancing for power and performance in cluster-based systems," *Proc. Workshop on Compilers and Operating Systems for Low Power*, September 2001.

[31] A. Manzanares, X. Ruan, S. Yin, and M. Nijim, "Energy-aware prefetching for parallel disk systems: Algorithms, models, and evaluation," *IEEE Int'l Symp. on Network Computing and Applications*, 2009.

[32] B. Schroeder and G. 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.*, 2007, p. 1.

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

[34] J. Elerath and M. Pecht, "Enhanced reliability modeling of raid storage systems," in *Proc. IEEE/IFIP Int'l Conf. Dependable Sys. and Networks*, 2007.

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

[36] S. Sivathanu, L. Liu, and C. Ungureanu, "Modeling the performance and energy of storage arrays," in *Proceedings of the International Conference on Green Computing*, ser. GREENCOMP '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 229–242. [Online]. Available: http://dx.doi.org/10.1109/GREENCOMP.2010.5598308

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:56:00 UTC from IEEE Xplore. Restrictions apply.