Reliability Analysis of An Energy-Aware RAID System

Shu Yin, Yun Tian, Jiong Xie, and Xiao Qin*
Department of Computer Science and Software Engineering
Auburn University, Auburn, AL 36849
Email: {szy0004, tianyun, jzx0009, and xqin}@auburn.edu

Mohammed Alghamdi
Department of Computer Science
Al-Baha University, Al-Baha City, Kingdom of Saudi Arabia
Email: mialmushilah@bu.edu.sa

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

Meikang Qiu
Department of Electrical and Computer Engineering
University of Kentucky, Lexington, Kentucky, 40506
Email: mqiu@engr.uky.edu

***Abstract*—We develop a mathematical model– MREED– to quantitatively evaluate the failure rate of energy-efficient parallel storage systems. The Power-Aware Redundant Array of Inexpensive Disk (PARAID) aims to reduce energy use of commodity server-class disks without specialized hardware. The goal of PARAID is to skewed striping pattern to adapt to the system load by changing the number of powered disks. By spinning down disks during light workloads, PARAID can reduce power consumption, while still meeting performance demands. We show that MREED can be used to estimate a five-disk PARAID-0 system. We validate the accuracy of MREED using the DiskSim simulator. Our approach shows that MREED can rely on file access pattern to estimate system utilization correctly. Furthermore, even thought PARAID may achieve reasonable reliability, our model shows that PARAID's reliability is affected by data locality.**

***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. In this paper, we first study the reliability of a parallel disk system equipped with the PARAID [1] technique by employing the **M**athematical **R**eliability model for **E**nergy- **E**fficient RAI**D** system called MREED. As a mathematical model, MREED shows its advantage of presenting the reliability trend of energy-aware storage systems. However, it is challenging to validate the MREED model. To address the correctness issue of MREED, we validate the access-rate-utilization model, which converts file access rate to utilization of the storage system, in MREED. Finally, we study impacts of the I/O load skewing technique –gear shifting – on the reliability of PARAID, a well known energy-aware data stripping 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 [2][3]. 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 [4][5][6], redundancy techniques [7][8][9][10], workload skew [11][12][13], and multi-speed settings [14][15]. 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 propose a reliability model MREED for Power-Aware RAID (i.e., an energy aware data-stripping parallel storage system);
2) We introduce Weibull distribution analysis to MREED. Using the utilization of a storage system as an input, we can estimate and forecast the annual failure rate (a.k.a, AFR) of this system;
3) We validate the access-rate-utilization model of MREED;
4) We study the impacts of the gear-shifting schemes on the reliability of PARAID.

We study impacts of the I/O load skewing technique especially on PARAID-0, which is an energy-aware RAID-0 system. 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 MREED model. In Section III, we apply MREED model to quantitatively estimate the reliability of PARAID. Secion IV demonstrates a solution to validate access-rate-utilization model in MREED. Section V presents experimental results and performance evaluation. In Section VI, the related work is discussed. Finally, Section VII 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. In the MREED framework, we evaluate the reliability impacts of a specific energy-saving technique - the Power-Aware RAID. 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. As such, we used MREED to study the reliability of a parallel disk system equipped with the PARAID technique.

978-1-4673-0012-4/11/$26.00 ©2011 IEEE

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 (e.g., PARAID [1]), 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 [16]. 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.

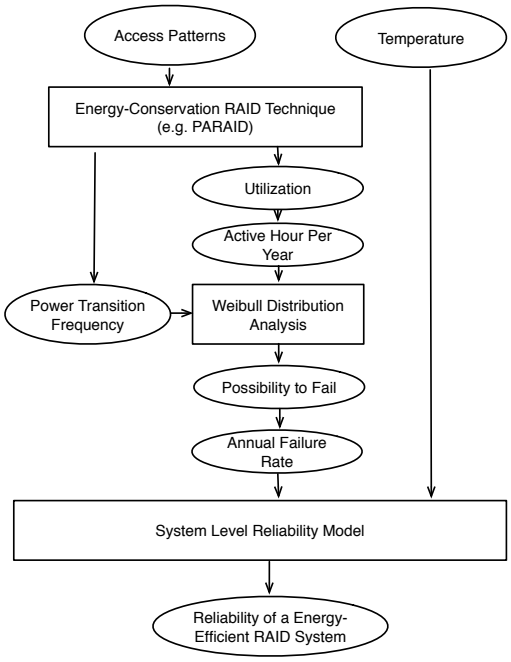

Fig. 1: Overview of the MREED reliability modeling methodology

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_Value} 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 [17]. 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. $TemperatureFacotr$ 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.

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

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).

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

$$R = R_{Base\_Value} * \tau + \alpha * R_{Frequency\_Adder} \quad (1)$$

where R_{Base_Value} is the baseline failure rate derived from disk utilization, τ is the temperature factor, α is a coefficient to reliability R, and $R_{Frequency\_Adder}$ is the power-state transition frequency adder to the baseline failure rate, which can be calculated by Eq. 2 [18].

$$R(f) = 1.51e^{-6}f^2 - 1.09e^{-5}f + 1.39e^{-2}, f \in [0, 500] \quad (2)$$

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%.

B. Weibull Distribution Analysis

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 [19]. It is now widely used reliability engineering and failure analysis including mechanical, electronic, materials, and human failures [20]. The Weibull reliability function describes the probability of survival as a function of time, and is described as follows in Eq. 3:

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

where β is the shape parameter or slope parameter ($0 < \beta < \infty$), and θ is the scale parameter or characteristic life ($0 < \theta < \infty$). Given a disk drive's total power-on hours per year, and the utilization calculated by Eq. 1, 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 $BaseValue$ in Eq. 1).

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 [21]. 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 [22] or PDC [23] 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 [1]. 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

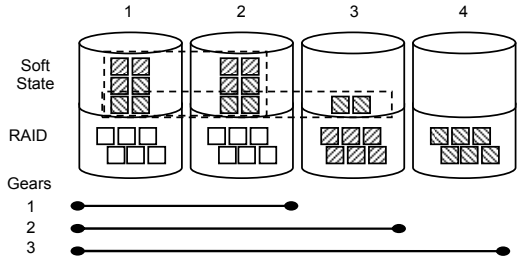

Fig. 2: Framework of PARAID: skewed striping of replicated blocks in soft state, creating 3 RAID gears over 4 disks [1]

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 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_{Base_Value} in Eq.1 shown in Section II). In this subsection, we develop a utilization model capturing behaviors of a RAID-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 L I/O phases. We denote T_l as the length or duration of the lth 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 $W_l = T_l/T$ where $T = \sum_{l=1}^{L} T_l$ is the total length of all the I/O phases. Suppose the utilization of a disk in the lth phase is ρ_l, 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

$$\rho = \sum_{l=1}^{L} (W_l \times \rho_l) = \sum_{l=1}^{L} (\frac{T_l}{T} \times \rho_l) \tag{5}$$

Since a PARAID system requires at least two disks to achieve the minimum I/O parallelism, the PARAID system consists of N disks has $(N-1)$ gears to shift. Assume that at the $G_{N-1}th$ gear, in which case all N disks of the system are kept spinning in order to offer the maximum parallelism, each single disk stores M blocks. When disk N is spun down, all its M blocks will be separated into $N-1$ sets in a way that each of the rest $N-1$ disks will handle making replicas for M blocks in disk N. Thus, we have:

$$F_{out\_G_{(N-1)(N-2)}} = M \tag{6}$$

and

$$if \mod (\frac{F_{out\_G_{(N-1)(N-2)}}}{N-1} = 0)$$

$$F_{in\_G_{(N-1)(N-2)}} = \frac{F_{out\_G_{(N-1)(N-2)}}}{N-1}$$

$$else \begin{cases} F_{in_{\_G(N-1)(N-2)}} = \left\lfloor \frac{F_{out\_G_{(N-1)(N-2)}}}{N-1} \right\rfloor + 1 \\ \quad for\ disk\ 1 \sim disk\ D \\ F_{in\_G_{(N-1)(N-2)}} = \left\lfloor \frac{F_{out\_G_{(N-1)(N-2)}}}{N-1} \right\rfloor \\ \quad for\ rest\ of\ (N-D) disks \end{cases} \tag{7}$$

where $D = mod(\frac{F_{out\_G_{(N-2)(N-3)}}}{N-2})$, $F_{out\_G_{(N-1)(N-2)}}$ represents replicas of the blocks moved out from the disk N when PARAID shifts down the gear from G_{N-1} to G_{N-2} due to the decreasing workload. $F_{in\_G_{(N-1)(N-2)}}$ represents the set of replicated blocks that moved into each of the $N-1$ disks. If M can be exactly divided by

$N-1$, each disk will handle $M/(N-1)$ blocks. Otherwise, the first remainder of $M/(N-1)$ disks will handle one extra block, while each of the rest disks will handle quotient of $M/(N-1)$ blocks.

Similarly, when PARAID shifts down from gear G_{N-2} to G_{N-3}, we have:

$$F_{out\_G_{(N-2)(N-3)}} = M + F_{in\_G_{(N-1)(N-2)}} \tag{8}$$

and

$$if \mod (\frac{F_{out\_G_{(N-2)(N-3)}}}{N-2} = 0)$$

$$F_{in\_G_{(N-2)(N-3)}} = \frac{F_{out\_G_{(N-2)(N-3)}}}{N-2}$$

$$else \begin{cases} F_{in\_G(N-2)(N-3)} = \left\lfloor \dfrac{F_{out\_G_{(N-2)(N-3)}}}{N-2} \right\rfloor + 1 \\ \qquad\qquad = \left\lfloor \dfrac{M + F_{out\_G_{(N-1)(N-2)}}}{N-2} \right\rfloor + 1 \\ \qquad for\ disk\ 1 \sim disk\ D \\ F_{in\_G_{(N-2)(N-3)}} = \left\lfloor \dfrac{F_{out\_G_{(N-2)(N-3)}}}{N-2} \right\rfloor \\ \qquad\qquad = \left\lfloor \dfrac{M + F_{out\_G_{(N-1)(N-2)}}}{N-2} \right\rfloor \\ \qquad for\ rest\ of\ (N-D) disks \end{cases} \tag{9}$$

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 $N-2$ disks move in more replicated blocks accordingly.

In general, when PARAID shifts down from gear G_j to G_i, where $j \in (3, ..., N-2)$, the number of blocks that the disk to be powered off must handle the following number of reads copy out is

$$\begin{aligned} F_{out\_G_{(j)(j-1)}} =& M + F_{in\_G_{(N-1)(N-2)}} + F_{in\_G_{(N-2)(N-3)}} + \\ &+ F_{in\_G_{(N-3)(N-4)}} \cdots + F_{in\_G_{(j+1)(j)}} \end{aligned} \tag{10}$$

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

$$if \mod (F_{out\_G_{(j)(j-1)}}/j) = 0$$

$$F_{in\_G_{(j)(j-1)}} = F_{out\_G_{(j)(j-1)}}/j$$

$$else \begin{cases} F_{in\_G_{(j)(j-1)}} = \left\lfloor F_{out\_G_{(j)(j-1)}}/j \right\rfloor + 1 \\ \quad for\ disk\ 1 \sim disk\ mod(F_{out\_G_{(j)(j-1)}}/(j-1)); \\ F_{in\_G_{(j)(j-1)}} = \left\lfloor F_{out\_G_{(j)(j-1)}}/j \right\rfloor \\ \quad for\ rest\ of\ disks. \end{cases} \tag{11}$$

where j represents the current gear number while $(j-1)$ indicated the gear number that the PARAID system is about to be shifted to, $\lfloor F_{out\_G_{(j)(j-1)}}/j \rfloor$ returns the integral part of $F_{out\_G_{(j)(j-1)}}$. 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 phase l as follows:

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

$$\rho_{power-off} = \frac{T_{I/O} + T_{read}}{T} \tag{12}$$

, while for the rest of disks, we have:

$$\rho_{power-on} = \frac{T_{I/O} + T_{write}}{T} \tag{13}$$

To improve the readability, Table II lists the notation used in our model.

TABLE II: List of Notations

Parameter	Description
R	Total Reliability
R_{Base_Value}	Reliability of Utilization
$R_{freq}(f)$	Reliability of Power Transition Frequency f
τ	Temperature Factor
α	Coefficient to R
β	Shape Parameter
θ	Scale Parameter
T_{active}	Active Time
T_{power_on}	Power-on Time
ρ	Disk Utilization
W_l	Relative Weight of l-th I/O phase
F_{out}	Copy Out File
F_{in}	Copy In File
N	Number of Disks
M	Number of Blocks
$T_{I/O}$	Service Time for I/O Requests
T_{read}	Service Time for Reading Duplicated Files
T_{write}	Service Time for Writing Duplicated Files

IV. Model Validation

A. The Validation Techniques

It is challenging to validate the accuracy of the MREED modeling framework, since we are unable to monitor PARAID running for a couple of decades. One way to address this problem is to maintain and analyze a large number of PARAID systems for a short period of time (e.g., 5 to 10 years). If one can track the systems over their entire service life, failure-rate data will be collected to validate reliability models. Even if we can test PARAID with 100 disks for five years, the sample size is small from a validation perspective.

To address this validation issue, we verify MREED using the Event Validity validation technique [24], which is a practical approach to verification and validation of reliability models. Events of occurrences of our MREED model are compared to those of the widely-used storage system simulator– DiskSim– to determine if our model and DiskSim agree with one another. In our validation process, we compared a file access trace in a real-world file system

Recall that MREED consists of two major components – a utilization model and a failure-rate model. The utilization model estimates disk utilization of the PARAID system based on I/O access rates. The failure-rate model relies on Weibull distribution analysis, parameters of which were derived from a hard drive disk manufacture's report (see [17]) to predict the possibility of disk failure from its utilization.

To validate MREED, we have to validate the utilization model and the failure-rate model. Since failure rates in this study are projections based on the failure-rate model derived from Seagate's empirical analysis (see [17]), we pay attention to the validation of the utilization model.

B. DiskSim Simulation

The DiskSim simulator, a powerful tool for the modeling and simulation of disk systems, is used widely for storage systems research [25]. Recent research projects using the DiskSim simulation environment include reducing disk I/O performance sensitivity and conserving energy in disk systems [26]. Although DiskSim is a

powerful simulation tool research, there is a lack of power models in DiskSim. The Sensitivity-Based Optimization of Disk Architecture introduced accurate power models into DiskSim, but this work was based on DiskSim 2.0 [27]. Another recent study on DiskSim and power models is the Dempsey project [28]. We are grateful to the author of the EEPF papar [29] who provided us with the source code of power models developed for a newer version (i.e., version 4.0) of DiskSim. This makes it possible for us to implement utilization and power transition models into DiskSim.

C. Simulation Framework

In order to complete our validation work via DiskSim, we integrate the following two major components in the system.

- DiskSim Simulator: It is in charge of simulating the operations of all disks and data blocks managements in the sytem.
- File to Block Translator: It is responsible for mapping files residing in the storage system into block-level data.

As shown in Fig. 3, files are mapped into blocks before being used as inputs to the DiskSim simulator. The file-to-block converter is critical, because data blocks are typically managed within a single node and a higher level mechanism is needed to manage data across different nodes in RAID systems. In the DiskSim simulator, we use the same

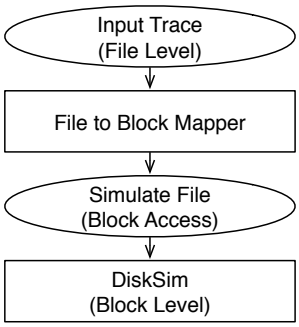

Fig. 3: File to Block Level Converter Outline

disk model (which is a Seagate ST3146855LW hard disk drive), the I/O throughput of which is significantly high than consumer level products. In order to avoid I/O transfer throughput bottlenecks, we modify a disk architecture in the DiskSim that each single disk has its own bus and controller (see in Fig. 4).

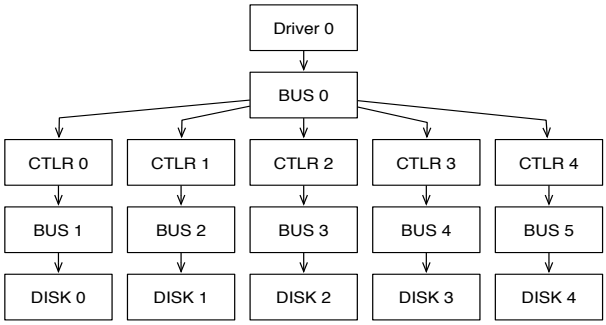

Fig. 4: Diagram of the Storage System Corresponding to the DiskSim Raid-0

D. UMass WebSearch Trace

The UMass WebSearch Trace [30] is used in the model validation process. This trace is obtained from the University of Massachusetts-Amherst (UMASS) website. The trace used in our experiments is WebSearch3.trace, which contains 4,261,709 read requests. The trace reply period is 298,715,395 milliseconds or 83 hours.

E. Validation Results

The Utilization-AFR model transfers the utilization of systems to the reliability. This model is employing the Weibull analysis by the same β and θ parameters (see Section II), so we only show the validation of utilization and power transition model in this subsection.

In order to make a clearer comparison between the MREED model and the trace-driven DiskSim simulations, we divided the comparison of utilization(see Fig. 5) and power transition (see Fig. 6). We observe that results obtained from the MREED model is similar to the simulation. Furthermore, the discrepancy between the model and the simulation is below 10%.

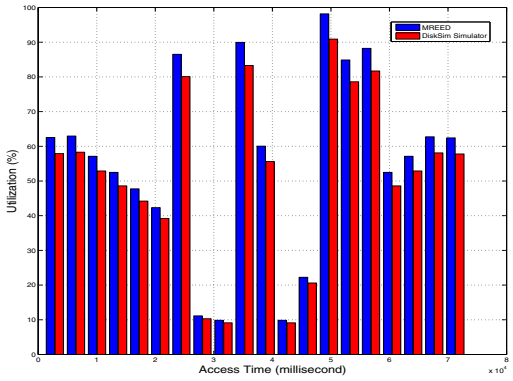

Fig. 5: Utilization Comparison Between MREED and DiskSim Simulator

After validating the Access Rate-Utilization sub-model, we further present the comparison results of Access Rate-Power Transition between the MREED model and the simulation results (as shown in Fig. 6). The figure shows that as time elapsed, the gear shifted accordingly as files access pattern changed. Fig. 6 illustrates that our model performs well in estimating gear-shift events.

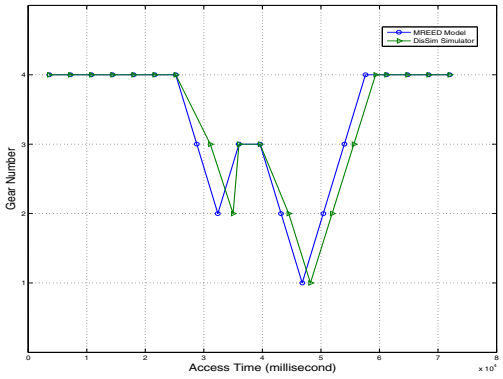

Fig. 6: Gear Shiftings Comparison Between MREED and DiskSim Simulator

V. Reliability Evaluation

A. Experimental Setup

We developed a simulator in which the PARAID-0 system (a.k.a Power-Aware RAID Level 0) is implemented. Table III shows the parameters of configurations for PARAID-0. We evaluate the reliability of a five-disk PARAID-0 system, in which the highest gear of the system is 4. In order to keep the RAID-0 configuration, there are two disks kept active at the lowest gear 1. The file access 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 III. Since Seagate's disks properties are introduced to our experimental

TABLE III: 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	5
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. 1 [17].

B. Disk Utilization

We first investigate the impacts of file access rate (λ in Poisson distribution) on utilization of PARAID-0. We set values of utilization to trigger gear-shifting are set to 60% for gear up while 30% for gear down. The PARAID-0 is assumed to be started at the top gear– all five disks are working . Fig. 7 plots the utilization comparison of PARAID-0 and RAID-0 within 24 hours. The average access rate is set to 20 per hour ($\lambda = 20$), which is relatively low. We observe from Fig. 7 that as time goes, the utilization of RAID-0 stays stable around 22%, while that of PARAID-0 increases twice then stays stable around 36%. Those two increasing points are caused by the gear-down shiftings hence the decreasing the number of active disks. Even though the utilization of PARAID-0 is 60% higher than that of RAID-0 at the end hour 24, the energy consumption of PARAID-0 is 40% lower than that of RAID-0 since there are only three active disks by then. Fig. 8 shows the utilization comparison of PARAID-0 and RAID-0 when the average access rate is set to 80 per hour ($\lambda = 80$), which is 3 times higher than that in Fig. 7. From the figure we notice that the utilization difference between PARAID-0 and RAID-0 is very vague. The major reason is that when the access rate is relatively high enough, the utilization of PARAID-0 keeps high (around 90% shown in Fig. 8) accordingly and, therefore a gear-shifting mechanism is not triggered. Hence at high access rate pattern, PARAID-0 behaves as similar as the regular RAID-0 system.

C. Annual Failure Rate

Fig. 9 illustrates the annual failure rates (AFR) of PARAID-0 and RAID-0 based on their utilization which is derived from Fig. 7. Results plotted in Fig. 9 show that AFR values of RAID-0 keeps increasing from 4.5% to 5.46% when hour lapses, while AFR of PARAID-0 increases by 4% at hour 2 and surges by another 8% at hour 3. We attribute this trend to the decreasing of the number

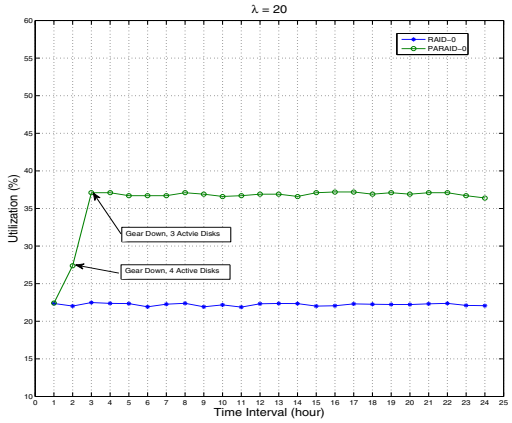

Fig. 7: Disks Utilization Comparison Between PARAID-0 And RAID-0 at A Low Access Rate(20 times per hour).

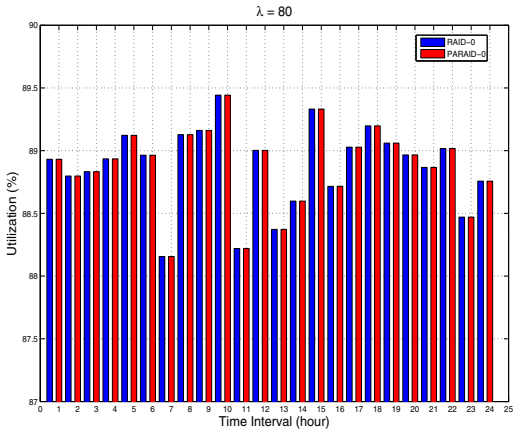

Fig. 8: Disks Utilization Comparison Between PARAID-0 And RAID-0 at A Low Access Rate(80 times per hour).

of active disks due to gear-down shifings. Since the utilization of PARAID-0 keeps the same as that of RAID-0 at high access rate, the AFR of the two systems are similar to each other accordingly. However, if the power transition issue is taken into account, AFR of PARAID-0 is different from that of RAID-0 even if their access rate are the same to each other. Fig. 10 reveals the AFR comparisons between RAID-0 and PARAID-0 starts from different gears within 24 hours. From the figure we observe that when the access rate increases shapely if PARAID-0 is not at the top gear, AFR of the system will suffer from the number of power transitions. Storage system at lower gear hvae relatively poor reliability. It is mainly because that more disks needs to be spun on to meet the needs of requests hence more number of power transitions will be counted.

VI. Related Work

A hard disk drive (HDD) is a complex dynamic system made up of various electrical, electronic, and mechanical components [31]. 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 [32][6] and multi-speed disks [5]), the operating system level (e.g., power-aware

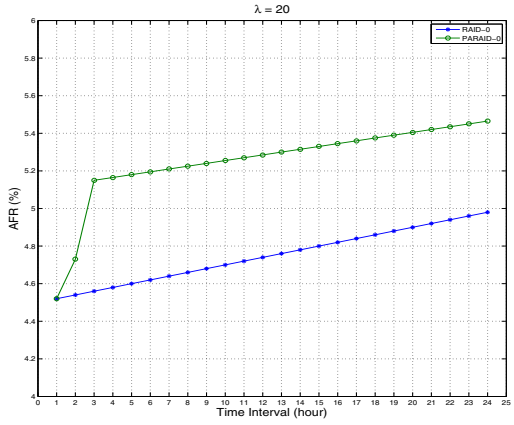

Fig. 9: AFR Comparison Between PARAID-0 And RAID-0 at A Low Access Rate(20 times per hour).

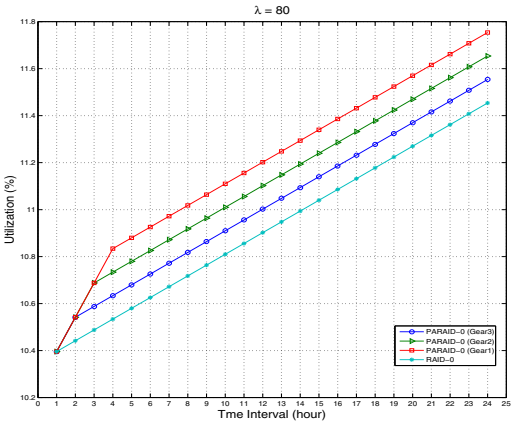

Fig. 10: AFR Comparison Between PARAID-0 And RAID-0 at A High Access Rate(80 times per hour).

caching/prefetching [8][15]), and the application level (e.g. software DMP [33] and cooperative I/O [34]). 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 [5][14][35]. 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 [23][22][36][1].

Recent studies show that both power management and workload skew schemes inherently impose adverse impacts on disk systems [2][3]. 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 [37], 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 [38], users have experienced a much lower MTBF from their field data [16]. 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 [16].

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 [39]; Elerath and Pecht proposed a flexible model for estimating reliability of RAID storage [40]; and Xin *et al.* developed a model to study disk infant mortality [41]. 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 *et. al* proposed a framework–e-PARAID–power optimization techniques in storage systems. In particular, the framework was applied to study the energy-efficiency of the PARAID system [42]. Our MREED approach differs from their e-PARAID in the sense that the goal of MREED is to evaluate reliability of energy-efficient parallel storage systems whereas e-PARAID aims to evaluate energy-efficiency of disk systems.

VII. Conclusion

This paper presents a reliability model called MREED to quantitatively study the reliability of energy-efficient parallel disk systems equipped with the PARAID technique. Note that PARAID is a newly developed energy-saving scheme for RAID systems. It aims to skew I/O load towards a few disks so that other disks can be transitioned to low power states to conserve energy. I/O load skewing techniques like PARAID inherently affect reliability of RAID disks, because disks keep working on low gears tend to have high failure rates, let alone the risk of failure caused by data duplicating during the *gear shifting*. Furthermore, once the number of failed disks exceeds the system's tolerance, data in the system are lost without any chance of being recovered. To address the model validation issue for MREED, we modified the DiskSim simulator, which is a widely-used storage system simulator, to validate our access-rate-utilizaiton sub-model of MREED by comparing the utilization of 5-disk PARAID system using a real-world disk I/O trace with the utilization that calculated from the MREED model using the same trace.

Future directions of this research can be performed in the following. First, we will extend the MREED model to investigate reliability of different levels (e.g., level 5) of PARAID in the future which introduces parity data technique to tolerate one disk failure. Second, we will investigate a fundamental trade-off between reliability and energy-efficiency in the context of energy-efficient RAID systems. A tradeoff curve will be used as a unified framework to justify whether or not it is wise to trade reliability for high energy efficiency. Last, we will evaluate and compare an array of energy-saving techniques with respect to specific application domains.

ACKNOWLEDGMENTS

This research was supported by the U.S. National Science Foundation under Grants CCF-0845257 (CAREER), CNS-0917137 (CSR), CNS-0757778 (CSR), CCF-0742187 (CPA), 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. Meikang Qiu's research is partially sponsored by University of Kentucky Start Up Fund and NSFC 61071061. Mohammed Alghamdi's research was supported by the King Abdulaziz City for Science and Technology (KACST) and AL-Baha University.

REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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

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

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

[14] 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.

[15] 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.

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

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

[18] 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.

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

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

[21] 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.

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

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

[24] R. G. Sargent, "Verification and validation of simulation models," in *Proceedings of the 37th conference on Winter simulation*, ser. WSC '05. Winter Simulation Conference, 2005, pp. 130–143. [Online]. Available: http://portal.acm.org/citation.cfm?id=1162708.1162736

[25] S. W. S. John S. Bucy, Jiri Schindler and G. R. Ganger, "The disksim simulation environment version 4.0 reference manual," 2008.

[26] 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.

[27] H. Shen, M. Kumar, S. Das, and Z. Wang, "Energy-efficient caching and prefetching with data consistency in mobile distributed systems," in *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, april 2004, p. 67.

[28] Q. Zhu, Z. Chen, L. Tan, Y. Zhou, K. Keeton, and J. Wilkes, "Hibernator: helping disk arrays sleep through the winter," *SIGOPS Oper. Syst. Rev.*, vol. 39, pp. 177–190, October 2005. [Online]. Available: http://doi.acm.org/10.1145/1095809.1095828

[29] A. C. Manzanares, "Energy efficient pre-fetching–models to implementation." Auburn University, April 2010.

[30] "Umass trace repository," http://traces.cs.umass.edu/index.php/Storage/Storage, December 2009.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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