REED: A Reliable Energy-Efficient RAID

Shu Yin, Xuewu Li, Kenli Li
School of Information Science & Engineering
Hunan University
National Supercomputing Center in Changsha
Changsha, China
Email:{shuyin, lixuewu2009, jt_lkl}@hnu.edu.cn

Jianzhong Huang
Dept. of Computer Science
Huazhong Univ. of Sci. and Tech.
Wuhan, China
Email: hjzh@hust.edu.cn

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

Xiaomin Zhu
Sci. & Tech. on Information System Eng. Lab
National University of Defense Technology
Changsha, China
Email: xmzhu@nudt.edu.cn

Wei Cao
Hubei Meteorological Information.
& Technology Support Center
Wuhan, China
Email: 44647874@qq.com

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

***Abstract*—Recent studies indicate that the energy cost and carbon footprint of data centers have become exorbitant. It is a demanding and challenging task to reduce energy consumption in large-scale storage systems in modern data centers. Most energy conservation techniques inevitably have adverse impacts on parallel disk systems. To address the reliability issues of energy-efficient parallel disks, we propose a reliable energy-efficient RAID system called REED, which aims at improving both energy efficiency and reliability of RAID systems by seamlessly integrating HDDs and SSDs. At the heart of REED is a high-performance cache mechanism powered by SSDs, which are serving popular data. Under light workload conditions, REED spins down HDDs into the low-power mode, thereby offering energy conservation. Importantly, during an I/O access turbulence (i.e., I/O load is dynamically and frequently changing), REED is conducive to reducing the number of disk power-state transitions by keeping HDDs in the low-power mode while serving requests with SSDs. We build a model to quantitatively show that REED is capable of improving the reliability of energy-efficient RAIDs. We implement the REED prototype in a real-world RAID-0 system. Our experimental results demonstrate that REED improves the energy-efficiency of conventional RAID-0 by up to 73% while maintaining good reliability.**

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

I. Introduction

Traditional energy-saving techniques in disk array (RAID) systems do not take into account adverse impacts of energy conservation mechanisms on system reliability. Existing solutions tend to frequently spin up and down hard disk drives (HDDs) or replace HDDs with solid state disks (SSDs). These solutions may cause severe mechanical malfunction of HDDs and lead to permanent data loss. Moreover, heavy write I/O loads shorten the lifetime of SSDs due to the limited numbers of erasure cycles of flash memory. SSDs are more expensive than HDDs; therefore, it is non-cost-effective to build a large-scale storage system using SSDs. In this paper, we propose a hybrid RAID system called *REED* aiming to improve reliability and energy efficiency of storage systems by integrating HDDs and SSDs. At the heart of *REED* is the hybrid disk technique that exploits SSDs' low power consumption as well as HDDs' large capacity and long lifetime to improve energy efficiency and reliability of RAID systems under dynamically changing I/O workloads.

A. Motivations

The following four factors motivate us to develop the REED system that is reliable and energy efficient.

- the high energy consumption of storage systems in data centers;
- the challenge of improving energy efficiency of RAID systems;
- the benefit of integrating solid state disks and hard drives; and
- the adverse impacts of existing energy-saving techniques on RAID system reliability.

Motivation 1. Increasing evidence indicates that as much as 27% of the energy in a modern data center is consumed by storage devices [4]. Even worse, such a fraction tends to go up as data storage capacity is dramatically rising by 60% annually [32]. Energy spent to operate disks leads to high heat dissipation, which poses a serious obstacle to the development of energy-efficient cooling systems. Improving energy efficiency of storage systems can substantially reduce operating costs of large-scale data centers; reducing energy consumption of storage systems offers potential economical and environmental benefits [7].

Motivation 2. Data centers that consume significant amount of energy deploy reliable RAID systems to meet the growing requirements of capacity and I/O performance. This trend makes it appealing and practical to improve the energy efficiency of RAID devices through the development of promising energy-conservation approaches. Due to the data-stripping mechanism, each disk in a RAID system stores only a portion of files; thus, all disks have to be power-on to service I/O

0190-3918/15 $31.00 © 2015 IEEE
DOI 10.1109/ICPP.2015.74

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

requests even under light workload conditions. Such characteristics of RAIDs make conventional energy-saving techniques, such as dynamic power management and workload skew techniques, inadequate for RAID systems.

Motivation 3. A few existing approaches build energy-efficient RAIDs for data archival systems [9], whereas others make good tradeoffs between performance and energy efficiency [30][50]. Controller-level cache management schemes leverage skew striping patterns [39] to reduce the number of power-on disks to meet system loads in RAID systems [21]. NAND-flash-based solid state disks (SSDs) are becoming popular storage devices thanks to their significant advantage of good shock resistance, fast data accessing, and lower energy consumption. However, SSD's drawback in low cost-efficiency and short lifetime prevent them from being widely applied to large data centers. Recent studies show that the erasure limit of SSDs is typically rated at 5,000 to 10,000 cycles per block, meaning that an SSD could be worn out within a few months under heavy write loads [3]. Hybrid techniques like HRAID6ML [48], HPDA [25], and CacheRAID [8] embrace the performance benefit of flash-based storage and the low cost of hard drives.

Motivation 4. 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 [6][42]. 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[45][47].

We have designed and measured the Reliable Energy-Efficient RAID (*REED*), which combines HDDs and SSDs to reduce negative effects on system reliability while reducing power consumption. *REED* aims to squeeze the hard drives' benefit in high cost-efficiency, high capacity, and long lifetime as well as making use of SSDs as the replica storage to take place sets of hard drives during light I/O workload. Compared to a conventional 5-disk RAID, a 6-hybrid-disk *REED* (5 HDDs and 1 SSD) can reduce power consumption by up to 73% while maintaining reasonable reliability and comparable power consumption under a fluctuating system workload. Moreover, *REED* reuses different RAID levels so that it can be applied to general RAID technologies.

The benefits of *REED* are three-fold.

- As *REED* uses hard drives as the primary storage method, it provides an un-comparable capacity budget for data centers;
- *REED* offers a better balance among energy consumption, reliability, and hardware budget under a fluctuating system workload;
- *REED* retains a better system reliability compared to flash-based RAID systems and provides a practical solution during the transition period from full-HDD storage to full-memory storage.

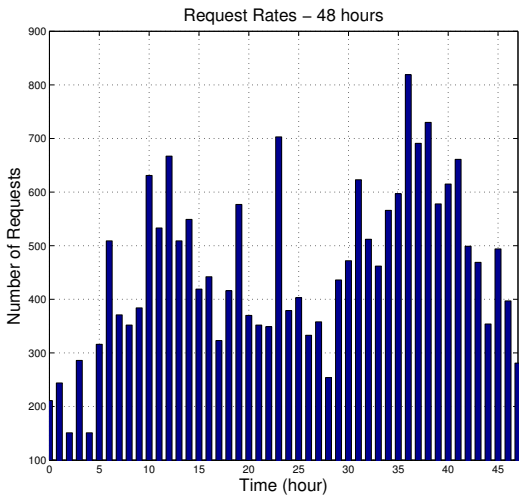

Fig. 1: Florida State University Web System Request Rate from June 13th through May 14th, 2005[1].

The remainder of this paper is organized as follows. Section II presents the observation of storage systems in term of workload, energy efficiency, and reliability. In Section III, we describe the architecture and design of *REED*. The experimental methodology and reliability analysis are presented in Section IV. Section V presents experimental results as well as performance analysis. In Section VI, we discuss the related work. Finally, Section VII concludes the paper with future research directions.

II. Observations

Increasing Needs for Storage: With the rapid growth of data generated in data-intensive applications, the requirement of storage space has been dramatically growing. Large-scale parallel I/O systems are widely used in high-performance computer systems. For example, the Facebook Photos application currently stores over 260 billion images, which account for more than 20 petabytes of data. Furthermore, up to 60 terabytes of image files are uploaded weekly [5].

Cyclic Load Fluctuation: An increasing number of online systems exhibit intriguing workload patterns that are cyclic fluctuations. Fig. 1 shows the collected request rates of an online server running for 48 hours at the Florida State University. The access patterns plotted in Fig. 1 implies that I/O load is periodically changing back and forth between light and heavy ones during the two-day observation interval. Such dynamically changing workload makes it possible to conserve energy by reducing the number of active disks under light load as long as performance needs are satisfied.

Solid State Storage: Solid State Devices or SSDs have emerged in the last few years as a viable replacement for hard drives in a wide variety of application settings. Commodity SSDs offer promissing random read and write performance, potentially eliminating I/O bottlenecks in high-performance data centers while driving down energy consumption [3]. OCZ

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

reports that SSDs speed up the performance of conventional hard drives (HDDs) by a factor of more than 100 [28]. More importantly, the power efficiency of SSDs is significantly higher than that of HDDs; for example, the power consumption of an SSD and an HDD at a peak load are 2W and 6W, respectively. Unfortunately, SSDs have severe disadvantages that prevent them from being widely applied in modern data centers. These drawbacks include low endurance limits and high cost. Very recently, a study demonstrates that the erasure limit of MLC devices is typically ranging anywhere from 5,000 to 10,000 cycles per block [41]. Although the SSD's cost per GB sharply drops down to 80¢, the cost-efficiency of SSDs is nowhere near that of HDDs, which cost as low as 5¢per GB [33][17]

Energy Efficiency vs. Reliability: Evidence on the reliability of RAID systems began to accumulate indicating that existing energy conservation techniques are inadequate for RAID systems due to the following three reasons [39]. First, no opportunity is offered to spin down any disk in a conventional RAID system due to the I/O load balancing across all the disks for maximized disk parallelism and performance. Second, hard disks were not designed for frequent power cycles, which significantly reduce HDD life expectancy. Third, energy-efficient caching and dynamic power management are inapplicable for RAIDs deployed in server systems, because the servers are too busy to have any long idle time period.

A reliability analysis (see, for example, [47]) on energy-efficient RAID systems suggests that an energy-saving mechanism greatly affects the healthiness of parallel disks, and the reason is two-fold: (1) the power-state-transition frequency caused by spinning up and down disks may lead to mechanical malfunctions in disks [46]; (2) flash-based disks are likely to exceed the erasure limits and have a high risk of breaking down with high frequent updates.

In order to take the full advantages of both SSDs and HDDs, we propose a solution called *REED* that fully utilizes HDDs under heavy workload conditions and hybrid disks (i.e., SSDs and HDDs are integrated) in light workload scenarios. When the workload of a storage system is heavy, no energy-saving mechanism can be applied to conserve energy by placing disks into the low-power mode. In this case, *REED* leverages HDDs as primary storage devices to achieve high I/O capacity and data reliability; *REED* also enlarges the lifetime of SSDs by keeping SSDs in the power-off state to avoid frequent updates on SSDs. Under light workload, SSDs are powered on to perform as a cache and a buffer for HDDs that are transitioning into the low-power mode to conserve energy. In doing so, SSDs boost the reliability of HDDs by eliminating the number of unnecessary power-state transitions in HDDs.

III. A Reliable Energy-Efficient RAID

The overarching goal of *REED* is to offer high system reliability and energy efficiency by allocating data stripes to disks. In this section, we first propose an approach to placing data strips in a way to conserve energy. Next, we outline the architecture of *REED*. Then, we articulate a way of modeling data popularity. Finally, we describe how to maintain high reliability for energy-efficient disk arrays.

A. Conserving Energy through Data Strip Allocation

REED leverages solid state disks, which are highly energy efficient, to cache and buffer data stripes in a skewed fashion. In doing so, *REED* strives to shut down a subset disks while maintaining the RAID structure. When disks are powered off, *REED* services I/O requests via a combination of blocks residing on power-on disks as well as on data replicas residing on the SSDs. *REED* identifies blocks that have been frequently accessed (a.k.a., hot data); replicas of hot data blocks are placed and managed into SSDs. Hence, in the case where a few disks are spun down, requests accessing hot data can naturally and quickly be served by SSDs. Evidence begins to accumulate showing that when workload is light, requests are likely to access hot data. Although a request might access data that are less popular than hot data, disks may be temporarily spun up to provide an I/O service under the light load. Recently accessed cold data will be cached into SSDs after disks are spun down again, thereby reducing unnecessary disk power-state transitions when the cold data is accessed in the not-long-distant future.

By replicating an appropriate amount of data strips to SSDs based on system workload, *REED* aims to aggressively conserve energy consumption by reducing the number of disks that are in the active mode. *REED* is maintaining high energy efficiency thanks to SSDs' low power consumption, which is one third of that of hard drives or HDDs. Even though SSDs are more expensive than HDDs in terms of price per GB, SSD-enabled *REED* becomes cost-effective in the long run by the virtue of SSDs' low energy usage and high I/O performance.

Fig. 2 shows an example of a REED-0 system (a.k.a Reliable Energy-Efficient RAID-0), which is comprised of five HDDs and a single SSD. When system workload in the REED-0 system is relatively high, all the disks are turned on. In this case, *REED* performs as a conventional RAID system offering maximized I/O parallelisms and performance (see also stage (a) in Fig. 2). During high workload, performance and quality of service receive a higher priority than energy efficiency. As I/O load decreases, *REED* spins down a disk (e.g., HDD5) as long as blocks stored in HDD5 are replicated into an SSD (see also stage (b) in Fig. 2). To address the drawback of limited capacity in SSDs, *REED* only keeps replicas of hot blocks in SSDs while storing cold blocks (i.e., unpopular data) in HDDs. If the load continues decreasing, *REED* will further shut down more disks (see also stages (c) and (d) in Fig. 2) until it reaches a state where only one disk and one SSD are servicing the lightest I/O load (see also state (e) in Fig. 2). Recognizing that an SSD delivers high I/O performance 100 times faster than hard drives [28], we argue that I/O performance of *REED* in the one-HDD-one-SSD state is superior to a RAID-0 system that contains three-HDDs.

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

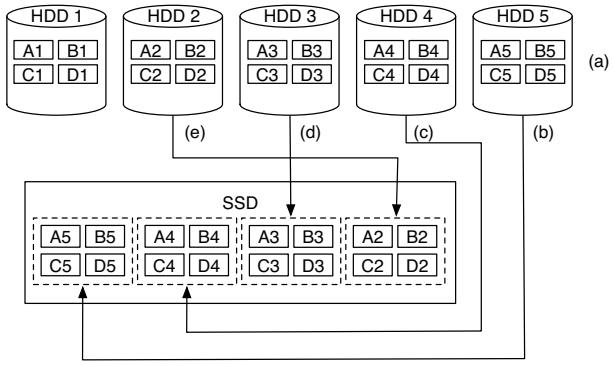

Fig. 2: The REED-0 system that is comprised of five HDDs and one SSD. Data strips are replicated copied from the HDDs to the SSD according to system I/O workload conditions. REED manages different storage setups over five disks under five stages (a)~(e). The details are delineated in III-A

B. The REED Architecture

Fig. 3 outlies the architecture of our proposed *REED*. The architecture is composed of both SSDs and HDDs, among which HDDs service as a primary storage whereas SSDs perform as a caching mechanism for data replicas. Recall that the goal of *REED* is to reduce system energy consumption under light workload. The number of SSDs and the capacity of SSDs are determined by the performance and space requirements of applications running in *REED*. Obviously, increasing the number of SSDs in *REED* leads to a large storage system with an improved I/O performance; on the other hand, an excessive number of SSDs inevitably makes *REED* less cost effective. *REED* consists of four key functional modules including a workload monitor, a director, an SSD manager, and a disk manager. The workload manager of *REED* is responsible for monitoring I/O access patterns and dynamically measuring system workload. Other than the monitor, the director is in charge of determining the number of HDDs that are in the active mode; the director judiciously determines which SSDs should be involved when a subset of disks are powered off during light system load. Importantly, the director is managing I/O distributions among active HDDs and proactively loading the replicas of data strips from HDDs to SSDs. When the system runs under heavy workload, *REED* relies on the original RAID configuration to preserve high I/O performance. In case of light I/O load, *REED* identifies the number of power-on HDDs and SSDs while shutting down a subset of HDDs to boost energy efficiency.

The workload monitor in *REED* continuously monitors system I/O workload, allowing the director to determine whether the hybrid mechanism should be activated. Once the dynamically changing workload reaches a preset utilization threshold, the director of *REED* is triggered to decide the number of active disks and to manage replicas cached in SSDs. Utilization thresholds (i.e., high threshold and low threshold)

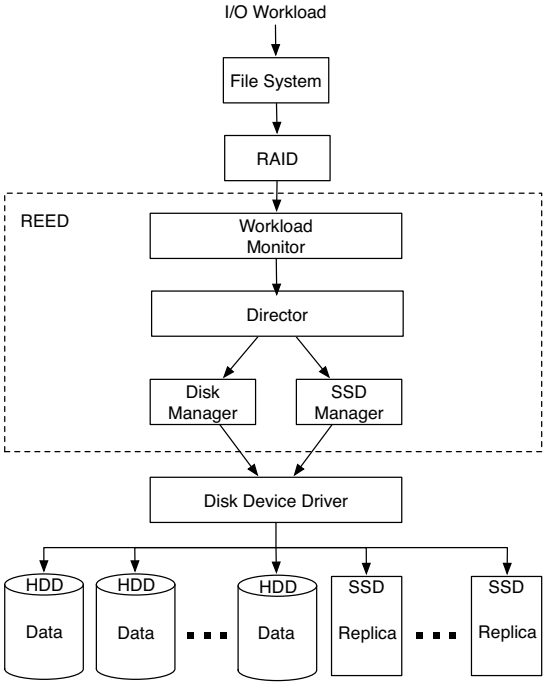

Fig. 3: The Architecture of REED.

are configurable in *REED*. When the system load reaches the high threshold, disks in the low-power mode are spun up to serve heavy workload; if the load is below a low threshold, a subset of disks are transitioned into the low-power mode to conserve energy. Section III-C discusses the details of the thresholds and the workload turbulence issue addressed by our threshold policy.

C. Data Popularity and Thresholds

The data popularity can be characterized using a Zipf-like distribution with high coefficient. The Zipf's law suggests that the popularity of a file measured in terms of I/O access frequency is the inverse of the file's rank. The relationship between a file's access frequency and its rank can be expressed as $P(r) = 1/(ln(N) * r)$ where N is the number of files, r is the rank of a file, $P(r)$ represents the file popularity [18]. For example, given 5000 files (i.e., $N = 5000$), the popularity or access frequency of the most popular file is $1/(ln(5000)*1) = 11.7\%$.

When it comes to power management, the director compares system workload with a threshold, based on which the director judiciously decides which disk should be power-on or power-off. *REED* applies the fixed-threshold technique, the principle of which is straightforward - a disk is spun down after the system's utilization is below a pre-specified threshold. The monitor of *REED* measures the system load in terms of I/O bandwidth within a time window. The fixed-threshold technique does not address the I/O burstiness incurred by making and placing replicas of popular data. Before turning off a disk, *REED* ensures that popular data stored on the

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

disk are duplicated to an SSD. In case the replicated data are not available on the SSD, *REED* initiates the data duplication process, which pushes up the system utilization. The increased utilization attributed to making replicas and turning off disks may become higher than the threshold, thereby triggering *REED* to turn on disks again. Such turbulent workload conditions inevitably and frequently power disks on and off. To tackle this problem, *REED* seamlessly integrates the multiple-threshold policy into the fixed-threshold scheme.

To address the aforementioned turbulent workloads, we configure in *REED* two thresholds (i.e., high and low thresholds) making three utilization zones called high-, medium-, and low-utilization zones (i.e., H-zone, M-zone, and L-zone). The power management goal is to keep the system utilization in the M-zone. Specifically, *REED* starts turning off disks when the utilization is in L-zone and turning on disks if the utilization is in H-zone. For example, we set the high and low thresholds to 80% and 30% given access patterns. When the utilization of a 4-disk system is in L-zone (i.e., load is below 30%), *REED* spins down a disk. As a result, the new 3-disk system's utilization goes up to 40%; the data duplication process contributes 20% load. The total utilization of 60% keeps the system load in M-Zone, avoiding the workload turbulence problem.

D. Improving Reliability

To improve reliability of energy-efficient RAID, we ensure that *REED* is capable of tolerating disk failures. In energy-efficient RAIDs, life expectancy of server-class disks is likely to be reduced due to power-state transitions. This problem is addressed in *REED* by facilitating RAIDs with data redundancy (i.e., placing replicas of popular data into SSDs).

REED can be envisioned as an intermediate layer sitting above a software RAID, which is supported by the operating system and physical storage devices. Thus, *REED* inherits data redundancy, striping granularity, and disk layouts. For example, the REED-0 system, which is a reliable energy-efficient RAID-0, is able to reconstruct a failed disk from parity blocks.

One challenge we are facing is how to maintain a high reliability of SSDs in *REED*. SSDs have limited erasure cycles; frequent data updates under light I/O load may shorten the lifetime of *REED* when it is in the SSD-HDD hybrid state. Furthermore, an excessive number of power-state transitions sabotages the reliability of hard disks. For example, suppose the transition frequency is 300 per month, the annual failure (AFR) rate is increased by 0.13% [45].

In order to maintain high reliability of SSDs, *REED* limits the number of erasure cycles of SSDs by disabling the SSDs under heavy I/O load. Enormous number of writes issued under high workloads tends to substantially shorten the lifespan of SSDs. Reducing the number of erasure cycles is achieved by keeping SSDs active when workload conditions are light. Apart from improving the reliability of SSDs, *REED* is capable of maintaining high reliability of HDDs. Reliable HDDs become possible in *REED*, because the number of power-state transitions of HDDs are reduced by SSDs under light and fluctuating workloads.

IV. Reliability Analysis

To provide a reliability analysis for *REED*, we implement *REED* in a disk array simulator by extending DiskSim [19], which is a widely used simulator for storage systems. Because there is a lack of a power model and SSD support in DiskSim, we incorporate simulated SSDs developed by Microsoft [26] and a power model built by Manzanares [24] into DiskSim. The updated disk simulator is driven by a web server trace that resembles read-intensive I/O traffic.

A. A Disk Model

We choose a disk model to simulate the real-world Seagate ST3146855LW [35] disks, because in our reliability model two major coefficient values of the Seagate disk can be obtained from a white paper [4]. The detailed parameters are summarized in Table I.

TABLE I: Disk Model Parameters

Disk Type	SEAGATE ST3146855LW
Capacity	146 GB
Cache Size	16 MB
Buffer to Host Transfer Rate	4 Gb/s (MAX)
Spindle Speed	15k RPM
Average Latency	2.0 ms
Average Seek Time	3.5(read)/4.0(write) ms
Average Transfer Rate	200 MB/s (MAX)
Active Power	14.2 Watts

B. Simulating SSDs

An SSD extension developed by Microsoft is patched to the DiskSim simulator to investigate SSDs. Specifically, we choose the OCZ Vector as the SSD module [27] because of two reasons. First, OCZ is one of the largest SSD vendors with a good reputation on SSD products [38]. Second, OCZ is one of few SSD companies that released details on their SSD product specifications. The detailed parameters are shown in Table II.

TABLE II: SSD Model Parameters

Disk Type	OCZ SSD VTR1-25SAT3-128G
Capacity	128 GB
NAND Components	25nm MLC
Ave. Latency	2.0 ms
Ave. Transfer Rate	550(read)/400(write) MB/s
Active Power	2.25 Watts

C. Extending DiskSim

We extend the DiskSim simulator by incorporating disk power state transitions. Because the disk models in the original DiskSim do not support power management, we add seven key parameters into the disk models to facilitate energy conservation techniques. In the extended DiskSim, we consider

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

two major power states – active and standby. Specifically, parameter hdd_power_on indicates power consumption when a disk is serving requests; parameter hdd_power_off is the power consumption when the disk is placed into the standby state. Parameters hdd_time_sd and hdd_time_su keep track of the amount of time spent in spinning down and up the disks, respectively. Parameters hdd_power_sd and hdd_power_su are the energy consumed to spin down and spin up the disk, respectively.

We modify DiskSim in a way to determine circumstances under which the disks should be spun down to conserve energy or be spun up to service requests when I/O load is surging up. We create an utilization tracker in DiskSim to calculate the system utilization. When the utilization exceeds the upper threshold (i.e., in H-zone), the tracker spins up standby disks in response to heavy workload; if the utilization drops below the lower threshold (i.e., in L-zone), the tracker spins down disks to offer energy savings. Such a power management feature is implemented by adding a UTI_Chk event to the DiskSim's event queue. In addition to the utilization tracker, we implement a data-replica manager in DiskSim to simulate I/O load introduced by duplicating popular data into an SSD (see also Section III-C). If the replica manager confirms that data are not available on the SSD, then the manager will simulate *REED*'s data duplication process by creating a list of I/O requests according to the amount of data to be duplicated.

D. Traces

The energy efficiency of *REED* is highly affected by workload characteristics. For example, under a continuously high load, there is the lack of opportunity for *REED* to save energy by powering off any disk. Nevertheless, no energy saving schemes can reduce energy consumption under heavy workloads. When the workload becomes light, *REED* is likely to turn off most of the hard drives to achieve optimized energy efficiency.

To resemble real-world scenarios, we make use of a web server trace to drive the extended DiskSim. We choose the web server trace to conduct our experiments; the reason is three-fold. First, the web trace exhibits the load-fluctuation feature, which demonstrates the strengths of *REED*. For example, the workload of multimedia web servers stays low during daytimes, surges up in evenings, and drops down after mid-nights. Second, most of the web service providers support the always-write-new policy rather than repeatly updating stored data. This policy avoids multiple writes to a data block in SSDs, thereby improving the lifetime of SSDs. Last, this study is focused on read-intensive workloads, which are represented by the web traces. Most datasets managed by web service providers (e.g., YouTube, Facebook, and Twitter) are write-once-read-multiple-times in nature. The web trace resembles the I/O workloads of popular real-world web services.

The web traces were collected by the Department of Computer Science at Florida State University. The trace captures activities from 4 AM, June 12, 2005 to 1AM, June 17, 2005. The trace collected during this time interval includes 47,562 requests accessing 14 Gbytes of data [1]. The relatively light workload, which captures the essence of read-intensive cyclic loads, sheds light on *REED*'s system performance in terms of energy efficiency and reliability.

E. Reliability Analysis

We apply *MREED* [47] to quantify the system reliability of our proposed *REED*. It is noteworthy that *MREED* is a reliability modeling framework for parallel disk systems coupled with 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. A second important module of *MREED* is to calculate the annual failure rate of each disk as a function of the disk's utilization and power-state transition frequency. Given the annual failure rate of each disk in a parallel disk system, *MREED* derives the reliability of the system.

MREED is composed of a Weibull-based disk reliability model, a system-level reliability model, and three reliability-affecting factors (i.e., temperature, power state transition 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 transition frequency and utilization. The Weibull-based disk reliability model derives individual disk's possibility of failure from utilization and power-on hours per year. Each disk's reliability is an input to the system-level reliability model evaluating the annual failure rate of parallel disk systems. Fig. 4 outlines the *MREED* reliability modeling framework.

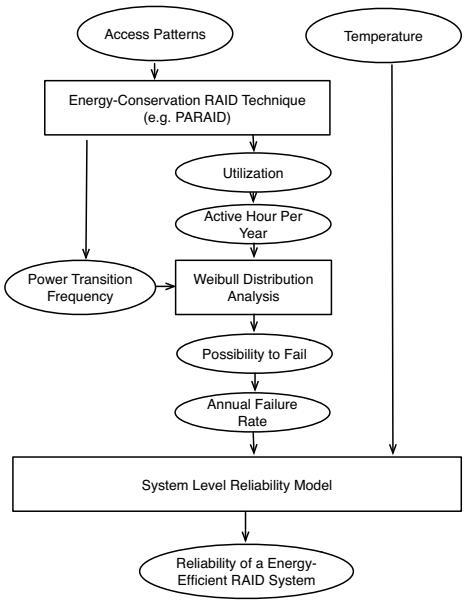

Fig. 4: Overview of the MREED reliability modeling framework.

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

The system reliability can be expressed as

$$R = R_{util} * \tau + \alpha * R_{freq} \tag{1}$$

where R_{util} is the baseline failure rate derived from disk utilization, which is examined by the Weibull distribution (see Eq. 2 [2]), τ is the temperature factor, α is a coefficient to reliability R, and R_{freq} is a power-state transition frequency adder to the baseline failure rate. We apply Eq. 3 [45] to calculate adder R_{freq}.

$$\begin{aligned} R(t) &= \int_t^{\infty} \frac{\beta(x)^{(\beta-1)}}{\theta^{\beta}} \exp[-(\frac{x}{\theta})^{\beta}]dx \\ &= \exp[-(\frac{t-\beta}{\theta})^{\beta}] \end{aligned} \tag{2}$$

$$R_{freq}(f) = 1.51e^{-6}f^2 - 1.09e^{-5}f + 1.39e^{-2}, f \in [0, 100] \tag{3}$$

where f is a power-state transition frequency and $R_{freq}(f)$ represents an adder to the base annual failure rate.

Thanks to the absence of mechanical moving units, SSDs are more reliable than hard drives in terms of traditional reliability measurements such as mean time to failure (i.e., MTTF) or mean time to data loss (i.e., MTTDL). The lifetime of SSDs is largely determined by bit error rate or BER. Growing evidence shows that bit errors occur due to writes, reads, and bit-rot over time [14]. In practice, hardware ECC is applied to reduce SSD error rates. There are two types of BER, namely, RBER (raw bit error rate) and UBER (uncorrectable bit error rate). RBER is a pre-ECC measure, whereas UBER is a post-ECC measure.

From the perspective of data loss, a storage system is reliable unless any data stored on the system can not be recovered. A REED system does not fail even if all SSD disks become faulty, because SSDs are holding replicas of popular accessed data stored on hard drives. In other words, replicated data kept on faulty SSDs can be straightforwardly recovered from corresponding hard disks. For the sake of simplicity without loss of generality, in our reliability analysis we pay attention to hard disk failures. In our future study, we will address the reliability issue of SSDs in *REED*.

V. Experimental Results

A. Experimental Setup

In our experiments, we focus on the comparisons between *REED* and RAID-0, which achieves high I/O performance by the virtue of the striping technique. The drawback of RAID-0 lies in the lack of a data recovery mechanism.

Apart from RAID-0, we compare *REED* against PARAID (i.e., Power-Aware RAID), which is a well-known energy-efficient RAID system. During the process of evaluation, we focus on energy savings as well as system reliability. Table III summarizes the configuration parameters of the disk systems used throughout our experiments. A baseline disk system is comprised of five disks, whereas the *REED* system incorporates an extra SSD as a cache disk performing under light workloads. An initial mode configured in both PARAID and *REED* is the full-disk mode, in which all hard disks are power-on and the SSD cache disk in *REED* is turned off.

TABLE III: Disk System Parameters

Disk Type	SEAGATE ST3146855FC
Capacity	146 GB
Cach Size	SATA 16MB
Buffer to Host Transfer Rate	4Gb/s (MAX)
Total Number of Disks	5(RAID) 5(PARAID) 5+1(REED)
Time Period	48 hours
Interval Time (Time Phase)	1 hour
Power On Hour Per Year	8760

We deploy Seagate's disks in *REED*, RAID-0, and PARAID. To resemble real-world Seagate disks, we set β and θ to 0.55 and 8410332 in the Weibull analysis model [47]; the temperature factor τ is set to 0.54 [10].

B. Power Consumption

Recognizing that the trace used in our experiments represents light I/O loads, we extrapolate the trace in two ways. First, we increase the arrival rate in terms of request per second. Second, we enlarge the data size of each I/O request. More specifically, in the first modification, we keep all the request sizes unchanged while substantially speeding up the arrival rate of the trace by a factor of 256, which are denoted as $256x$. Such a manipulation results in a high arrival rate of 58 requests per second, respectively. As for the second modification, the request size is increased to 32MB, which correspond to a high throughput of 27 MB per second, respectively.

Figs. 5 plots the power consumption of RAID-0, PARAID-0, and REED-0 under a web I/O density. Due to the lower I/O density (see Fig. 5), both PARAID-0 and REED-0 are offered ample opportunities to shut down a number of hard disks while maintaining acceptable I/O response times. Hence, PARAID-0 and REED-0 exhibit higher energy efficiency than the traditional RAID-0 system. Thanks to SSDs' low power consumption and high I/O bandwidth, REED-0 fully utilizes one hard drive and one SSD for a long time period under light load. As a result, REED-0 reduces the energy consumption of PARAID-0 by more than 40%.

Now we evaluate the performance of *REED* in the context of multimedia systems. We manipulate the trace modification by setting the file size to 32MB, because the file size resemble real-world video streaming services like YouTube and Hulu. Figs. 6 reveals the power consumption of RAID-0, PARAID-0 and REED-0 under a two file-size settings. When I/O throughput are dramatically changing, PARAID-0 spares no effort to switch the number of power-on disks to achieve high energy efficiency and acceptable I/O bandwidth. In contrast, REED-0 takes the full advantage of SSDs to reduce energy consumption. Fig. 6 shows that without frequent power-state transitions, REED-0 conserves energy by 70%.

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

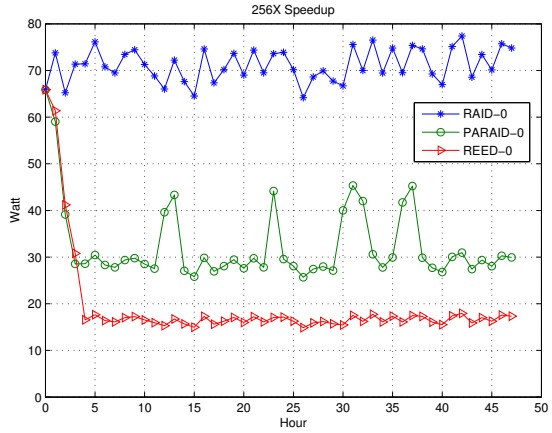

Fig. 5: Power consumption of RAID-0, PARAID-0, and REED-0. (Speedup of the Web trace: 256x).

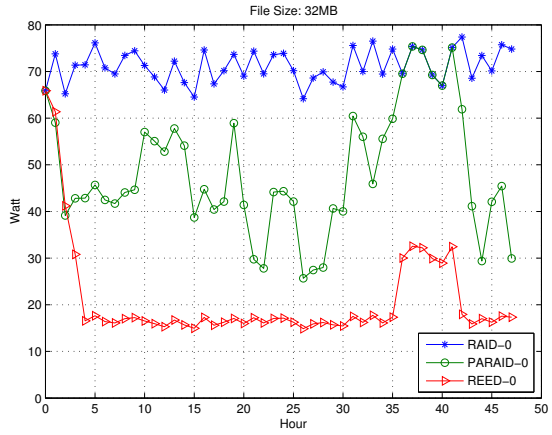

Fig. 6: Power consumption of RAID-0, PARAID-0, and REED-0. (Size of files in the Web trace: 32MB).

C. Reliability Analysis

Figs. 7-8 illustrate the annual failure rates (AFR) of RAID-0, PARAID-0, and REED-0 driven by the real-world Web trace. RAID-0's AFR behavior stands out from all four figures, primarily due to the lack of energy-saving mechanisms. PARAID-0 and REED-0 exhibit higher AFRs than those of RAID-0, because of the reduced number of power-state-transitions affecting the reliability of energy-efficient storage systems [47]. Furthermore, REED-0 lowers the AFR value of PARAID-0 thanks to the minimized number of power-transitions.

Although REED-0's system reliability is slightly lower than that of RAID-0, REED-0 is superior to RAID-0 in terms of energy efficiency. Moreover, REED-0 exhibits a lower annual failure rate than that of PARAID-0 – a well-known parallel storage system powered by an energy-saving scheme. Table IV summarizes the reliability of a parallel disk system managed by REED-0 and PARAID-0.

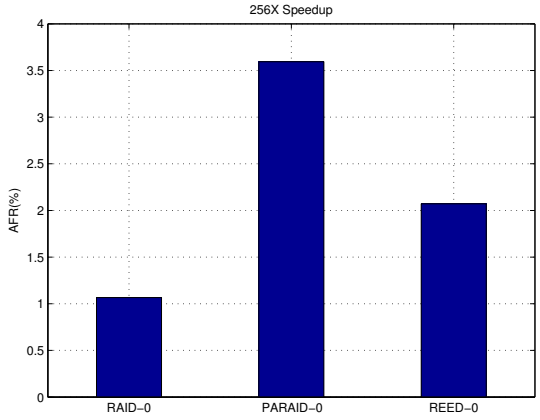

Fig. 7: Annual failure rate (AFR) of RAID-0, PARAID-0, and REED-0. (Speedup of the Web trace: 256x).

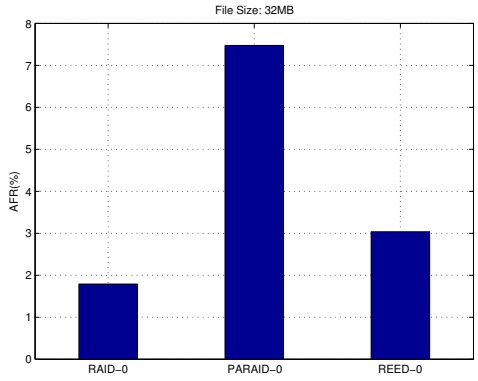

Fig. 8: Annual failure rate (AFR) of RAID-0, PARAID-0, and REED-0. (Size of files in the Web trace: 32MB).

TABLE IV: Reliability improvement of REED-0 over PARAID-0.

256x	32MB
42%	59%

VI. Related Work

A hard disk drive (HDD) is a complex dynamic system made up of various electrical, electronic, and mechanical components [44]. A handful of techniques were developed to conserve energy in a single HDD. Energy dissipation in hard drives can be reduced at the I/O level (e.g., dynamic power management[11][22] and multi-speed disks[16]), the operating system level (e.g., power-aware caching/prefetching[49][36]), and the application level (e.g. software DMP[37] and cooperative I/O[40]). The existing energy-saving techniques designed for parallel disks often rely on one of the two basic ideas -

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

power management and workload skew. Power management schemes conserve energy by turning disks into the standby mode after a period of idle time. Although multi-speed disks have not yet been widely adopted in storage systems, power management was incorporated into multi-speed disks to provide energy savings [16][15][20]. 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 [30][9][31][39].

Growing evidence shows that the existing energy-saving schemes inherently impose adverse impacts on disk systems [6][42][47]. For example, the power management schemes are likely to result in an excessive number of disk spin-downs and spin-ups, which significantly shorten hard disk lifetimes. The workload skew techniques dynamically migrate frequently accessed data to a subset of disks [23], which inevitably have a higher risk of breaking down than other disks that are usually in the standby mode. Disks storing popular data tend to have high failure rates due to extremely unbalanced workloads. Hence, the popular data disks have a strong likelihood to become a reliability bottleneck.

A malfunction of any component in a hard disk drive leads to a disk failure. Reliability — one of the key characteristics of disks — can be measured in terms of mean-time-between-failure (MTBF). More often than not, disk manufacturers investigate MTBFs of disks either by laboratory testing or mathematical modeling. Although hard drive manufacturers claim that the MTBF value of disks is longer than one million hours [34], users have experienced a much lower MTBF from field testing data [12]. More importantly, it is challenging to measure MTBF due to a wide range of contributing factors like disk age, utilization, temperature, and power-state transition frequency [12].

A handful of reliability models were 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 [29]; Elerath and Pecht proposed a flexible model for estimating reliability of RAID systems [13]; and Xin *et. al* developed a model to study disk infant mortality [43]. Unlike these reliability models tailored for conventional parallel and distributed disk systems, the *MREED* model developed by Yin *et. al* pays special attention to reliability of parallel disk systems coupled with energy-saving mechanisms [47].

VII. Conclusion

In this paper, we designed a parallel storage system called *REED*, which is highly reliable and energy-efficient in nature. *REED* incorporates solid state drives (SSDs) into disk arrays to form a hybrid system reducing energy consumption. Importantly, *REED* aims to improve energy efficiency while maintaining good reliability of hard and solid state drives. More specifically, *REED* not only minimizes the number of power-state transitions in hard disks, but also reduces frequent data updates in solid state disks. *REED* is conducive to boosting the reliability of both hard drives and solid state disks by taking fluctuating I/O workloads into account. *REED* aggressively applies SSDs to serve requests under light workloads, thereby allowing hard disks to be power-off to conserve energy. Thanks to the high bandwidth of SSDs, there is no need to frequently change hard disks' power states in the presence of fluctuating workloads. Energy savings become impossible under heavy loads. In this case, all hard drives are kept active while turning off SSDs to decrease the number of data erasures that seriously affect SSD reliability. Our experimental results confirm that compared with PARAID, our *REED* achieves high energy efficiency while maintaining good reliability of storage systems.

In the not-too-distant future, we plan to address the following open issues. First of all, we will extend *REED* to meet the needs of RAID-1 and RAID-5. We will investigate how to make *REED* cache RAID-1's popular data in SSDs while managing data replica stored in mirroring disks. When it comes to RAID-5, we will propose a way of caching data as well as parity blocks in SSDs in *REED*. Second, we intend to design an online algorithm to automatically set thresholds for L-, M-, and H-zones. The goal of the online algorithm is to reduce power management overhead. Last but not the least, besides read-intensive workloads, write-intensive scenarios will be addressed in *REED*. If the write workloads are frequently changing, then *REED* will direct writes to SSDs rather than hard drives. Otherwise, writes will be serviced by hard drives. We will develop an algorithm to determine which write requests are redirected to the SSDs.

VIII. Acknowledgment

The authors would like to extend their appreciation to Mojen Lau for proofreading the manuscript. Shu Yin's research was supported by the Hunan Province Natural Science Foundation under Grant 14JJ4025, the National Natural Science Foundation of China under Grant, the Scientific Research Foundation for the Returned Overseas Chinese Scholars under Grant [2013]1792, Ministry of Education of the People's Republic of China, and Hunan University under a start-up grant. Xiao Qin's work is 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).

References

[1] Florida state university traces. http://www.cs.fsu.edu/~awang/, June 2005.

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

[3] M. Balakrishnan, A. Kadav, V. Prabhakaran, and D. Malkhi. Differential RAID. *ACM Transactions on Storage*, (2):1–22, July 2010.

[4] B. Battles, C. Belleville, S. Grabau, and J. Maurier. Reducing data center power consumption through efficient storage. *Netapp white paper*, February 2007.

[5] D. Beaver, S. Kumar, H. Li, J. Sobel, and P. Vajgel. Finding a needle in haystack: Facebook's photo storage. In *Proc. 9th USENIX OSDI*, 2010.

[6] 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.

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.

[7] A. Bianzino, C. Chaudet, D. Rossi, and J. Rougier. A survey of green networking research. *Communications Surveys Tutorials, IEEE*, 14(1):3–20, 2012.

[8] T.-Y. Chen, T.-T. Yeh, H.-W. Wei, Y.-X. Fang, W.-K. Shih, and T.-s. Hsu. CacheRAID: An Efficient Adaptive Write Cache Policy to Conserve RAID Disk Array Energy. In *2012 IEEE Fifth International Conference on Utility and Cloud Computing*, pages 117–124, Nov. 2012.

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

[10] G. Cole. Estimating drive reliability in desktop computers and consumer electronics systems. Technical report, Seagate Personal Storage Group, Longmont, Colorado, 2000.

[11] 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*, pages 23–23, Berkeley, CA, USA, 1994.

[12] J. Elerath. Specifying reliability in the disk drive industry: No more mtbf's. In *Proc. Reliability and Maintainability Symp.*, pages 194–199, 2000.

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

[14] L. Grupp, A. Caulfield, J. Coburn, S. Swanson, E. Yaakobi, P. Siegel, and J. Wolf. Characterizing flash memory: Anomalies, observations, and applications. In *Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on*, pages 24–33, 2009.

[15] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke. Drpm: dynamic speed control for power management in server class disks. In *Computer Architecture, 2003. Proceedings. 30th Annual International Symposium on*, pages 169–179, June 2003.

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

[17] T. Iyer. Hdds return to pre-flood prices. http://www.tomshardware.com/news/Thailand-Flood-Storage-Price-SSD,22150.html, April 2013.

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

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

[20] P. Krishnan, M. P. Long, and S. J. Vitter. Adaptive disk spindown via optimal rent-to-buy in probabilistic environments. *Algorithmica*, 23(1):31–56, 1999.

[21] D. Li and J. Wang. Eeraid: energy efficient redundant and inexpensive disk array. In *Proceedings of the 11th workshop on ACM SIGOPS European workshop*, EW 11, 2004.

[22] 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*, pages 22–22, Berkeley, CA, USA, 1994.

[23] 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.

[24] A. C. Manzanares. *Energy Efficient Pre-Fetching–Models to Implementation*. "Ph.D. dissertation", Auburn University, April 2010.

[25] B. Mao, H. Jiang, S. Wu, L. Tian, D. Feng, J. Chen, and L. Zeng. HPDA: a hybrid parity-based disk array for enhanced perforamnce and reliablity. *ACM Transactions on Storage*, (1):1–20, Feb. 2012.

[26] Microsoft. Ssd extension for disksim simulation environment. http://research.microsoft.com/en-us/downloads/b41019e2-1d2b-44d8-b512-ba35ab814cd4/, March 2009.

[27] OCZ. Ocz vector sepcifications. http://www.ocz.com/consumer/vector-7mm-sata-3-ssd/specifications.

[28] OCZ. Ssd vs hdd - why solid state drive. http://www.ocz.com/consumer/ssd-guide/ssd-vs-hdd.

[29] 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.

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

[31] 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.

[32] E. Pinheiro, R. Bianchini, and C. Dubnicki. Exploiting redundancy to conserve energy in storage systems. *SIGMETRICS Perform. Eval. Rev.*, 34(1):15–26, June 2006.

[33] A. Richi. Ssd price crash:80 ¢per gb! http://h30565.www3.hp.com/t5/Mobility-Matters/SSD-price-crash-80-per-GB/ba-p/5032, February 2012.

[34] 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.*, page 1, 2007.

[35] Seagate. Data sheet: Cheetah 15k.5, mainstream enterprise hard drive storage. http://www.seagate.com/files/docs/pdf/en-GB/datasheet/disc/ds-cheetah-15k-5-gb.pdf.

[36] 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*, pages 105–114, New York, NY, USA, 2006.

[37] 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.

[38] Storagesearch.com. The top ssd companies. http://www.storagesearch.com/ssd-top10-25.html, 2013.

[39] 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*, pages 30–30, Berkeley, CA, USA, 2007.

[40] 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*, pages 117–129, New York, NY, USA, 2002.

[41] G. Wu and X. He. Delta-ftl: improving ssd lifetime via exploiting content locality. In *Proceedings of the 7th ACM european conference on Computer Systems*, EuroSys '12, pages 253–266, 2012.

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

[43] 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.

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

[45] 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.

[46] S. Yin, X. Ruan, A. Manzanares, X. Qin, and K. Li. Mint: A reliability modeling framework for energy-efficient parallel disk systems. *Dependable and Secure Computing, IEEE Transactions on*, 11(4):345–360, July 2014.

[47] 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*, pages 1–8, 2011.

[48] L. Zeng, D. Feng, J. Chen, Q. Wei, B. Veeravalli, and W. Liu. Hraid6ml: A hybrid raid6 storage architecture with mirrored logging. In *Mass Storage Systems and Technologies (MSST), 2012 IEEE 28th Symposium on*, pages 1–6, 2012.

[49] 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 *Software, IEE Proceedings-*, pages 118–118, Washington, DC, USA, Feb. 2004.

[50] X. Zhu, C. He, K. Li, and X. Qin. Adaptive energy-efficient scheduling for real-time tasks on dvs-enabled heterogeneous clusters. *J. Parallel Distrib. Comput.*, 72(6):751–763, June 2012.

Authorized licensed use limited to: Auburn University. Downloaded on November 17,2026 at 22:59:56 UTC from IEEE Xplore. Restrictions apply.