Sacrificing Reliability for Energy Saving: Is It Worthwhile for Disk Arrays?

The 22th IEEE International Parallel and Distributed Processing Symposium, 2008

Tao Xie and Yao Sun, Department of Computer Science, San Diego State University, San Diego, CA 92182

Introduction

Mainstream energy conservation schemes for disk arrays inherently affect the reliability of disks. A thorough understanding of the relationship between energy saving techniques and disk reliability is still an open problem, which prevents effective design of new energy saving techniques and application of existing approaches in reliability-critical environments. As one step towards solving this problem, this paper presents an empirical reliability model, called Predictor of Reliability for Energy Saving Schemes (PRESS). Fed by three energy-saving-related reliability-affecting factors, operating temperature, utilization, and disk speed transition frequency, PRESS estimates the reliability of entire disk array. Further, a new energy saving strategy with reliability awareness called Reliability and Energy Aware Distribution (READ) is developed in the light of the insights provided by PRESS. Experimental results demonstrate that compared with existing energy saving schemes, MAID and PDC, READ consistently performs better in performance and reliability while achieving a comparable level of energy consumption.

The PRESS model

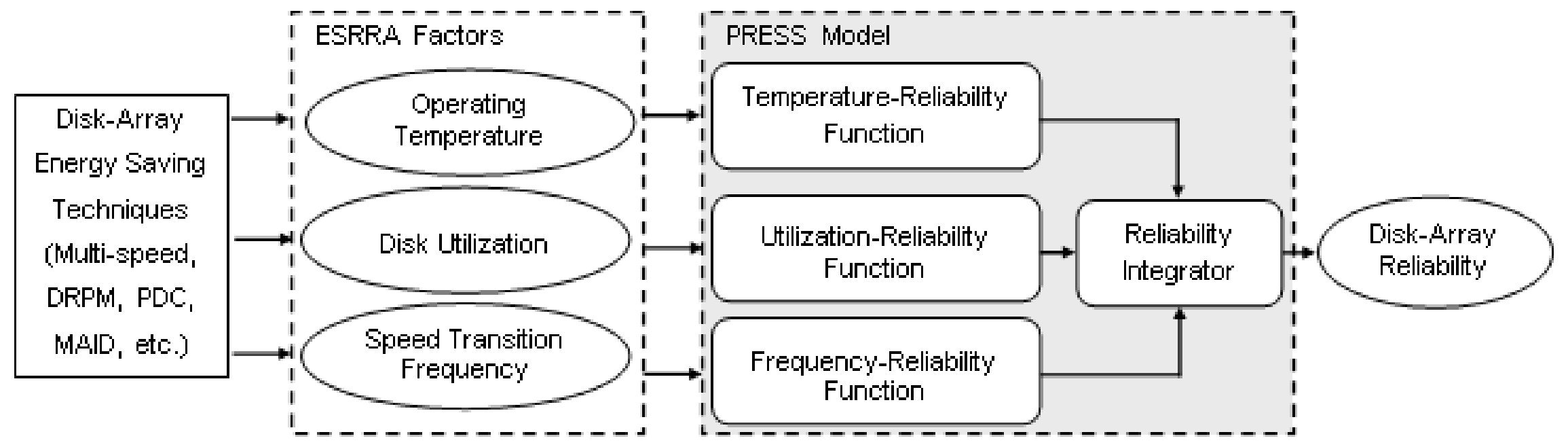
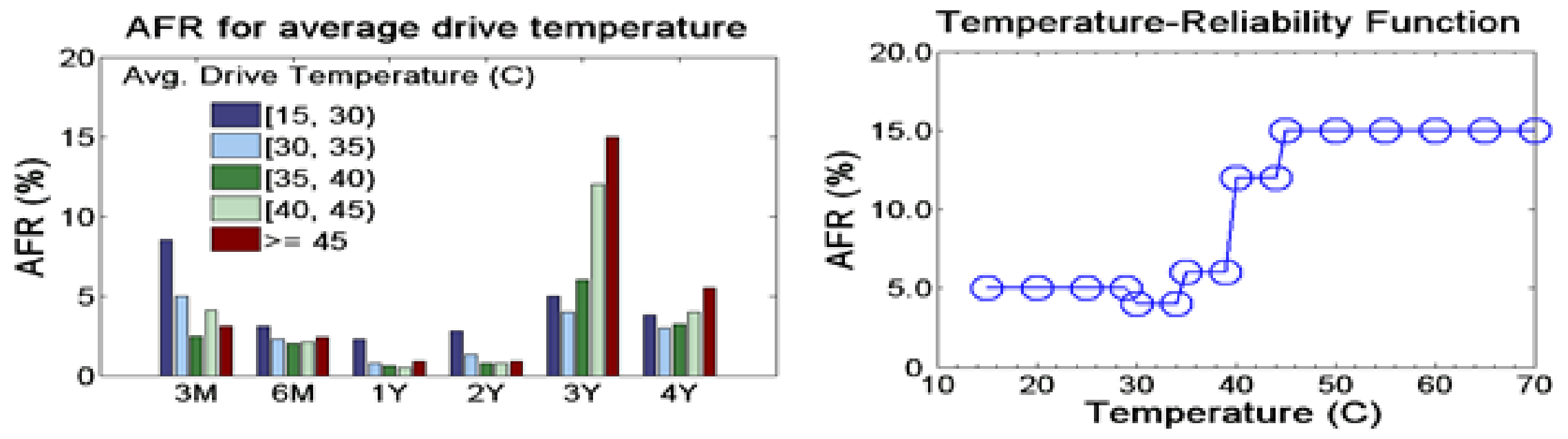

Disk-Array Energy Saving Techniques (Multi-speed, DRPM, PDC, MAID, etc.)
ESRRA Factors
Operating Temperature
Disk Utilization
Speed Transition Frequency
PRESS Model
Temperature-Reliability Function
Utilization-Reliability Function
Frequency-Reliability Function
Reliability Integrator
Disk-Array Reliability

Figure 1. Overall architecture of the PRESS model

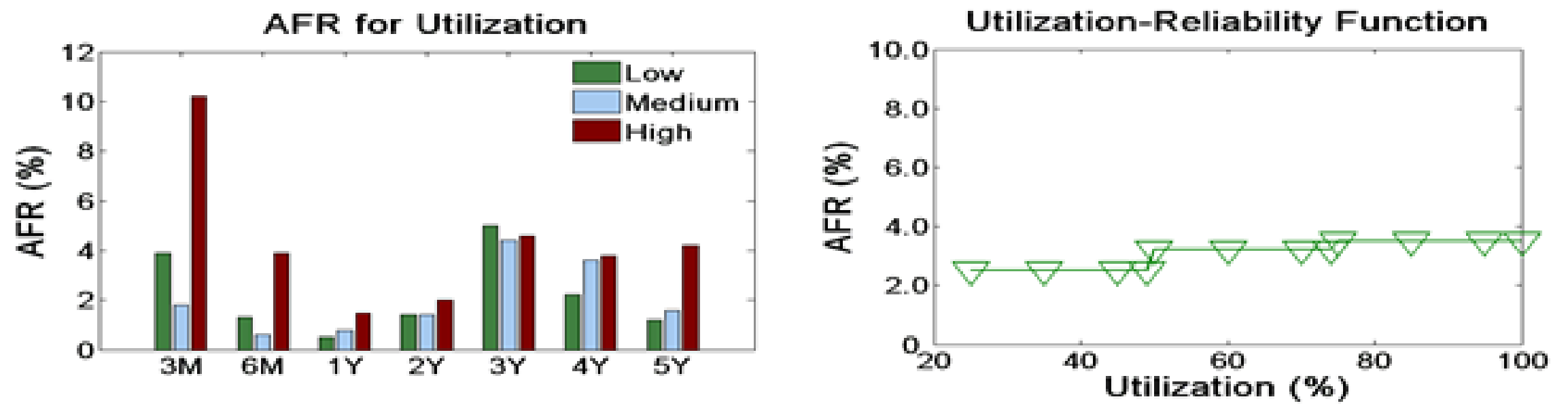
Figure 1 depicts the overall architecture of the PRESS model. Energy-saving schemes such as DRPM, PDC, and MAID inherently affect either part of the three ESRRA factors or all of them. Each of the three ESRRA factors is then fed into a corresponding reliability estimation function within the PRESS model. The PRESS model is composed of a reliability integrator module and three functions: temperature-reliability function, utilization-reliability function, and frequency-reliability function. While the former two functions are derived based on Google's results in, the last one is built from the spindle start/stop failure rate adder suggested by the IDEMA standards and the modified Coffin-Manson model.



(a) (b)

Figure 2. (a) Temperature impacts on AFR from ; (b) The temperature-reliability function

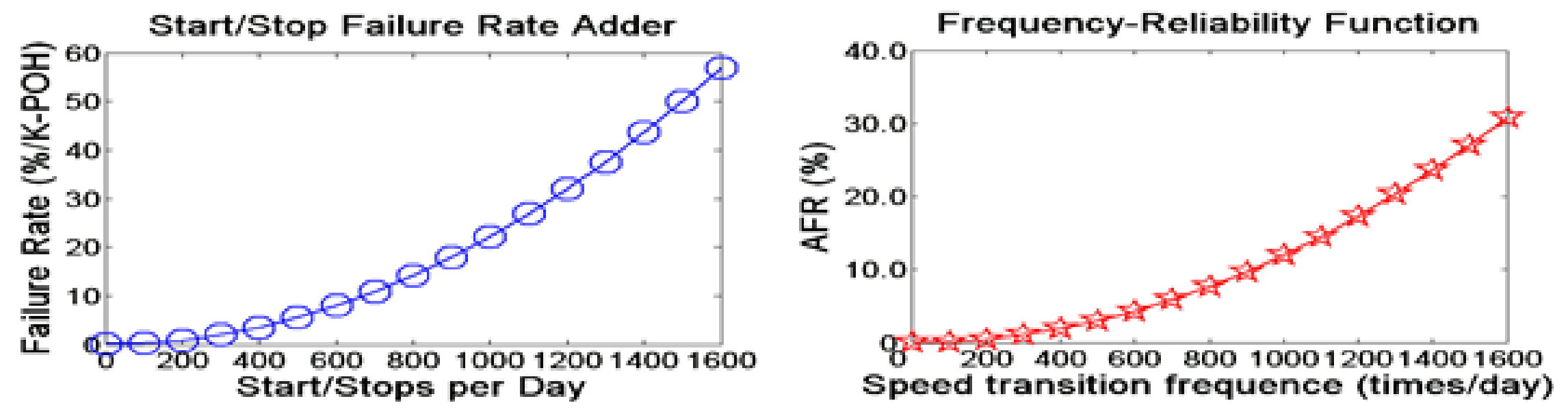
High temperature was discovered as a major culprit for a number of disk reliability problems. One such problem is off-track writes, which could corrupt data on adjacent cylinders. Even worse, spindle motor and voice coil motor running at high temperatures can lead to head crash. There are two different avenues to establishing a temperature-reliability relationship function. One is using mathematical modeling and laboratory testing techniques and the other is employing user field data. In this paper we select the latter. From the Figure 2, we can observe that First, higher temperatures are not associated with higher failure rates when disks are less than 3 years old. Second, the temperature effects on disk failure rates are salient for the 3-year-old and the 4-year-old disks, especially when temperatures are higher than 35 C.



(a) (b)

Figure 3. (a) Utilization impacts on AFR from; (b) The utilization-reliability function

Disk utilization is defined as the fraction of active time of a drive out of its total power-on-time. higher utilizations in most cases affect disk reliability negatively has been generally confirmed by two widely recognized studies. One is a classical work from Seagate, which utilized laboratory testing and mathematical modeling techniques . The other is a new breakthrough, which analyzes the utilization impacts on disk reliability based on field data from Google. Figure 3 shows that disk drives in their middle ages (2 or 3 years) are strong enough in both electronic and mechanical parts to resist the effects of higher utilizations. Therefore, AFR of disks in these two age groups has little correlation with utilization.



(a) (b)

Figure 4. (a) Start/Stop failure rate adder; (b) The frequency-reliability function

The disk speed transition frequency (hereafter called frequency) is defined as the number of disk speed transitions in one day. Establishing a frequency-reliability function is the most difficult task in this research primarily because multi-speed disks have not been largely manufactured and deployed. Thus, no result about the impacts of frequency on disk reliability has been reported so far. Our frequency-reliability function is built on a combination of the spindle start/stop failure rate adder suggested by IDEMA and the modified Coffin-Manson model. We derive our frequency-function based on Figure 4a and the modified Coffin-Manson model, which is listed as below:

$$N_f = A_0 f^{-\alpha} \Delta T^{-\beta} G(T_{\max})$$

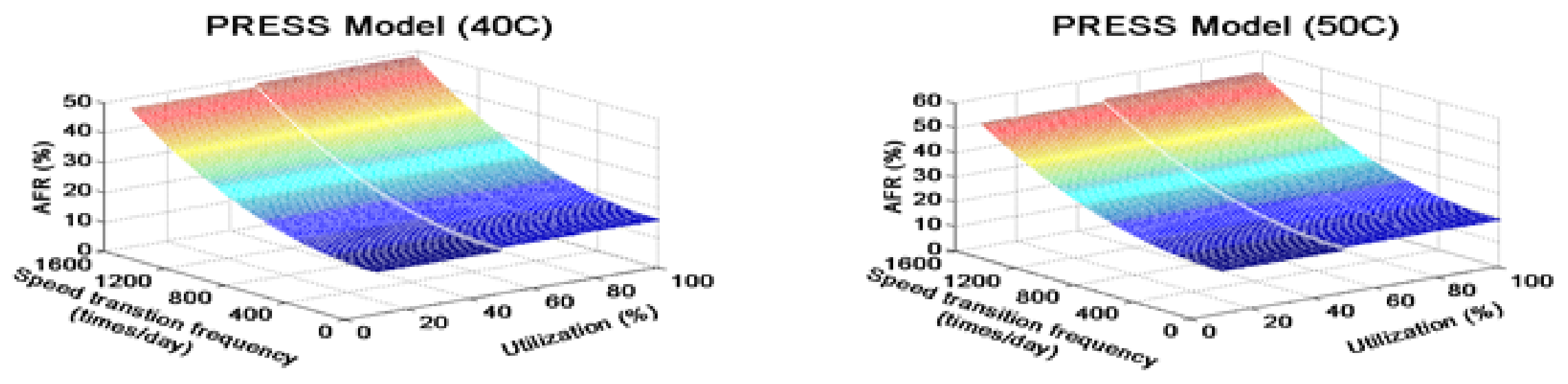
$G(T_{\max})$ is an Arrhenius term evaluated at the maximum temperature reached in each cycle. It can be calculated by the following Arrhenius equation :

$$G(T) = Ae^{(-E_a / KT)}$$

we scale down the spindle start/stop failure rate adder curve (Figure 4a) by half and change the unit of the *X* axis to times per day to obtain our frequency-reliability function (Figure 4b). The expression based on quadratic curve fitting for the reliability-frequency function is:

$$R(f) = 1.51e^{-5} f^2 - 1.09e^{-4} f + 1.39e^{-4},\ f \in [0,1600]$$

we present two 3-dimennsional figures to represent the PRESS model at operating temperature 40 C (Figure 5a) and 50 C (Figure 5b), respectively.



(a) (b)

Figure 5. (a) The PRESS model at 40 C; (b) The PRESS model at 50 C

The PRESS model yields several important insights on how to make trade-offs between energy-saving and reliability when developing energy conservation techniques for disk array systems. First, disk speed transition frequency is the most significant reliability-affecting factor among the three ESRRA factors. Second, operating temperature is the second most significant reliability-affecting factor. A high temperature can be caused by a long time running at high speed. Finally, since the differences in AFR between high utilizations and medium utilizations are slim, an uneven utilization distribution in an array should not be overly concerned.

The READ strategy

```
Input: A disk array $D$ with $n$ 2-speed disks, a collection of $m$ files in the set $F$, an epoch $P$, idleness threshold $H$, a disk maximum allowed times of speed transitions per day $S$, speed transition times for each disk $T(n)$, and the skew parameter $\theta$
Output: A file allocation scheme $X(m)$ for each epoch $P$
1. Use Eq. 4 to compute the number of popular files and the number of unpopular files
2. Use Eq. 5 to compute $\gamma$, the ratio between the number of hot disks and the number of cold disks
3. Hot disk number , cold disk number $CD = n - HD$, $d_h$=1, $d_c$=1
4. Configure $HD$ of $n$ disks to high speed mode and set $CD$ of $n$ disks to low speed mode
5. Sort all files in file size in a non-decreasing order
6. Assign all popular files onto the hot disk zone in a round-robin manner
7. Assign all unpopular files onto the cold disk zone in a round-robin manner
8. for each epoch $P$ do
9.     Keep tracking number of accesses for each file
10.    Re-sort all files in number of accesses during the current epoch
11.    Re-calculate the skew parameter $\theta$ and re-categorize popular and unpopular for each file
12.    for each previously hot file that becomes unpopular do
13.        Migrate it to the cold disk zone
14.        Update its record in the allocation scheme $X$
15.    end for
16.    for each previously cold file that becomes popular do
17.        Migrate it to the hot disk zone
18         Update its record in the allocation scheme $X$
19.    end for
20.    for each disk $d_i \in D$ do
21.        if $S/2 \le T(d_i)$      // Still has room in terms of disk speed transitions to spin down
22.            H=2H;      // Double the idleness threshold H to reduce future disk speed transitions
23.        end if
24.    end for
25. end for
```

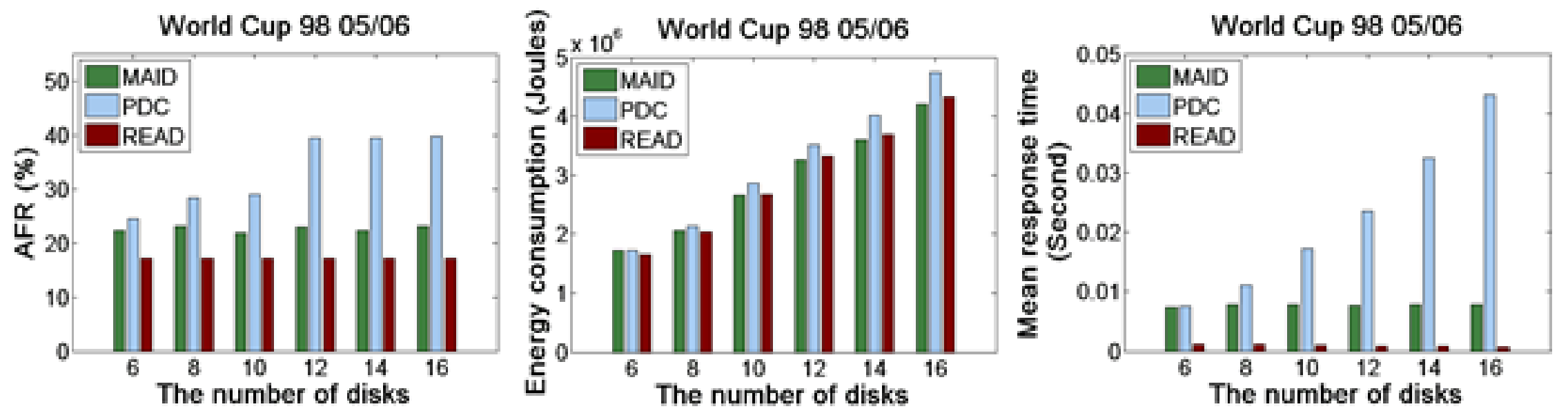
Figure 6. The READ strategy

The general idea of READ is to control disk speed transition frequency based on the statistics of the workload so that disk array reliability can be guaranteed. Also, READ employs a dynamic file redistribution scheme to periodically redistribute files across a disk array in an even manner to generate a more uniform disk utilization distribution. A low disk speed transition frequency and an even distribution of disk utilizations imply a lower AFR based on our PRESS model.

Figure 6 depicts the READ algorithm. READ assigns sorted popular files in F_p onto the hot disk zone in a round-robin manner with the first file (supposed most popular one) onto the first disk, the second file onto the second disk, and so on. Similar file assignment strategy is applied for sorted unpopular file in F_u onto the code disk zone. After all files in F have been allocated, READ launches an Access Tracking Manager (ATM) process, which records each file's popularity in terms of number of accesses within one epoch in a table called File Popularity Table (FPT). The FPT table with the latest popularity information for each file will be used later by the File Redistribution Daemon (FRD). At the end of each epoch, FRD re-orders all files based on their access times recorded during the current epoch in the FPT table and then redefine popular file set F_p and unpopular file set F_u accordingly. A hot file will be migrated to the cold disk zone if its new position in the entire re-sorted file set is out of the newly defined hot file set range. It will stay in the hot zone, otherwise. Similarly, a previous cold file will be migrated to the hot disk zone if its new ranking is within the new hot file set scope.

Performance evaluation

We developed an execution-driven simulator that models an array of 2-speed disks. The same strategy used in to derive corresponding low speed mode disk statistics from parameters of a conventional Cheetah disk is adopted in our study. The number of disks in the simulated disk array varies from 6 to 16. The performance metrics by which we evaluate system performance include *mean response time* (average response time of all file access requests submitted to the simulated 2-speed parallel disk storage system), *energy consumption* (energy consumed by the disk systems during the process of serving the entire request set, and *AFR* (Annualized failure rate of a disk array). Each disk has an AFR calculated based on the PRESS model. The highest one is used to designate the AFR of the entire disk array.



(a) Reliability (b) Energy consumption (c) Mean response time

Figure 7. An overall comparison of the three algorithms

The READ algorithm consistently outperforms MAID and PDC algorithms in reliability by up to 39.7% and 57.5%, respectively. READ constrains each disk's number of speed transitions so that it cannot be larger than *S*, which is set to 40 in our study. READ accomplishes this by gradually enlarging the idleness threshold value. In our implementation, we simply double the idleness threshold value once READ finds that a disk's current number of speed transitions reaches half of *S*.

Conclusions

In this paper, we establish an empirical reliability model PRESS, which can be utilized to estimate reliability impacts caused by the three ESRRA factors. The PRESS model is built on a state-of-the-art work and our own investigation on the relationship between disk speed transition frequency and reliability. In particular, our frequency-reliability function reveals that it is not a good idea to save energy if disk speed transition frequency is always higher than 65 times per day. Further, with the light shed by the PRESS model, we develop and evaluate a novel energy saving strategy with reliability awareness called READ. The READ strategy exploits popularity locality of I/O workload characteristics and an adaptive idleness threshold to limit each disk's speed transition times per day to provide a good reliability. Besides, it generates a more even load distribution to further alleviate reliability side-effect

Acknowledgement

This work was supported by the US National Science Foundation under Grant No. CCF-0742187.

References

[1] D. Anderson, J. Dykes, and E. Riedel, "More Than an Interface - SCSI vs. ATA," *Proc. 2nd USENIX Conf. File and Storage Technologies*,, 2003.
[2] M. Arlitt and T. Jin, *1998 World Cup Web Site Access Logs*, August 1998, Available at http://www.acm.org/sigcomm/ITA/.
[3] E. V. Carrera, E. Pinheiro, and R. Bianchini, "Conserving Disk Energy in Network Servers," *Proc. Int'l Conf. Supercomputing*, pp. 86-97, 2003