

Is Traditional Power Management + Prefetching == DRPM for Server Disks?

Vivek Natarajan Sudhanva Gurumurthi Anand Sivasubramaniam
Department of Computer Science and Engineering,
The Pennsylvania State University, University Park, PA 16802, USA
{vnataraj, gurumurt, anand}@cse.psu.edu

Abstract—The I/O subsystem on servers consumes considerable energy leading to cost, reliability and environmental concerns. Hence, there is a need for reduction of energy consumption of server disks. Most of the idle periods for server disks are shorter than the total time taken to spin down the disks and bring them back up. Traditional Power Management (TPM) schemes, which completely shut the disks down during periods of inactivity, are therefore ineffective. In a previous study, it has been shown that Dynamic Rotations per Minute (DRPM), a power management scheme that modulates disk rotation speed based on request arrival patterns is an effective solution to this problem. However, DRPM disks do not exist yet. This paper intends to evaluate both TPM schemes combined with I/O Prefetching and DRPM. Using both synthetic and real workloads and both idealistic and realistic versions of TPM, DRPM and Prefetching, we have conducted simulations which reiterate the necessity of alternate techniques such as DRPM for server power management.

I. INTRODUCTION

Data centers are required to provide high processing capacities, processing speeds, storage capacities, fault tolerance, reliability and availability. Unfortunately, these capabilities for data centers also lead to increased power consumption and cooling requirements [1]. These additional factors are of paramount importance in data center design. Storage demands for data centers would undoubtedly increase in the near future [2] and so would their power consumption requirements which is estimated to be in excess of 200W/ft² [1]. This would in turn lead to the energy costs at data centers being a substantial part of their total cost of ownership [2]. Furthermore, power consumption due to storage is a very high percentage of the overall power consumption at data centers [3]. These challenges have motivated research towards efficient energy management of server disks [4], [5], [6], [7].

Traditionally, there have been efforts towards efficient energy management of mobile/laptop devices in order to extend their battery life [8], [9], [10], [11], [12], [13], [14], [15], [16]. Most mobile/laptop disks have low spin-up and spin-down times and their traffic is usually not very I/O intensive. Energy management schemes for such devices shut the disks down during periods of inactivity, that are predicted based on prior history of lengths of idle periods.

On the other hand, most server disks have high spin-up and spin-down times and their traffic is much more I/O intensive. Response time degradation toleration levels are also

quite low in server environments [5]. Therefore, Traditional Power Management (TPM) schemes, which are effective in mobile/laptop environments, are rendered ineffective in server environments, even if accurate prediction of the start and duration of idle periods is possible [5].

In a previous study, it has been shown that Dynamic Rotations per Minute (DRPM), [4], [17], [6], [7] a power management scheme that modulates disk rotation speed based on request arrival patterns, is an effective solution to this problem in server environments. However, design of a DRPM disk would be complex and such disks are not yet available in the market.

Therefore, in order to improve existing TPM schemes, the idle period lengths for server disks need to be extended. One way to extend the idle period lengths is to predict, prefetch and cache future requests. Prefetching can create I/O burstiness if performed aggressively and accurately. Using both synthetic and real storage workloads and both idealistic and realistic versions of TPM, DRPM and prefetching, we conduct simulations to evaluate the effectiveness of TPM coupled with prefetching vis-a-vis DRPM.

The rest of this paper is organized as follows. In the next section, a comprehensive overview of various disk power management schemes is provided. Subsequently, the simulation environment is described and the simulation results are examined in section 3. Finally, the contributions of this work are summarized in section 4.

II. DISK POWER MANAGEMENT OVERVIEW

Many current hard disks offer different power modes of operation. A disk is in active mode when it is servicing a read or write request. It is in idle mode when it is spinning but not servicing any requests. It is in standby mode when it is neither spinning nor servicing requests. Active and idle modes of operation for a disk consume the highest amount of power. The standby mode consumes comparatively less power. To transition to the standby mode from the active mode, the disk needs to be spun down and to transition from the standby mode to the active mode, the disk needs to be spun up.

Figure 1 shows various disk power modes, the power consumed in these modes, the time taken and the power consumed to transition between these modes. This data is for

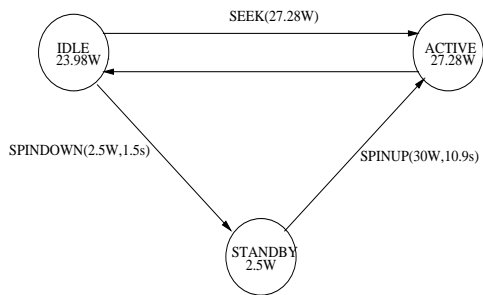


Fig. 1. Disk Power Modes

an IBM Ultrastar 36Z15 [18], [7] type disk that is used in several server environments.

Periods of time between consecutive read/write request arrivals to the disk are called idle periods. If an idle period is detected to be long enough to outweigh disk spin-up and spin-down times and power costs, it is considered to be a suitable idle period for spin-down of the disk. Managing the energy consumption of disks involves detecting lengths of idle periods and spinning the disk down during idle periods that have the potential to facilitate energy savings. If a request arrives when the disk is in the standby mode, it needs to be spun up to the active mode before it could service the request that incurs additional latency and power costs.

Idle time predictors facilitate detection of idle periods. They track the history of past idle period lengths to make predictions about lengths of future idle periods. Golding et al. [13] have conducted a detailed study of idle time predictors and their effectiveness in disk power management. During idle periods that were predicted to last longer than the total spin-down and spin-up times, the disk could be spun down and proactively spun up before its next request arrival, provided that prediction of lengths of idle periods is accurate. However, not many prior studies have focused on this methodology. Lu et al. [10] provide an experimental comparison of several disk power management schemes proposed in literature on a single disk platform.

Broadly, disk power management schemes fall into two categories:

- 1) Traditional Power Management.
- 2) Dynamic Rotations per Minute.

A. TPM

A disk power management scheme would be effective if it ensures that the disk is in standby mode as often as possible. Traditional Power Management (TPM) schemes transition the disk to the standby mode during idle periods that last longer than a threshold value. This value is set to less than 20 seconds for a mobile hard disk drive [19]. Alternatively, this value could also be adaptively varied during execution of programs [8], [9]. However, a transition from the standby mode to the active mode needs to take place before the disk is able to service any subsequent requests. It is clear that TPM would be efficient if the frequency of occurrence of long idle periods is high.

B. DRPM

Dynamic Rotations per Minute (DRPM) [4] is a scheme that dynamically modulates disk angular velocity to save the energy expended in the spindle motor driving the disk platters. The disk spindle motor angular velocity directly impacts the idle power consumption of the disk that is a very high percentage of its total power consumption [4]. Specifically, the idle power consumption of the disk has a quadratic relation with the rotation speed of the disk [4]. To take advantage of this fact, DRPM proposes a range of active/idle modes of operation for a disk in addition to the standby mode. Low RPM modes consume less power than high RPM modes and hence, during periods of idleness, instead of spinning at the highest RPM, the disk could transition to any one of several low RPM modes depending on the lengths of idle periods to save energy.

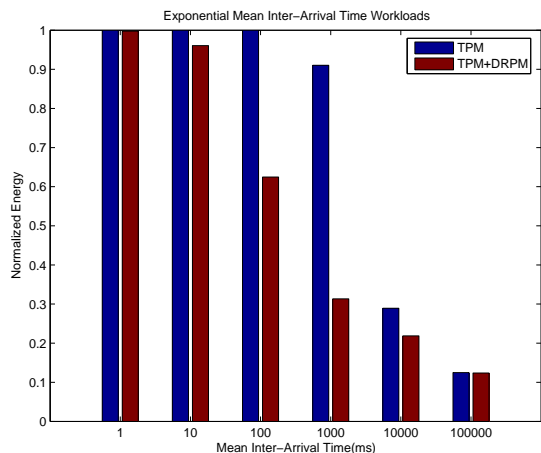
The time taken for the disk to service a read/write request is the sum of the seek time, the rotational latency and the data transfer time. The rotational latency and the data transfer time have a linear relation to the disk's angular velocity. The seek time is independent of the disk's angular velocity. Hence, the time taken to service requests at low RPM modes is higher than that in high RPM modes. The disk should service requests at the highest possible RPM mode to minimize the service time. The time taken to transition from one RPM mode to another has a linear relation to the difference in their RPM ratings [4]. Since the power consumed has a quadratic relation to the rotation speed of the disk, and both the request service time and the mode transition time have a linear relation to the rotation speed of the disk, it is possible to benefit more from energy savings than the loss in performance with DRPM.

C. TPM versus DRPM

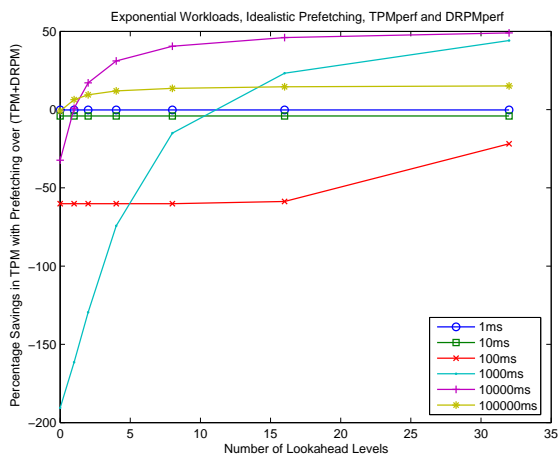
Several components of the disk such as the spindle motor, which spins the platters, the actuator, which moves the disk read/write head, the disk cache and the electrical components contribute to its overall power consumption. Of these, the spindle motor accounts for a major fraction of the power consumption [4]. Both TPM and DRPM intend to optimize the spindle motor power consumption.

TPM might require longer idle periods to spin the disk down, remain in the standby mode and to spin the disk back up without delaying subsequent disk requests. DRPM is more fine-grained to exploit shorter idle periods to save energy due to the existence of intermediate power modes. Also, with DRPM, the disk need not be spun up to its full speed before servicing requests as is done in TPM. Instead, the disk could service requests at one of the intermediate power modes. Opting to service the request at a speed less than full speed would stretch the request service time although the transition time to that power mode would be lower than the transition time to the standby mode.

The energy savings, either with TPM or with DRPM, is directly related to the distribution of the lengths of idle periods for a disk. An ideal version of TPM that provides maximum energy savings and always services requests at full speed is



(a) Normalized TPM and (TPM+DRPM) Energy Consumption for Exponential Workloads



(b) Percentage Savings in TPM with Prefetching over (TPM+DRPM) for Exponential Workloads

Fig. 2.

called perfect TPM or *TPM_{perf}* [4]. In this scheme, the disk transitions to the standby mode only if the idle period is long enough to accommodate both spin-down and spin-up times and if the total energy for the idle period is minimized. An ideal version of DRPM that provides maximum energy savings and always services requests at full speed is called perfect DRPM or *DRPM_{perf}* [4]. In this scheme, the disk transitions to a low power mode only if the idle period is long enough to accommodate both ramp-up and ramp-down times and if the total energy for the idle period is minimized. Both these schemes assume perfect knowledge of future idle periods.

A *Combined* scheme determines which of the two schemes, TPM or DRPM, provides maximum energy savings for an idle period and implements the chosen scheme for that

idle period. We shall henceforth refer to the *Combined* scheme as (TPM+DRPM). The reason we have considered (TPM+DRPM) is due to the fact that if DRPM is feasible for a disk, (TPM+DRPM) should also be feasible for the disk.

To investigate the potential benefits of both *TPM_{perf}* and *DRPM_{perf}*, we have performed an experiment to simulate *TPM_{perf}* and *DRPM_{perf}* using random block access patterns, request sizes and read/write behavior. We have used exponential mean inter-arrival time synthetic workloads for this experiment. As is well understood, exponential arrivals model a purely random Poisson process and to a large extent model a regular traffic arrival behavior (without burstiness). Figure 2(a) is a plot of TPM and (TPM+DRPM) total energy consumption normalized with respect to full speed energy consumption for exponential workloads with different mean inter-arrival times. The left bar for each workload depicts the TPM total energy consumption and the right bar depicts the (TPM+DRPM) total energy consumption.

It is apparent from Figure 2(a) that both TPM and (TPM+DRPM) consume similar amounts of energy for very long and very short mean inter-arrival time workloads. If idle periods are very long, TPM performs better than DRPM. This is so because TPM completely stops spinning the disk during such idle periods whereas DRPM spins the disk, albeit at a low speed. If idle periods are very short, neither TPM nor DRPM has the opportunity to transition the disk to a low power or standby mode. But, (TPM+DRPM) outperforms TPM for workloads with mean inter-arrival times in the entire range between these two extreme values. TPM does not provide much scope for energy savings for these workloads since most of the idle periods are not long enough to enable the disk to transition to the standby mode. But, DRPM transitions the disk to one of several intermediate low power modes and hence saves energy.

Very simple versions of a DRPM type disk drive have started appearing in the market. Hitachi's Deskstar 7K400 [20] is an example of one such disk drive that provides both power and acoustics management. It has four power modes, the Normal mode, the Standby mode and two additional low power modes called the Unload mode and the Low RPM mode. Each of these modes has a different rate of energy consumption and recovery time to the Normal mode which is where all requests are serviced. Still, full fledged DRPM disk design is complex and is more of only a proposed technique. Frequent disk spin-up and spin-down operations decrease the mean time between failures for server disks. Other issues with respect to the physical realization of a DRPM disk such as maintaining the read/write head fly height, read/write head positioning servo design and data channel design have been discussed in some detail in [4].

All the discussions in this section suggest that, either a different power management scheme other than TPM and DRPM should be implemented, or the effectiveness of TPM itself should be improved. Figure 2(a) suggests that in order to improve energy savings with TPM, idle periods need to be extended to enable the disk to transition to the standby mode

Parameter	Value
Number of RPM Levels	5
Maximum RPM Value	15000
Minimum RPM Value	3000
Spin-up Power	30 W
Spin-down Power	2.5 W
Active Power	27.28 W
Seek Power	27.28 W
Idle Power @15000 RPM	23.98 W
Idle Power @3000 RPM	5.89 W
Standby Power	2.5 W
Spin-up Time	10.9 sec
Spin-down Time	1.5 sec
Time taken to transition to the 12000 RPM power mode from full speed	145.8 ms
Time taken to transition to the 9000 RPM power mode from full speed	291.6 ms
Time taken to transition to the 6000 RPM power mode from full speed	437.4 ms
Time taken to transition to the 3000 RPM power mode to full speed	583.2 ms
Disk Controller Cache Size	16 MB

TABLE I
SIMULATION PARAMETERS

more often. To achieve this end, TPM schemes could be coupled with prefetching. The latter tends to create burstiness if performed aggressively and accurately that leads to stretching of idle periods. In the rest of the paper, we leverage TPM with prefetching to evaluate its effectiveness with respect to DRPM.

III. EXPERIMENTAL SETUP, WORKLOADS AND RESULTS

In this section, we describe the simulation platform, workloads and present the results of combining TPM and prefetching and compare it with DRPM.

A. Simulation Environment

We have conducted all experiments using the DiskSim simulator infrastructure [21] augmented with a disk power model [4]. DiskSim provides a large number of timing and configuration parameters to specify the disks, controllers and buses for the I/O interface that has been shown to be accurate [22]. The disk power model records the energy consumption of the disks during operations such as data transfers, seeks or when just idling. It also accounts for queuing and service delays caused by changes in the RPM of the disks. We have also incorporated a realistic sequential prefetching scheme on top of the simulator’s cache module. We have used disk controller caches within the simulator and have set the disk controller cache size to 16 MB [23].

A complete list of our simulation parameters are listed in Table I. As mentioned before, this data is for an IBM Ultrastar 36Z15 [18], [7].

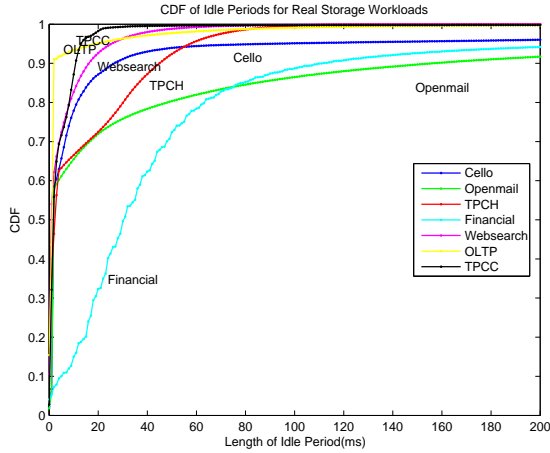
B. Exponential Workloads, Prefetchperf, TPMperf and DRPMperf

If perfect knowledge of future requests is available, prefetch accuracy would be 100 % (Prefetchperf). We performed an experiment to evaluate the performance of a combination of TPMperf and Prefetchperf with respect to that of DRPMperf

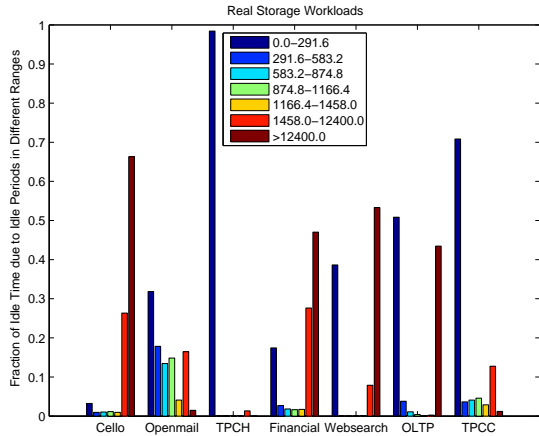
for exponential mean inter-arrival time workloads. We obtained the idle period profiles for these workloads using the simulator. These idle period profiles contain information about idle periods of all the disks. We simulated Prefetchperf for various lookahead levels on these idle periods profiles to obtain new idle period profiles assuming 100 % accuracy in prediction of future requests. Lookahead levels are the levels of prefetching in terms of disk requests. We then simulated TPMperf and DRPMperf on these new idle period profiles to estimate their total energy consumption with TPM and (TPM+DRPM).

The total energy consumption is the sum of the ramp-down, hold, ramp-up and active energy consumption. The ramp-down energy is the energy consumed to transition the disk from full speed to either one of the intermediate low power modes or to the standby mode. This energy might be different from the spin-down energy which is the energy consumed to transition the disk from full speed to the standby mode. The ramp-down power is the power rating of the mode the disk transitions to. The hold energy is the energy consumed by the disk to remain idle at a power mode. The ramp-up energy is the energy consumed to transition the disk from either one of the intermediate low power modes or the standby mode to full speed. This energy might be different from the spin-up energy which is the energy consumed to transition the disk from the standby mode to full speed. The ramp-up power is the power rating at full speed. The active energy is the active power times the active time for the disk.

Figure 2(b) is plot of percentage savings in TPM total energy consumption with prefetching over (TPM+DRPM) total energy consumption for various lookahead levels for exponential workloads. For very short (1 ms,10 ms) mean inter-arrival time workloads, TPM energy consumption is very close to (TPM+DRPM) energy consumption for various lookahead levels. This is so because most of the idle periods before prefetching are not long enough to be exploited either by TPM or by DRPM and most of the idle periods after prefetching are not long enough to be exploited by TPM. For very long (100000 ms) mean inter-arrival time workloads, TPM performs uniformly slightly better compared to (TPM+DRPM) for various lookahead levels. This is so because most of the idle periods before prefetching are long enough to be exploited equally well by both TPM and DRPM and most of the idle periods after prefetching are long enough to be exploited by TPM. For the 100 ms mean inter-arrival time workload, TPM does not break-even with (TPM+DRPM) even for higher lookahead levels although the gap between them reduces considerably for very high lookahead levels. This gap before prefetching is considerable since DRPM is able to exploit most of the idle periods that TPM is not able to exploit. After prefetching, this gap reduces since TPM is able to exploit most of the idle periods. For the 1000 ms mean inter-arrival time workload, with close to 12 lookahead levels, TPM breaks even with (TPM+DRPM) and its performance scales well with further increase in the number of lookahead levels for similar reasons. For the 10000 ms mean inter-arrival



(a) Cumulative Distribution Function (CDF) of Idle Periods for Real Workloads



(b) Fraction of Total Idle Time due to Idle Periods in Different Ranges for Real Workloads

Fig. 3.

time workload, with close to 2 lookahead levels, TPM breaks even with (TPM+DRPM) and its performance scales well with further increase in the number of lookahead levels, again for similar reasons.

It is evident that the number of lookahead levels required for TPM to break-even with (TPM+DRPM) is different for different workloads. We have seen that the workloads with mean inter-arrival time longer than 1000 ms provide good energy savings with TPM and sufficiently high lookahead levels.

C. Real Workloads

The real storage workloads we have used are described below.

Real Workload	Storage	Mean Arrival Time	Inter-90th Percentile of Inter-Arrival Times
HPL Cello 99		278.35 ms	27 ms
HPL Openmail		70.73 ms	156 ms
TPC-H		14.77 ms	45 ms
Umass Financial		222.36 ms	112 ms
Umass Websearch		15.78 ms	17 ms
OLTP		9.29 ms	2 ms
TPC-C		6.83 ms	11 ms

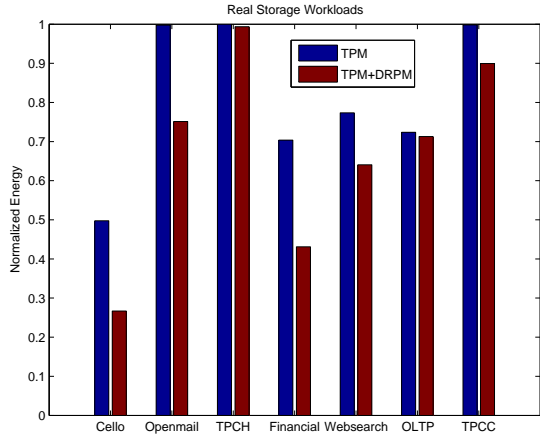
TABLE II

MEAN AND 90TH PERCENTILE OF INTER-ARRIVAL TIMES FOR REAL STORAGE WORKLOADS

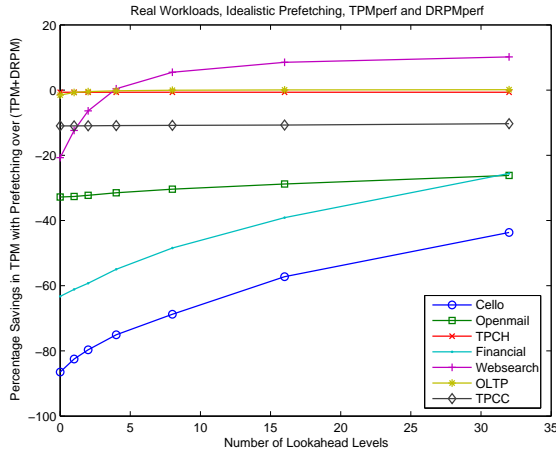
- **HPL Cello 99** [24] was collected on a news server named Cello at HP labs in 1999. Cello was a K570 class machine with 4 CPUs running HP-UX 10.20 with about 2GB of main memory.
- **HPL Openmail** [25] was run on Atlanta Response Center OpenMail Servers with a 640GB message store on EMC 3700 disk drives.
- **TPC-H** [26], [5] is a benchmark that is used to capture decision-support transactions on a database. There are 22 queries in this workload, and these queries typically read the relational tables to perform analysis for decision-support. The workload was collected on an IBM Netfinity SMP server with 8700 Mhz Pentium III processors 15 IBM Ultrastar 10K RPM disks running EEE DB-2 on Linux.
- **Umass Financial** [27] was obtained by running OLTP applications in a financial institution.
- **Umass Websearch** [28] was obtained from a popular search engine.
- **OLTP** [29], [7] is an On-Line Transaction Processing benchmark that was collected from a VI-attached database storage system connected to a Microsoft SQL Server via a storage area network. The Microsoft SQL Server Client connects to the Microsoft SQL Server via Ethernet and executes the TPC-C benchmark for 2 hours.
- **TPC-C** [30], [5] is an On-Line Transaction Processing (OLTP) benchmark. It simulates a set of users who perform transactions such as placing orders, checking the status of an order etc. Transactions in this benchmark are typically short, and involve both read and update operations. The tracing was performed for a 20-warehouse configuration with 8 clients. The traced system was a 2-way Dell PowerEdge SMP machine with Pentium-III 1.13 GHz processors with 4 10K RPM disks running IBM's EEE DB-2 [31] on the Linux operating system.

D. Real Workloads, Prefetchperf, TPMperf and DRPMperf

The cumulative distribution function (CDF) of the idle periods for the real workloads is shown in Figure 3(a). Figure 3(b) is a plot of the fraction of total idle time due to idle periods in different ranges for these workloads. These ranges have been chosen based on the times taken to transition from full speed to each of the intermediate power modes



(a) Normalized TPM and (TPM+DRPM) Energy Consumption for Real Workloads



(b) Percentage Savings in TPM with Prefetching over (TPM+DRPM) for Real Workloads

Fig. 4.

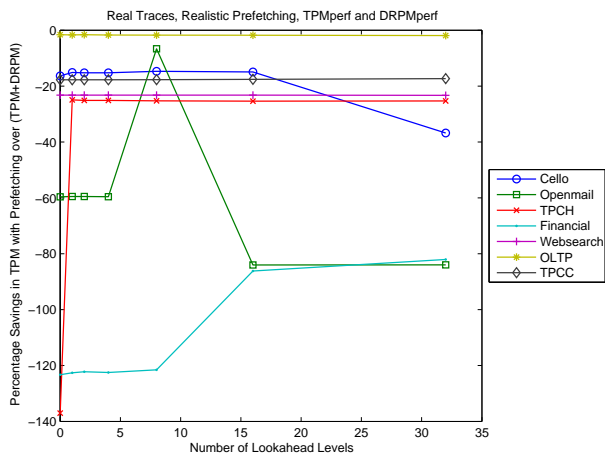
and to the standby mode with (TPM+DRPM) which are also listed in Table I. Figure 4(a) is a plot of TPM and (TPM+DRPM) total energy consumption normalized with respect to full speed energy consumption for real workloads. The left bar for each workload depicts the TPM total energy consumption and the right bar depicts the (TPM+DRPM) total energy consumption. We notice that for HPL Cello 99, HPL Openmail, Umass Financial and Umass Websearch workloads, there is a considerable gap between TPM and (TPM+DRPM) that could be bridged with prefetching. For all the other real workloads, TPM energy consumption is very close to (TPM+DRPM) energy consumption. Figure 4(b) is a plot of percentage savings in TPM total energy consumption with prefetching over (TPM+DRPM) total energy consumption for

various lookahead levels.

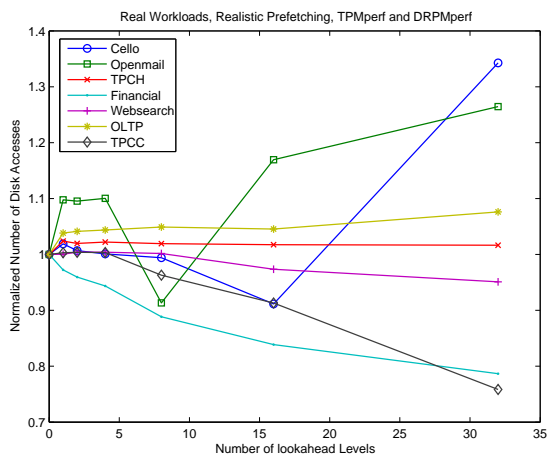
For the HPL Cello 99 workload, the gap between TPM and (TPM+DRPM) shortens with higher lookahead levels. This is so because the fraction of total idle time due to longer idle periods (1166.4 ms and higher) is high for this workload. With prefetching, these longer idle periods are further extended and are effectively exploited by TPM, especially for higher lookahead levels. The Umass Financial workload shows a similar trend for similar reasons. The fraction of total idle time due to very short idle periods (0.0 ms to 291.6 ms) is very high for The TPC-H workload. TPM and (TPM+DRPM) provide uniformly similar energy savings for various lookahead levels since neither TPM nor DRPM is able to exploit these very short idle periods and even with aggressive prefetching, these idle periods do not extend enough to be exploited by TPM.

The trend is similar for the TPC-C workload. The fraction of total idle time due to very short idle periods (0.0 ms to 291.6 ms) is very high for this workload. There is almost a constant gap between TPM and (TPM+DRPM). This is due to the idle periods in the 291.6 ms to 1166.4 ms range that are exploited by DRPM but not by TPM. This gap remains constant even with higher lookahead levels since the idle periods do not extend enough to be exploited by TPM. For the HPL Openmail workload, the gap between TPM and (TPM+DRPM) is approximately 30 percent without prefetching. This is so because DRPM is able to exploit the idle periods in the 291.6 ms to 1166.4 ms range that form a substantial fraction of the total idle time for this workload. TPM is not able to exploit these idle periods. Furthermore, this gap does not reduce much with higher lookahead levels since the idle periods are not extended enough to be exploited by TPM.

For the Umass Websearch workload, the gap between TPM and (TPM+DRPM) is approximately 20 percent without prefetching. TPM breaks even with (TPM+DRPM) for close to 4 lookahead levels and energy savings scale further with increasing lookahead levels. This is so because the fraction of total idle time due to very long idle periods is considerable for this workload that are exploited by TPM for increasing lookahead levels. For the OLTP workload, the gap between TPM and (TPM+DRPM) is negligible for various lookahead levels. This is understandable for zero lookahead levels since this workload has idle periods primarily in two ranges, between 0.0 ms and 291.6 ms and beyond 12400.0 ms. The former range is exploited neither by TPM nor by DRPM. The latter range is exploited both by TPM and (TPM+DRPM). Although the fraction of the total idle time due to idle periods longer than 12400.0 ms is high for this workload, even with higher lookahead levels, both TPM and (TPM+DRPM) provide similar amounts of energy savings. The reason for this could be due to the fact that the longer idle periods created as a result of prefetching are not exploited any better by TPM than are the longer idle periods before prefetching by (TPM+DRPM).



(a) Percentage Savings in TPM with Prefetching over (TPM+DRPM) for Real Workloads



(b) Normalized Number of Disk Accesses for Real Workloads

Fig. 5.

E. Real Workloads, Realistic Prefetching, TPM_{perf} and DRPM_{perf}

In the previous subsection, we assumed the existence of an idealistic prefetcher that has perfect knowledge of the future and hence performs accurate prefetching. In practice, such a prefetcher does not exist. Consequently we used a realistic prefetching scheme in our experiments. We performed sequential read-ahead in our simulator for every request reaching the disk and also stored the prefetched data in the disk controller cache in this process. The effectiveness of this sequential prefetching scheme depends upon the extent of sequentiality of the workloads [32] and the controller cache behavior and management [7]. We once again performed the TPM_{perf} and DRPM_{perf} analysis on the idle period profiles that we obtained

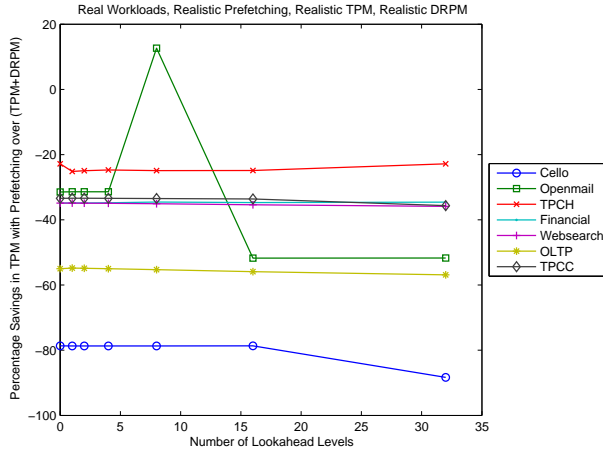
with the realistic prefetcher.

Figure 5(a) is a plot of percentage savings in TPM total energy consumption with prefetching over (TPM+DRPM) total energy consumption for various lookahead levels. Figure 5(b) is a plot of the number of disk accesses for various lookahead levels normalized with respect to the number of disk accesses without lookahead. For the HPL Cello 99 workload, TPM energy consumption is almost at a constant offset from the (TPM+DRPM) energy consumption for up to 16 lookahead levels and for 32 lookahead levels, this gap increases. This is so because for 32 lookahead levels, the number of disk accesses is much higher than the number of disk accesses with no lookahead. For the HPL Openmail workload, TPM energy consumption is more than the (TPM+DRPM) energy consumption for moderate (up to 4) lookahead levels. TPM performs almost as well as (TPM+DRPM) for 8 lookahead levels. This is so because for 8 lookahead levels, the number of disk accesses is much lower than the number of disk accesses with no lookahead. However, for higher (16 and 32) lookahead levels, the number of disk accesses increases again and the energy consumption for TPM also increases.

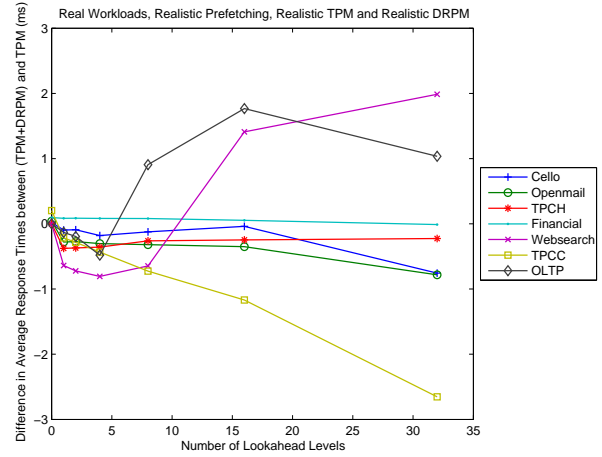
For the Umass Financial workload, the TPM energy consumption is more than the (TPM+DRPM) energy consumption but the gap between the two reduces with increasing lookahead levels since the number of disk accesses reduces with increasing lookahead levels. For the TPC-H workload, the TPM energy consumption is almost at a constant offset from (TPM+DRPM) energy consumption for various lookahead levels. The number of disk accesses is similar for various lookahead levels for this workload. The trend is similar for the Umass Websearch and the OLTP workloads. For the TPC-C workload, the TPM energy consumption is almost at a constant offset from the (TPM+DRPM) energy consumption for various lookahead levels. Even though the number of disk accesses reduces with increasing lookahead levels for this workload, this reduction does not cause a major change in the TPM energy consumption.

F. Real Workloads, Realistic Prefetching, Realistic TPM and Realistic DRPM

In the previous subsection, we performed the TPM_{perf} and DRPM_{perf} analysis on the idle period profiles that we obtained with the realistic prefetching scheme. Subsequently, we performed experiments with realistic TPM and DRPM schemes. The realistic DRPM scheme we used in this experiment is a heuristic DRPM algorithm [4] that dynamically modulates disk speed by setting tolerance levels for response time degradation that finally leads to amplification in power savings. In this scheme, the array controller communicates a set of operating RPM values to the individual disks based on how the system response time evolves and subsequently each disk uses local information to decide on the RPM transitions. The realistic TPM scheme we used in this experiment transitions the disk from the idle mode to the standby mode if an idle period lasts longer than 20 seconds [19] and transitions the disk back to the active mode upon the next request arrival.

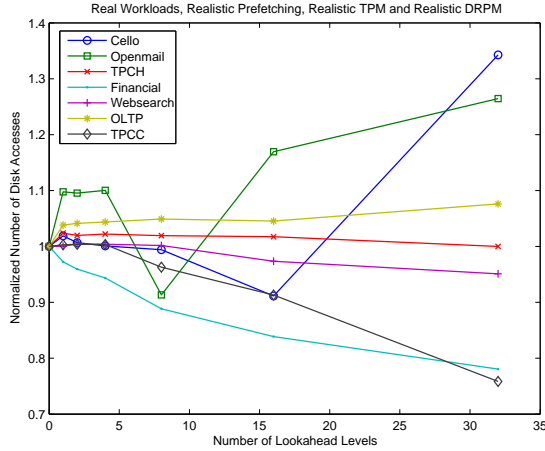


(a) Percentage Savings in TPM with Prefetching over (TPM+DRPM) for Real Workloads



(a) Difference in Average Response Times between (TPM+DRPM) and TPM for Real Workloads

Fig. 7.



(b) Normalized Number of Disk Accesses for Real Workloads

Fig. 6.

Figure 6(a) is a plot of percentage savings in TPM total energy consumption with prefetching over (TPM+DRPM) total energy consumption for various lookahead levels. Figure 6(b) is a plot of the number of disk accesses for various lookahead levels normalized with respect to the number of disk accesses without lookahead. Figure 7(a) is a plot of the difference in average response times between (TPM+DRPM) and TPM for various lookahead levels. For the HPL Cello 99 workload, the TPM energy consumption is almost at a constant offset from the (TPM+DRPM) energy consumption for up to 16 lookahead levels and for 32 lookahead levels, this gap increases. This is so because for 32 lookahead levels, the number of disk accesses is much higher than the number of disk accesses with no lookahead for this workload. The average response time also increases slightly for 32 levels of lookahead due

to higher number of disk accesses. For the HPL Openmail workload, the TPM energy consumption is more than the (TPM+DRPM) energy consumption for moderate (up to 4) lookahead levels. TPM performs better than (TPM+DRPM) for 8 lookahead levels. This is so because for 8 lookahead levels, the number of disk accesses is much lower than the number of disk accesses with no lookahead for this workload. However, for higher (16 and 32) lookahead levels, the number of disk accesses increases again and the energy consumption for TPM also increases. For higher lookahead levels, the average response time also increases for this workload due to higher number of disk accesses.

For the Umass Financial workload, the TPM energy consumption is more than the (TPM+DRPM) energy consumption and the gap between the two remains almost constant with increasing lookahead levels. The number of disk accesses reduces with increasing lookahead levels for this workload but this reduction does not cause a major change in either the energy consumption or the average response time. For the TPC-C workload, the number of disk accesses decreases with higher lookahead levels. The average response times increase with higher lookahead levels. However, the TPM energy consumption value is almost at a constant offset from the (TPM+DRPM) energy consumption value for various lookahead levels. For the TPC-H workload, the TPM energy consumption is almost at a constant offset from (TPM+DRPM) energy consumption for various lookahead levels. The number of disk accesses is similar for various lookahead levels for this workload and the average response time values are also similar. For the Umass Websearch workload, the number of disk accesses decreases with higher lookahead levels and this causes the average response times to improve although the energy consumption does not change considerably with

various lookahead levels. For the OLTP workload, the average response time value for 16 lookahead levels is lower than that for 32 levels of lookahead due to slightly increased number of disk accesses from 16 levels to 32 levels. However, the TPM energy consumption value is almost at a constant offset from the (TPM+DRPM) energy consumption value for various lookahead levels.

We notice that for most of the real workloads, with realistic prefetching, the gap between TPM and (TPM+DRPM) is much higher for realistic TPM and DRPM schemes than for *TPMperf* and *DRPMperf*. This is due to the effectiveness of the realistic DRPM scheme over *DRPMperf* in terms of saving energy.

IV. CONCLUSIONS

This paper has conducted a thorough examination of Traditional Power Management (TPM) schemes and Dynamic Rotations per Minute (DRPM) in terms of their effectiveness in saving energy for server disks. Simulation studies were conducted using DiskSim employing both synthetic and real storage workloads. While theoretically, prefetching can enhance burstiness for better power savings with Traditional Power Management schemes, the prefetching that is needed for such savings turns out to be extremely aggressive. At such aggressive levels of prefetching, the effects can in fact turn out to be detrimental when considering the implementation in a practical setting where there are bound to be inaccuracies. Consequently, the results of this paper reiterate the necessity of alternate techniques such as DRPM for server disk power management.

Acknowledgments: This research has been supported in part by NSF grants 0429500, 0325056, 0130143, 0103583, 0097998, and an IBM Faculty Award.

REFERENCES

- [1] B. Moore, "Taking the Data Center Power and Cooling Challenge, Energy User News, Aug 27, 2002."
- [2] F. Moore, "More Power Needed, Energy User News, Nov 25, 2002."
- [3] "Power, Heat and Sledgehammer. White Paper, Maximum Institution Inc." <http://www.max-t.com/downloads/whitepapers/SledgehammerPowerHeat20411.pdf>, 2002.
- [4] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, "DRPM: Dynamic Speed Control for Power Management in Server Class Disks," in *Proceedings of the 30th Annual International Symposium on Computer Architecture*. ACM Press, 2003, pp. 169–181.
- [5] S. Gurumurthi, J. Zhang, A. Sivasubramaniam, M. Kandemir, H. Franke, N. Vijaykrishnan, and M. J. Irwin, "Interplay of Energy and Performance for Disk Arrays Running Transaction Processing Workloads," in *Proceedings of the International Symposium on Performance Analysis of Systems and Software*, 2003.
- [6] E. Carrera, E. Pinheiro, and R. Bianchini, "Conserving Disk Energy in Network Servers," in *Proceedings of the 17th International Conference on Supercomputing*, 2003.
- [7] Q. Zhu, F. David, C. Devaraj, Z. Li, Y. Zhou, and P. Cao, "Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management," in *Proceedings of the 10th International Symposium on High Performance Computer Architecture*, 2004.
- [8] F. Douglis, P. Krishnan, and B. Bershad, "Adaptive Disk Spindown Policies for Mobile Computers," in *Proc. 2nd USENIX Symp. on Mobile and Location-Independent Computing*, 1995.

- [9] D. P. Helmbold, D. D. E. Long, T. L. Sconyers, and B. Sherrod, "Adaptive Disk SpinDown for Mobile Computers," *Mob. Netw. Appl.*, vol. 5, no. 4, pp. 285–297, 2000.
- [10] Y. Lu and G. D. Micheli, "Adaptive Hard Disk Power Management on Personal Computers," *Proceedings of the IEEE Great Lakes Symposium on VLSI*, pp. 50–53, 1999.
- [11] K. Li, R. Kumpf, P. Horton, and T. E. Anderson, "A Quantitative Analysis of Disk Drive Power Management in Portable Computers," in *Proceedings of the USENIX Winter Conference*, 1994, pp. 279–291.
- [12] T. Heath, E. Pinheiro, J. Hom, U. Kremer, and R. Bianchini, "Application Transformations for Energy and Performance-Aware Device Management," in *Proceedings of the 2002 International Conference on Parallel Architectures and Compilation Techniques*. IEEE Computer Society, 2002, pp. 121–130.
- [13] R. A. Golding, P. B. II, C. Staelin, T. Sullivan, and J. Wilkes, "Idleness is Not Sloth," in *Proceedings of the USENIX Winter Conference*, 1995, pp. 201–212.
- [14] Y.-H. Lu, E.-Y. Chung, T. Simunic, L. Benini, and G. D. Micheli, "Quantitative Comparison of Power Management Algorithms," in *Proceedings of the Design Automation and Test in Europe*. Stanford University, March 2000, pp. 20–26.
- [15] A. Papathanasiou and M. Scott, "Increasing Disk Burstiness for Energy Efficiency, Technical Report 792, Department of Computer Science, University of Rochester," 2002.
- [16] A. Weissel, B. Beutel, and F. Bellosa, "Cooperative I/O: a Novel I/O Semantics for Energy-Aware Applications," *SIGOPS Oper. Syst. Rev.*, vol. 36, no. SI, pp. 117–129, 2002.
- [17] X. Li, Z. Li, F. David, P. Zhou, Y. Zhou, S. Adve, and S. Kumar, "Performance Directed Energy Management for Main Memory and Disks," *Proceedings of the SIGARCH Comput. Archit. News*, vol. 32, no. 5, pp. 271–283, 2004.
- [18] "IBM Ultrastar 36Z15 Hard Disk Drive," <http://www.hitachigst.com/tech/techlib.nsf/techdocs>.
- [19] "Adaptive Power Management for Mobile Hard Drives - Adaptive Battery Life Extender, A Self-Managed Approach to Saving Energy, Storage Systems Division, IBM Corporation," http://www.almaden.ibm.com/almaden/mobile_hard_drives.html, 1999.
- [20] "Quietly Cool-White Paper, Hitachi Power and Acoustic Management," <http://www.hitachigst.com/tech/techlib.nsf/techdocs>, 2004.
- [21] G. R. Ganger, B. L. Worthington, and Y. L. Patt, "The DiskSim Simulation Environment - Version 2.0 Reference Manual," <http://www.pdl.cmu.edu/DiskSim/diskim2.0.html>.
- [22] G. Ganger, "System-Oriented Evaluation of I/O Subsystem Performance, PhD thesis, University of Michigan, Ann Arbor," 1995.
- [23] "IBM ServeRAID Ultra 160 SCSI Controller User's Reference Version 4.80," <http://www-307.ibm.com/pc/support/site.wss>.
- [24] "HPL Cello 99," <http://www.hpl.hp.com/research/ssp/software/>.
- [25] "HPL Openmail," <http://www.hpl.hp.com/research/ssp/software/>.
- [26] "TPC-H, Transaction Processing Performance Council," <http://www.tpc.org/tpch>.
- [27] "Umass Trace Repository, Financial," <http://signl.cs.umass.edu/repository>.
- [28] "Umass Trace Repository, Websearch," <http://signl.cs.umass.edu/repository>.
- [29] S. T. Leutenegger and D. M. Dias, "A Modeling Study of the TPC-C Benchmark," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*, P. Buneman and S. Jajodia, Eds. ACM Press, 1993, pp. 22–31.
- [30] "TPC-C, Transaction Processing Performance Council," <http://www.tpc.org/tpcc>.
- [31] "IBM DB2," <http://www-306.ibm.com/software/data/db2/>.
- [32] K. Keeton, G. Alvarez, E. Riedel, and M. Uysal, "Characterizing I/O-intensive Workload Sequentiality on Modern Disk Arrays, Hewlett-Packard Labs Storage Systems Program," in *Proceedings of the 4th Workshop on Computer Architecture Evaluation using Commercial Workloads*, 2001.