Power-efficient Server-class Performance from Arrays of Laptop Disks

Athanasios E. Papathanasiou and Michael L. Scott

The University of Rochester
Computer Science Department
Rochester, New York 14627

Technical Report 837

May 2004

Abstract

The disk array of a server-class system can account for a significant portion of the server's total power budget. Similar observations for mobile (e.g. laptop) systems have led to the development of power management policies that spin down the hard disk when it is idle, but these policies do not transfer well to server-class disks. On the other hand, state-of-the-art laptop disks have response times and bandwidths within a factor of 2.5 of their server class cousins, and consume less than one sixth the energy. These ratios suggest the possibility of replacing a server-class disk array with a larger array of mirrored laptop disks. By spinning up a subset of the disks proportional to the current workload, we can exploit the latency tolerance and parallelism of typical server workloads to achieve significant energy savings, with equal or better peak bandwidth. Potential savings range from 50% to 80% of the total disk energy consumption.

This work was supported in part by NSF grants numbers EIA-0080124, CCR-9988361, and CCR-0204344, by DARPA/AFRL contract number F29601-00-K-0182, and by Sun Microsystems Laboratories.

1 Introduction

Data centers capable of providing Internet, application, database and network services are an increasingly important component of the world's computing infrastructure. In 1995 there were 20,000 servers in the world. As of June 2001, that number had reached six million [Maximum Throughput Inc, 2002]. Most existing research on data center design has aimed to improve performance, reliability and availability. Recently, however, researchers have begun to recognize the importance of energy efficiency [Chase and Doyle, 2001; Bohrer *et al.*, 2002]. Scientists at the Lawrence Berkeley National Laboratory have estimated that the annual energy used by servers, minicomputers, and mainframes exceeded 19 terawatt-hours (TWh) in 1999 [Kawamoto *et al.*, 2000] (equivalent to the total output of four large commercial power plants). Moreover, the increase in demand for data center services suggests that by 2005 new data centers will require approximately 40TWh or $4B at $100 per MWh unless they become more energy efficient [Chase and Doyle, 2001]. Increased data center power consumption translates directly into higher total cost of ownership, attributable to operating power, cooling, and decreased reliability.

A recent white paper suggests that disk drives in a data center can account for 27% of total electric consumption [Maximum Throughput Inc, 2002]. In some configurations the fraction can be significantly higher. A Dell PowerEdge 6650 [DELL, 2003], for example, comes equipped with 4 Intel Xeon 2.0 GHz processors and 292 15 KRPM hard drives. The processors are rated at 58W each, while an operational SEAGATE ST318453 15KRPM 18GB server-class disk drive consumes 15W [Seagate, 2003]. In such a configuration the hard disks consume 15 times more power than the processors.

Previous work on the energy efficiency of storage systems has focused on the hard drives used in mobile environments. Such drives support multiple non-operational low power modes that can provide significant energy savings during relatively long periods of inactivity (on the order of tens of seconds). The key idea in this work is to place the hard drive into a non-operating low power mode during relatively long periods of inactivity [Douglis *et al.*, 1995; Douglis *et al.*, 1994; Helmbold *et al.*, 1996; Li *et al.*, 1994]. Unfortunately, such policies do not transfer in an obvious way to the server environment. Server-class disks require tens of seconds to move from an inactive to an active state, and consume tens of Watts while "spinning up". In order to save energy by spinning down, they must be idle for significantly longer than their laptop-class counterparts. At the same time, server workloads are characterized by disk access patterns that have significantly shorter request inter-arrival times than are typical in mobile workloads. Finally, server-class disks are not designed for frequent stop-start cycles; heavy use of standby mode can be expected to significantly lower their mean time to failure.

To improve the energy efficiency of server-class storage systems Gurumurthi et al. have suggested the use of DRPM [Gurumurthi *et al.*, 2003], an approach that dynamically modulates disk speed depending on current workload, decreasing the power required to keep the platters spinning when the load is light. Given the load fluctuations typical of web and transaction processing [Bohrer *et al.*, 2002], such an approach can save significant amounts of energy during periods of low activity while maintaining adequate response times during peaks. Unfortunately, DRPM disks are not yet commercially available. As an alternative, we suggest replacing each server-class disk in an array with a modest number (e.g. 3) of mirrored, energy-efficient laptop-class disks. This alternative can provide comparable or even improved throughput during workload peaks by exploiting read parallelism, while consuming significantly less energy when the load is light, by spinning down disks that represent excess capacity.

The rest of the paper is structured as follows. Section 2 describes the characteristics of laptop-class and server-class disks and makes the case for using the former in server disk arrays. Section 3 describes an array of laptop disks designed to minimize energy consumption. Section 4 shows the energy efficiency potential of the proposed design. Finally, section 5 describes related work and section 6 presents our conclusions.

Disk	DK23DA-F30	TravelStar 7K60	UltraStar 15K73	Cheetah 15K.3
Price	$127	$210	$195	$199
Capacity	$30GB$	$60GB$	$73.9GB$	$73GB$
Form Factor	$2.5inch$	$2.5inch$	$3.5inch$	$3.5inch$
Cache Size	$2048KB$	$8192KB$	$8192KB$	$8192KB$
Rotational Speed	$4200RPM$	$7200RPM$	$15037RPM$	$15000RPM$
Max. Transfer Rate	$35MB/s$	$64MB/sec$	$129MB/s$	$112MB/s$
Avg. Seek Time	$13ms$	$10ms$	$3.9ms$	$3.8ms$
Max. Seek Time	$25ms$	$16ms$	7.2	$6.7ms$
Operating Shock	$180G/2ms$	$200G/2ms$	$15G/11ms$	$60G/2ms$
Recording Density	$600KBPI$	$624KBPI$	N/A	$533KBPI$
Track Density	$55.0KTPI$	$88.2KTPI$	N/A	$64KTPI$
Avg. Active Power	$2.1W$	$2.5W$	$15.6W$	$14.9W$
Performance Idle	$1.6W$	$2W$	$12.0W$	$12.2W$
Active Idle	N/A	$1.3W$	N/A	N/A
Low Power Idle	$0.6W$	$0.85W$	N/A	N/A
Standby	$0.25W$	$0.25W$	N/A	N/A
Sleep	$0.1W$	$0.1W$	N/A	N/A
Standby to Ready	$3sec$	$3sec$	$25sec$	$20sec$

Table 1: Characteristics of four mobile and server-class hard drives. Hitachi *DK23DA-F30* [Hitachi, 2001] is a 2001 mobile hard drive, Hitachi *TravelStar 7K60* [Hitachi, 2003a] is the newest mobile hard drive implemented by the Hitachi Global Storage Technologies and the first mobile hard drive with a rotational speed of 7200 RPM. Hitachi *UltraStar 15K73* [Hitachi, 2003b] and Seagate *Cheetah 15K.3* [Seagate, 2003] represent modern server-class drives. The prices presented were the lowest prices retrieved from http://www.pricegrabber.com on April 22, 2004. The rest of the parameters have been taken from the respective hard drive specifications. The Transfer rate represents the maximum media-to-buffer transfer rate. For the two mobile drives the row "Standby to Ready" represents the time to transition from the standby low power mode (heads are parked and platters are spun down) to ready (performance idle mode), while for the server-class disks the time from power on to ready.

2 Disk Arrays: Mobile or Server-class Disks?

The constant need of data-centric services for higher throughput and shorter response times is driving server-class hard drives toward higher rotational speeds and shorter seek latencies. State-of-the-art server class disks have rotational speeds of 15 thousand RPM and seek times shorter than 4ms. As shown in Table 1, such performance characteristics come at the cost of significantly increased power budgets. Even when idle, a 15 KRPM hard drive consumes more than 12 Watts.

In comparison to their server-class counterparts, modern hard disks for mobile systems have worse performance characteristics. However, the need of the mobile system market for longer battery lifetimes makes energy efficiency a very important factor in the development of mobile hard drives. Even when active, modern mobile hard drives typically consume less than 3 Watts. They also typically support several non-operational low power modes (Table 1): Active Idle, Low Power Idle, Standby and Sleep[1]. Non-operational

[1]In the Active Idle state the head servoing mechanism and portion of the electronics are off. In the Low Power Idle state the disk is still spinning, but the electronics may be partially unpowered, and the heads may be parked or unloaded. In the Standby state the disk is spun down. The Sleep state powers off all remaining electronics; a soft reset is required to return to higher states.

low power modes can save significant energy during periods of disk inactivity.

A comparison of the hard disks parameters presented in Table 1 shows that a modern 7200 RPM mobile hard disk, such as the Hitachi TravelStar 7K60, has the same cache size as, and similar capacity to, a 15 KRPM server-class disk. The server-class disks provide up to 2.5 times better performance, but consume 6 times as much power, even when the low power modes of the mobile disk are not being used. Such differences suggest that by replacing each high performance disk in a server environment with a mirrored (RAID Level 1) array [Patterson *et al.*, 1988] of several (Table 1 suggests 3) mobile hard disks, one can achieve similar or higher I/O throughput at significantly lower power. Individual request response times will be higher, but for most large secondary storage systems aggregate I/O throughput is (within limits) more important [Chen *et al.*, 1994].

The lower power consumption of mobile hard drives has additional advantages. First, it can lead to improved reliability. Increased temperatures caused by heat dissipation in densely packed storage systems can lead to component malfunctions: for every degree that the operating temperature of a hard disk exceeds the recommended level the failure rate increases two to three percent [Maximum Throughput Inc, 2002]. Over time, a hard disk operating at just five degrees above its recommended temperature is 10 to 15 percent more likely to fail. The resistance of mobile disks to operating shocks (Table 1) also increases their reliability. Second, reduced power consumption can lead to a smaller cost for cooling equipment. Third, the smaller form factor, combined with lower heat dissipation, means that several (certainly three) mobile disks can be squeezed into the same physical space as their server-class counterpart, while generating much less acoustic noise.

We believe that the attractiveness of mobile hard drives as an alternative to server-class disks will increase over time. The notebook market is growing faster than the desktop and server markets. Rising demand, together with the trend toward higher performance mobile processors, operating systems, interfaces, and buses, is fueling the development of ever faster mobile hard drives. Recent technological advancements, such as adaptive formating [Laroia and Condon, 2003], antiferromagnetically-coupled (AFC) media [IBM, 2001], fluid dynamic motors [Blount, 2001], and fempto sliders [Best *et al.*, 2003], have led to faster, more reliable, higher capacity mobile hard drives at (almost) the same power budget. The 2004 TravelStar 7K60 hard drive, when compared to the 2001 Hitachi DK23DA hard drive (Table 1), has twice the capacity, 1.7 times the rotational speed, and improved seek times, while its idle power consumption has increased by only 25%. The same technologies that have lead to the development of the 7K60 promise additional future improvements. For example the antiferromagnetically-coupled media [Laroia and Condon, 2003] and the fempto slider [Best *et al.*, 2003] suggest future areal densities of $100 Gbit/inch^2$, compared to roughly $70 Gbit/inch^2$ for current commercial products.

The above trends make arrays of mobile disks an attractive alternative to high performance server class disks. The principal disadvantage of such an array is its price: with mobile and server-class disks of comparable capacity costing roughly the same amount (Table 1), replacing each server-class disk with several mobile disks will increase the initial investment several fold. This investment may be partially offset by reduced electric bills. In addition, the redundancy of mirroring should eliminate the need for parity disks, allowing an array of $n+1$ server-class disks to be replaced by $3n$ mobile disks, rather than $3n+3$, where n is the parity width. (Striping may still be desirable for high bandwidth sequential reads.) Finally, economies of scale suggest that mobile disk prices may improve faster than those of server-class disks.

3 Design Issues in an Array of Laptop Disks

The idea of using arrays of laptop disks in place of server-class disks has as a goal to minimize power consumption while maintaining acceptable aggregate I/O throughput. Based on the discussion in section 2,

replacing each server-class disk in a large scale storage system with a mirrored array of three 7200 RPM mobile hard disks will provide similar or even better aggregate throughput at half the power budget, even if the disks remain active constantly. Additional power savings can be achieved by taking advantage of the varying workload intensity of web server systems and transaction processing systems [Bohrer *et al.*, 2002]. During low intensity periods, only the portion of the hard disks that is necessary in order to sustain acceptable throughput need actually be active. Depending on the exact characteristics of the workload, the additional disks may enter one of several low power modes. The choice of mode depends on the intensity of the workload and more specifically on the rate at which the secondary disks accept requests.

Traditional mirrored disk array systems aim to maximize aggregate throughput without regard to power consumption. Hence, common policies used to select the disk to service a request attempt to balance the load evenly across all mirrored disks. Examples of such policies include random selection, round-robin selection, or selection of the disk with the shortest request queue. Such load balancing schemes are inappropriate for power efficiency: the disk array controller may keep all disks active even during light workloads by submitting requests to all disks even during low intensity workloads.

A more power-friendly approach would be to use a policy that starts using secondary disks only when individual response times exceed a certain threshold. Such a policy has the advantage of increasing the request inter-arrival time to secondary disks, allowing them to drop into low power modes when the load is low. At the same time, by tracking the response times of individual requests and spinning up additional disks when those times exceed some acceptable threshold, we can guarantee not to damage aggregate throughput. A simple way to limit worst case response time is to track the number of pending requests in the queue of each mirrored disk. New requests can be scheduled to one of the currently active disks until the number of requests in the queue exceeds a certain some queuing threshold, at which point one of the secondary disks can be activated in order to service additional requests. If all disks in the system have exceeded their queuing threshold, a traditional load balancing scheme will become appropriate.

Unfortunately, while reads can be spread across disks, writes must be performed on all copies. This may lead to increased response times in write intensive workloads, since the aggregate write throughput is limited to that of a single disk. Power consumption may also increase with a decrease in the length of secondary disk idle intervals, which can lead to inefficient use of low power modes. Fortunately, Internet content delivery is characterized mostly by read activity. It may also be possible to reduce the power impact of writes (though not their performance impact) by updating only those disks that are currently active. Idle disks may be synchronized periodically using data from the primary disks or from a disk array write cache. We plan to explore such options in our future work.

4 Energy Efficiency Potential of Laptop Disk Arrays

In this section we evaluate the idea of using laptop disk arrays in server-class storage systems. We conducted our experiments using the Disksim simulator [Ganger *et al.*, 1999; Ganger, 1995]. We augmented the simulator with code to model energy consumption. Part of the code is based on the Dempsey simulator [Zedlewski *et al.*, 2003]. The disk configuration parameters used in the simulations are based on the Hitachi TravelStar 7K60 and Hitachi UltraStar 15K73 disks (Table 1). In the simulations of the TravelStar 7K60 disk we assume a simplified disk power model with the characteristics shown in Table 2[2].

[2]We do not model the active idle power mode, since we do not have information about its transition time parameters. Our results are hence conservative, since the use of the active low power mode can provide additional savings even during short idle intervals. In addition, we do not model the Sleep mode.

State	Mobile Disk
Active	2.50 W
Performance Idle	2.00 W
Low Power Idle	0.85 W
Standby Power	0.25 W
Low Power Idle to Active Energy	1.45 J
Low Power Idle to Active Time	0.53 s
Active to Low Power Idle Energy	2.15 J
Active to Low Power Idle Time	0.85 s
Standby to Active Energy	9.00 J
Standby to Active Time	3.00 s
Active to Standby Energy	2.94 J
Active to Standby Time	1.25 s

Table 2: Abstract disk model parameters used calculate energy consumption. Values are based on the TravelStar 7K60 and UltraStar 15K73 disk.

Workload Type	Bandwidth (Rqs/sec)	Fraction of Max. *Srv* Bandwidth
<*Exp-8*>	124.73	78.75%
<*Exp-10*>	99.79	63.00%
<*Exp-1000*>	1.00	0.63%
<*Par-10*>	99.75	62.98%
<*Par-50*>	19.95	12.60%
Max. *Srv* Bandwidth	158.39	100.00%
Max. *Mbl* Bandwidth	211.66	133.63%

Table 3: Aggregate bandwidth in requests per second imposed by each workload tested and its fraction of the maximum bandwidth supported by the server-class disk (*Srv*). The maximum aggregate bandwidth of the laptop-disk array (*Mbl*) is also shown.

In our evaluation, we compare the performance and energy consumption of a server-class disk with that of mirrored disk array system consisting of three mobile hard drives. Since we want to evaluate the proposed idea under workloads with various degrees of intensity we have conducted an open system simulation using synthetic workloads. In contrast to a closed system simulation, in an open system simulation requests are considered to be independent of each other. In all the experiments, one million requests are issued. We use two types of distributions in order to model the inter-arrival time among requests generated by the synthetic workloads: an exponential distribution that leads to a "smoother" workload and a Pareto distribution that leads to a more "bursty" workload. We experiment with a wide range of mean inter-arrival times. In the remainder of the paper, we will follow the naming convention <*Distribution-Mean*> in order to refer to a specific workload. For example <*Exp-10*> represents a workload with request inter-arrival times that follow an exponential distribution with a mean of 10 ms, while <*Par-50*> represents a workload with request inter-arrival times that follow a Pareto distribution with a mean of 50 ms. Table 3 presents the aggregate bandwidth in requests per second imposed by each workload to the storage system. Similar workload configuration parameters were used in the evaluation of the DRPM approach [Gurumurthi *et al.*, 2003].

Across all experiments we present results for five systems. *Srv* represents a storage system consisting of a single server-class disk. The remaining four systems are variations of the laptop disk based array. We

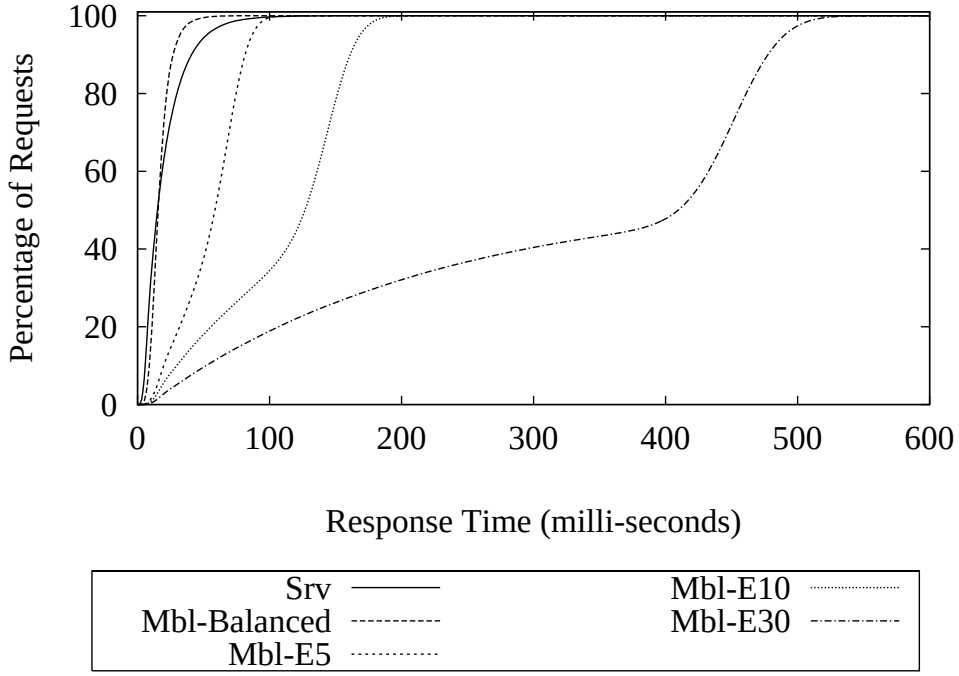

Figure 1: Request Response Time Distribution for a read-only <*Exp-8*> workload. Higher queuing delays in the *Srv* system lead to slightly worse response times than the *Mbl-Balanced* system.

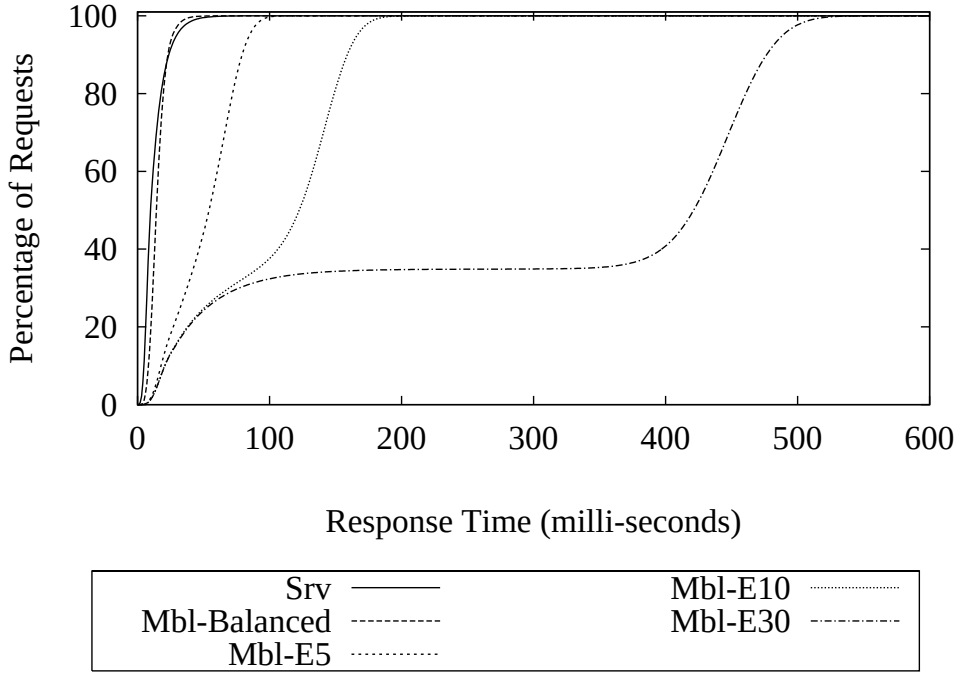

Figure 2: Request Response Time Distribution for a read-only <*Exp-10*> workload.

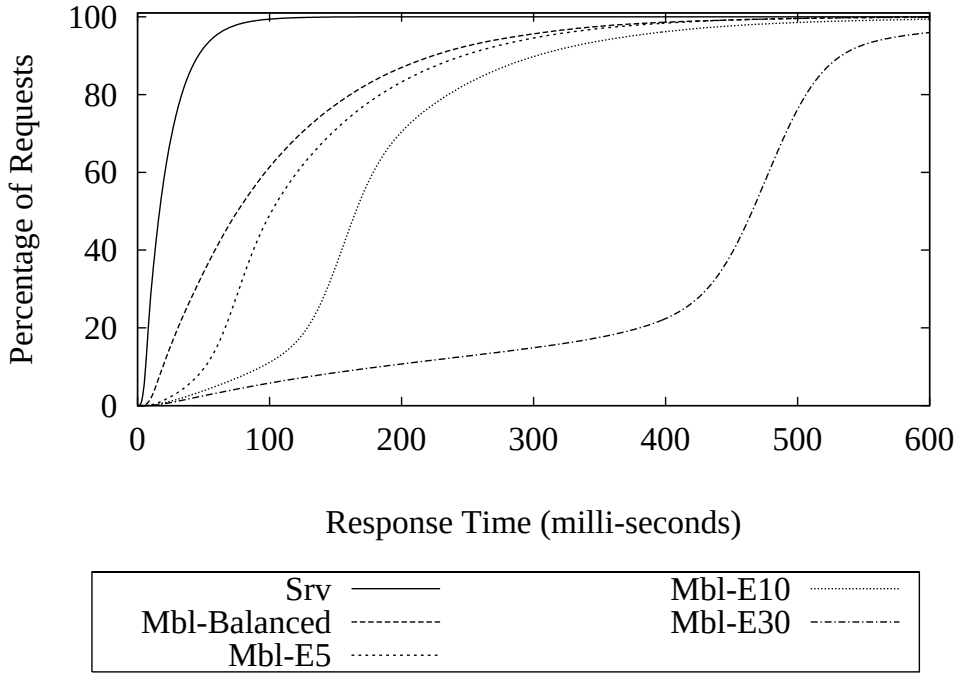

Figure 3: Request Response Time Distribution for <*Exp-8*> workload. 20% of the requests are writes.

experiment with four different policies for selecting which disk among the replicas in the disk array should service a given request. The first disk selection policy, *Mbl-Balanced*, represents a traditional load balancing scheme that selects the disk in the array with the shortest queue. The remaining three policies represent the more power-conscious disk selection policy discussed in Section 3 with various queuing thresholds. The queuing threshold is 5 in the *Mbl-E5* system, 10 in the *Mbl-E10* system and 30 in the *Mbl-E30* system. The disk selection policy represents a tradeoff between energy efficiency and worst case response time for individual requests. Depending on the worst case response time that a server system can tolerate, we can use a traditional load balancing scheme in the array in order to achieve better performance or a power-conscious disk selection policy with a higher queuing threshold to achieve a more power efficient system.

Our first set of experiments evaluates the performance of the traditional and laptop-disk systems under two intensive read-only workloads, <*Exp-8*> and <*Exp-10*>. Figures 1 and 2 present the results. Since the laptop disk array is able to support a higher maximum aggregate throughput than the server-class disk (212 versus 159 requests per second), the *Mbl-Balanced* system slightly outperforms the server-class disk during the most intensive <*Exp-8*> workload (Figure 1). It leads to an average response time of 17.08 requests per second versus 20.01. The power-conscious arrays exhibit worse response times. Larger queuing thresholds lead to larger queuing delays for a significant portion of the requests and hence a larger average response time. Across all experiments, however, response time remains below half a second, a reasonable worst-case value for a web server or similar system. Results for the <*Exp-10*> distribution are similar. The main difference is that the server-class disk performs slightly better than the balanced disk array. Since inter-arrival times are relatively longer the *Srv* system experiences shorter queuing delays, which lead to shorter response times.

In Figures 3-7 we present cumulative response time distributions for workloads in which 20% of the requests are writes. <*Exp-8*>, <*Exp-10*> and <*Par-10*> are intensive workloads, while <*Exp-1000*> and <*Par-50*> are relatively light workloads. As expected, the performance of the laptop disk array degrades in

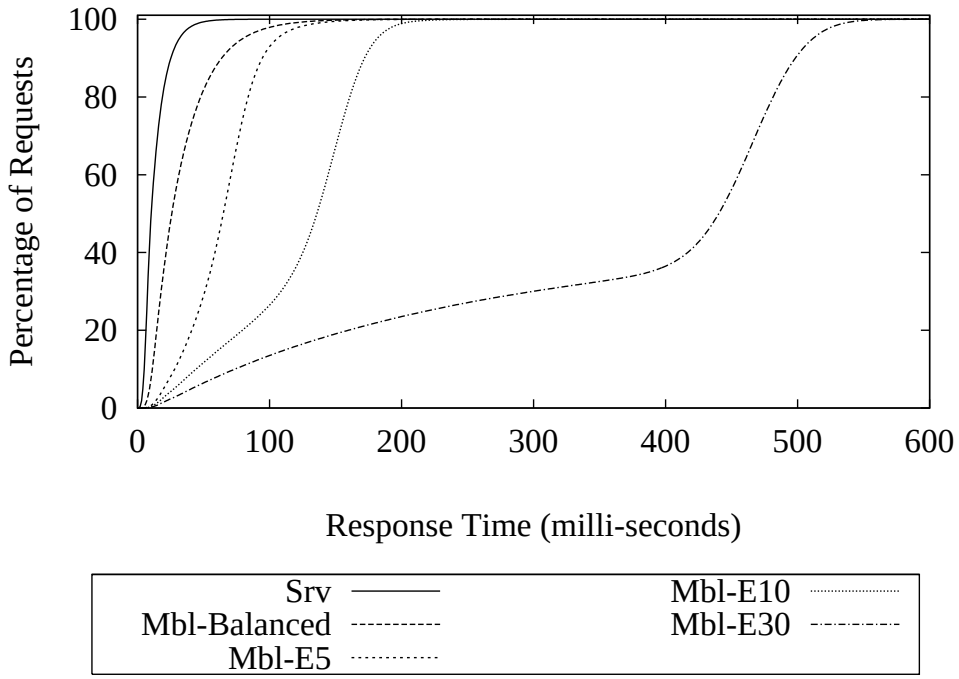

Figure 4: Request Response Time Distribution for <*Exp-10*> workload. 20% of the requests are writes.

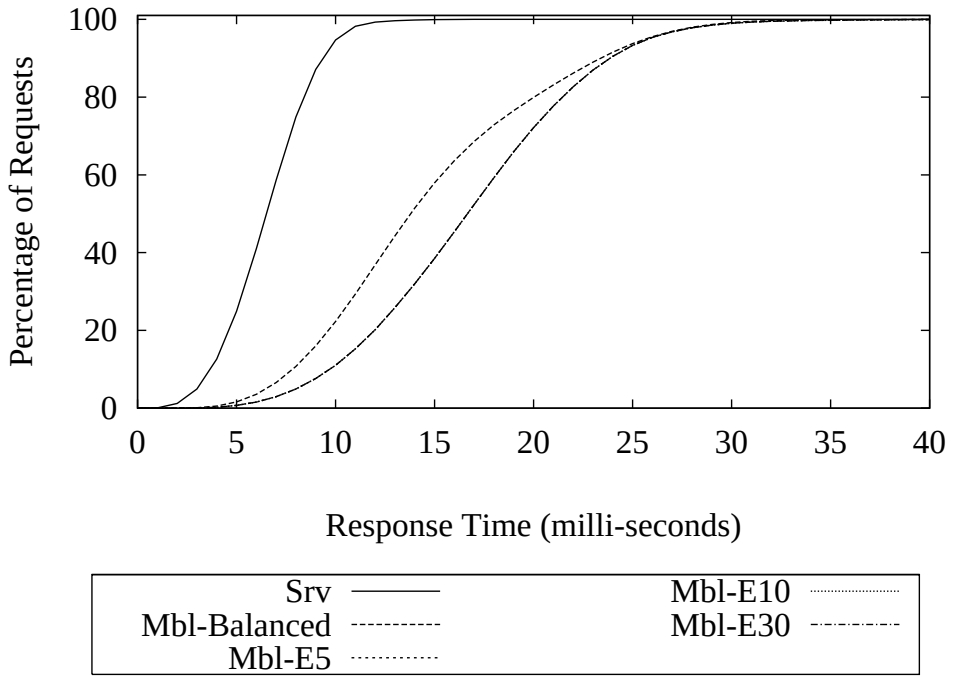

Figure 5: Request Response Time Distribution for <*Exp-1000*> workload. 20% of the requests are writes.

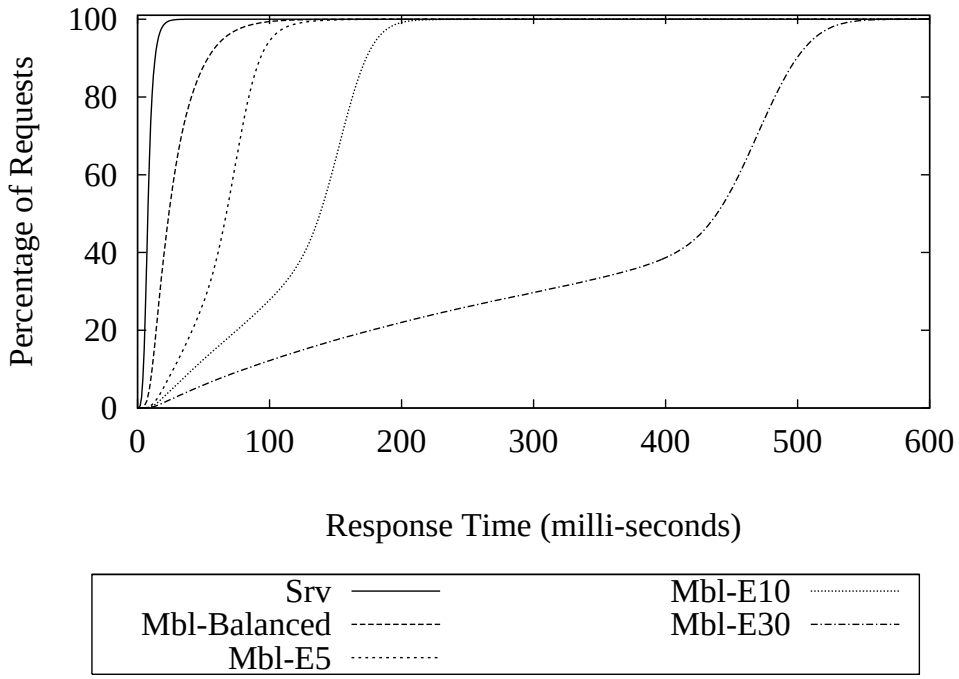

Figure 6: Request Response Time Distribution for <*Par-10*> workload. 20% of the requests are writes.

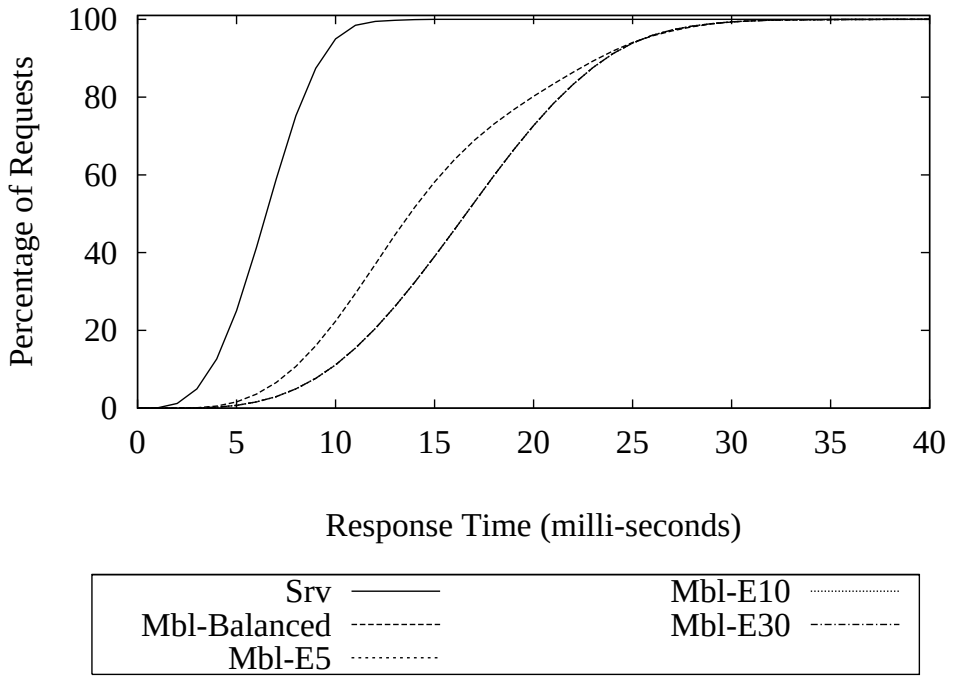

Figure 7: Request Response Time Distribution for <*Par-50*> workload. 20% of the requests are writes.

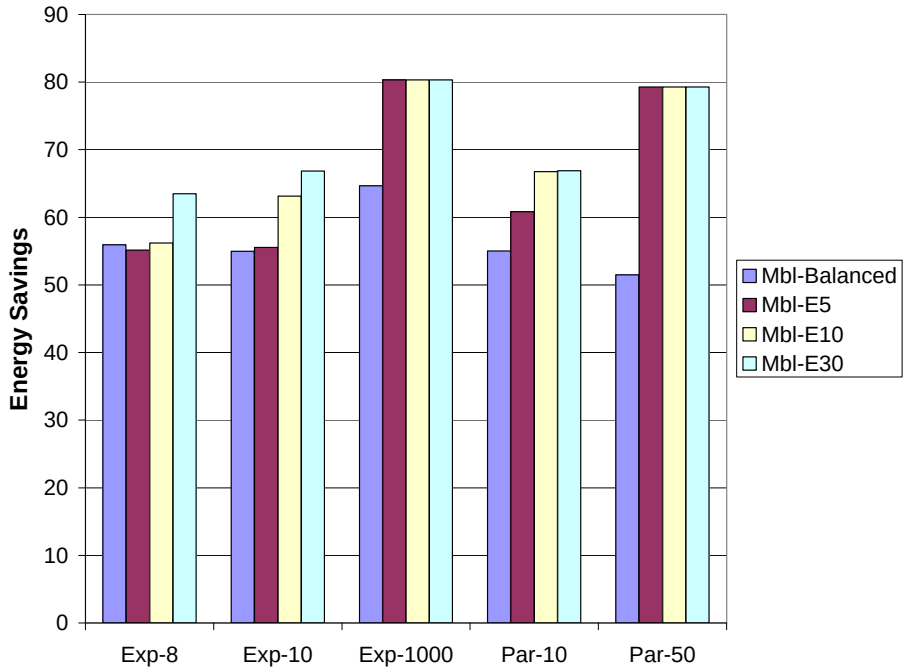

Figure 8: Energy Savings of the array of laptop disks during read-only workloads. The base case for the comparison is the energy consumption of the server-class disk for the respective workloads.

the presence of write activity since writes have to be performed on all replicas. The performance degradation is more pronounced in the intensive workloads (Figures 3, 4 and 6). In the light workloads (Figures 5 and 7) all systems exhibit very low response times: (below 17 ms).

Figures 8 and 9 present the energy savings that can be achieved by the array of laptop disks. The base case for the comparison is the energy consumption of a server-class disk (*Srv*) system. We present results for all workloads and disk selection policies. We assume an optimal power management policy that always chooses the most efficient mode and also preactivates the disks. Our goal is to estimate the energy efficiency potential of the proposed design. The laptop-disk array consumes at least 50% less energy than the server-class disk. These savings come from the fact that each laptop disk consumes 1/6*th* of the energy consumed by the server-class disk, even when low power modes are not being used. In the absence of write requests, (Figure 8) energy savings can reach up to 80%, when the power-conscious disk selection policies are being used. A higher queuing threshold leads to higher savings. Figure 9 presents results for workloads with 20% write requests. Write activity reduces significantly the energy savings achieved by the laptop-disk array across all workloads other than the lightest one, $<$*Exp-1000*$>$. Write requests have to be issued to all replicas, and hence idle interval lengths are reduced for all disks in the array. The reduced idle interval lengths lead to reduced energy savings since low power modes cannot be used efficiently. In our future work we plan to explore methods for postponing write requests to idle replicas.

5 Related Work

Power Management for Mobile Systems. The research community has been very active in the area of power-conscious systems during the last few years. Ellis et al. [Ellis, 1999] emphasized the importance of energy efficiency as a primary metric in the design of operating systems. ECOSystem [Zeng *et al.*, 2002] provides a model for accounting and for fairly allocating the available energy among competing

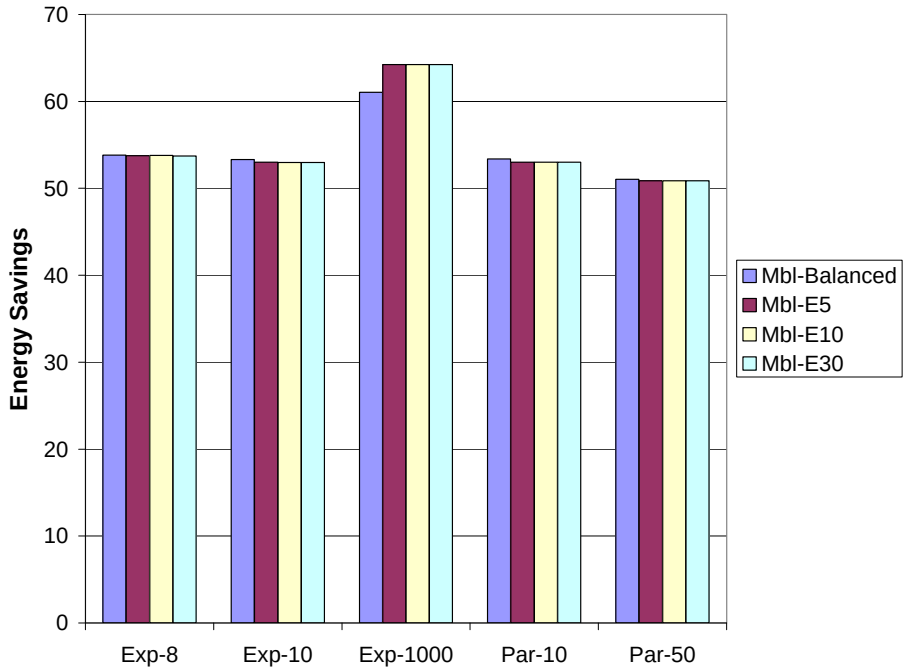

Figure 9: Energy Savings of the array of laptop disks during workloads with 20% write requests. The base case for the comparison is the energy consumption of the server-class disk for the respective workloads.

applications according to user preferences. Odyssey [Flinn and Satyanarayanan, 1999; Noble *et al.*, 1997] provides operating system support for application-aware resource management. Several policies have been proposed for decreasing the power consumption of processors that support dynamic voltage and frequency scaling. The key idea is to schedule so as to "squeeze out the idle time" in rate-based applications. Several researchers have proposed voltage schedulers for general purpose systems [Flautner and Mudge, 2002; Govil *et al.*, 1995; Weiser *et al.*, 1994; Pering *et al.*, 2000].

Power Management for Server Systems. The importance of power management in server systems has been noted by several researchers [Bohrer *et al.*, 2001; Bohrer *et al.*, 2002; Chase and Doyle, 2001]. Bohrer et al. [Bohrer *et al.*, 2002] explore the potential of dynamic voltage scaling for power savings in web servers, while Elnozahy et al. [Elnozahy *et al.*, 2003] suggest the combination of dynamic voltage scaling with "request batching", a technique that groups requests and executes them in batches, to reduce the processor power consumption of web servers. Pinheiro et al. [Pinheiro *et al.*, 2001] propose turning machines on and off in server clusters in order to reduce energy consumption. Chase et al. [Chase *et al.*, 2001] explore a similar idea and introduce an economic approach to resource management by allowing web servers to "bid" for resources.

Power Management for Storage Systems. The energy efficiency of hard disks is not a new topic. The cost and risks of standby mode played a factor in the early investigation of hard-disk spin-down policies [Douglis *et al.*, 1995; Douglis *et al.*, 1994; Helmbold *et al.*, 1996; Li *et al.*, 1994] for mobile systems. Unfortunately such spin-down policies are not effective in server-class storage systems. Server-class disks are not designed for rapid spin-down and spin-up, and server workloads typically have access patterns with very short idle intervals, which would limit the utility of low-power modes even if they were available.

Recently, several groups have begun to investigate the deliberate generation of *bursty* access patterns that provide more opportunities the a hard disk to transition to a low power mode. Heath et al. [Heath *et al.*, 2002] and Weissel et al. [Weissel *et al.*, 2002] have proposed user-level mechanisms to increase the

burstiness of I/O requests from individual applications. Papathanasiou et al. [Papathanasiou and Scott, 2003; Papathanasiou and Scott, 2004] explore operating system prefetching and caching techniques to "shape" the disk access pattern of the system as a whole. We believe that such techniques, though initially intended for mobile systems, could improve the efficiency of servers as well.

Colarelli et al. [Colarelli and Grunwald, 2002] explore massive arrays of idle disks, or MAID, as an alternative to conventional mass storage systems for scientific computing; they demonstrate performance comparable to that of a conventional RAID system at significantly lower power. Gurumurthi et al. [Gurumurthi *et al.*, 2003] suggest the use of DRPM [Gurumurthi *et al.*, 2003], an approach that dynamically modulates disk speed depending on current workload, decreasing the power required to keep the platters spinning when the load is light. Carrera et al. [Carrera *et al.*, 2003] compare various design schemes for conserving disk energy in network servers and conclude that the DRPM approach leads to a reduced power consumption when compared to other approaches. Since multi-speed hard disk drives are not yet commercially available, we suggest the use of a modest number of mirrored, power-efficient laptop-class disks as a practical alternative. Finally, Zhu et al. [Zhu *et al.*, 2004] propose several power-aware storage cache management algorithms that provide additional opportunities for a disk power management system to save energy. Such algorithms might work well in conjunction with mirrored mobile disks, leading to higher energy efficiency than could be achieved with either approach alone.

6 Conclusions

Given the increasing importance of power management in server farms and the rapid pace of technological improvement in small form factor hard disks, we have raised the possibility of replacing an array of server-class disks with a larger array of "mobile" (2.5") disks. Current technological design points suggest a ratio of three to one.

The principal disadvantage of "mobile" disk array is its initial cost. A secondary disadvantage is higher latency for individual requests when the load is light. Potential advantages include significantly lower operational power, lower cooling needs, potentially denser packaging, lower noise, potentially higher peak bandwidth, potentially higher mean time to data loss (due to mirroring), and the opportunity to ride the faster development curve for commodity laptop disks. Back-of-the-envelope calculations (confirmed by simulations) suggest a baseline power savings of 50% when all mobile disks are active. Simulations confirm that significant additional savings, up to 80%, can be achieved when the load is light by exploiting the non-operational low-power modes supported by mobile disks. Under high load, simulations also indicate a modest *improvement* in average latency, as requests that would queue for a single server-class disk are serviced by separate mobile disks.

Many questions remain for future work. The relative costs of replacing failed drives in conventional and mobile-array based storage systems will depend not only on the MTTF of various devices, but on the extent to which MTTF itself depends on thermal management. More definitive evaluation will require not only the use of actual devices (rather than simulations), but also the development of realistic on-line power management policies (the optimal policy assumed in our simulations is not feasible in practice).

As mentioned in Section 3, we plan to consider policies in which writes are flushed only to drives that are currently spinning. Such policies obviously have an impact not only on energy and performance, but on failure resilience as well. The mirroring inherent in our proposal effectively introduces an extra dimension in the RAID design space; we will want to consider the interaction between our mirroring and both routine and recovery-mode file striping. We also plan to consider the staging of writes in non-volatile solid-state cache.

References

[Best *et al.*, 2003] John Best, Sandy Bolasna, Lee Dorius, John Kotla, Yasuhiro Iihara, Tsuyoshi Matsumoto, Randy Simmons, Atsushi Tobari, and Hiroyasu Tsuchida, "White Paper: The Femto Slider in Travelstar 7K60," May 2003.

[Blount, 2001] Walker C. Blount, "White Paper: Fluid Dynamic Bearing Spindle Motors," February 2001.

[Bohrer *et al.*, 2001] P. Bohrer, D. Cohn, E.N. Elnozahy, T. Keller, M. Kistler, C. Lefurgy, R. Rajamony, F. Rawson, and E. V. Hensbergen, "Energy Conservation for Servers," In *Proc. of the IEEE Workshop on Power Management for Real-Time and Embedded Systems at the IEEE Real-Time Technology and Applications Symposium*, pages 1–4, May 2001.

[Bohrer *et al.*, 2002] Pat Bohrer, Elmootazbellah N. Elnozahy, Tom Keller, Michael Kistler, Charles Lefurgy, Chandler McDowell, and Ram Rajamony, "The Case for Power Management in Web Servers," In *Power Aware Computing*, pages 261–289. Kluwer Academic Publishers, 2002.

[Carrera *et al.*, 2003] Enrique V. Carrera, Eduardo Pinheiro, and Ricardo Bianchini, "Conserving Disk Energy in Network Servers," In *Proc. of the 17th Annual ACM International Conference on Supercomputing (ICS'03)*, pages 86–97, June 2003.

[Chase *et al.*, 2001] Jeffrey S. Chase, Darrell C. Anderson, Prachi N. Thakar, Amin M. Vahdat, and Ronald P. Doyle, "Managing Energy and Server Resources in Hosting Centers," In *Proc. of the 18th ACM Symposium on Operating Systems Principles*, pages 103–116. ACM Press, October 2001.

[Chase and Doyle, 2001] Jeffrey S. Chase and Ronald P. Doyle, "Balance of Power: Energy Management for Server Clusters," In *Proc. of the 8th Workshop on Hot Topics in Operating Systems (HotOS VIII)*, May 2001.

[Chen *et al.*, 1994] Peter M. Chen, Edward K. Lee, Garth A. Gibson, Randy H. Katz, and David A. Patterson, "RAID: High-Performance, Reliable Secondary Storage," *ACM Computing Surveys (CSUR)*, 26(2):145–185, 1994.

[Colarelli and Grunwald, 2002] Dennis Colarelli and Dirk Grunwald, "Massive Arrays of Idle Disks for Storage Archives," In *Proc. of the 2002 ACM/IEEE Conference on Supercomputing (SC'02)*, pages 1–11, November 2002.

[DELL, 2003] "Dell PowerEdge 6650 Executive Summary," January 2003, Available: http://www.tpc.org/results/individual_results/Dell/dell_6650_010603_es.pdf.

[Douglis *et al.*, 1995] Fred Douglis, Pillaipakkamnatt Krishnan, and Brian Bershad, "Adaptive Disk Spindown Policies for Mobile Computers," In *Proc. of the 2nd USENIX Symposium on Mobile and Location-Independent Computing*, April 1995.

[Douglis *et al.*, 1994] Fred Douglis, Pillaipakkamnatt Krishnan, and Brian Marsh, "Thwarting the Power-Hungry Disk," In *Proc. of the 1994 Winter USENIX Conference*, pages 293–306, January 1994.

[Ellis, 1999] Carla S. Ellis, "The Case for Higher Level Power Management," In *Proc. of the 7th Workshop on Hot Topics in Operating Systems (HotOS VII)*, March 1999.

[Elnozahy *et al.*, 2003] Elmootazbellah N. Elnozahy, Michael Kistler, and Ramakrishnan Rajamony, "Energy Conservation Policies for Web Servers," In *Proc. of the 4th USENIX Symposium on Internet Technologies and Systems (USITS'03)*, March 2003.

[Flautner and Mudge, 2002] Krisztian Flautner and Trevor Mudge, "Vertigo: Automatic Performance-Setting for Linux," In *Proc. of the 5th USENIX Symposium on Operating Systems Design and Implementation (OSDI'02)*, pages 105–116, December 2002.

[Flinn and Satyanarayanan, 1999] Jason Flinn and Mahadev Satyanarayanan, "Energy-Aware Adaptation for Mobile Applications," In *Proc. of the 17th ACM Symposium on Operating Systems Principles*, pages 48–63, December 1999.

[Ganger, 1995] Gregory R. Ganger, *System-Oriented Evaluation of I/O Subsystem Performance*, PhD thesis, Department of Computer Science and Engineering, University of Michigan, June 1995, Published as Technical Report CSE-TR-243-95, Department of EECS, University of Michigan, Ann Arbor, June 1995.

[Ganger *et al.*, 1999] Gregrory R. Ganger, Bruce L. Worthington, and Yale N. Patt, "The DiskSim Simulation Environment Version 2.0 Reference Manual," December 1999.

[Govil *et al.*, 1995] Kinshuk Govil, Edwin Chan, and Hal Wasserman, "Comparing Algorithms for Dynamic Speed-Setting of a Low-Power CPU," In *Proc. of the 1st Annual International Conference on Mobile Computing and Networking (MobiCom'95)*, November 1995.

[Gurumurthi *et al.*, 2003] Sudhanva Gurumurthi, Anand Sivasubramaniam, Mahmut Kandemir, and Hubertus Franke, "DRPM: Dynamic Speed Control for Power Management in Server Class Disks," In *Proc. of the 30th International Symposium on Computer Architecture (ISCA'03)*, pages 169–181. ACM Press, June 2003.

[Heath *et al.*, 2002] Taliver Heath, Eduardo Pinheiro, Jerry Hom, Ulrich Kremer, and Ricardo Bianchini, "Application Transformations for Energy and Performance-Aware Device Management," In *Proc. of the 11th International Conference on Parallel Architectures and Compilation Techniques (PACT'02)*, September 2002.

[Helmbold *et al.*, 1996] David P. Helmbold, Darrell D. E. Long, and Bruce Sherrod, "A Dynamic Disk Spin-down Technique for Mobile Computing," In *Proc. of the 2nd Annual International Conference on Mobile Computing and Networking (MobiCom'96)*, November 1996.

[Hitachi, 2001] "DK23DA-40F/30F/20F/10F Disk Drive Specifications REV.3," August 2001.

[Hitachi, 2003a] "Hard Disk Drive Specification; Hitachi Travelstar 7K60 2.5 inch ATA/IDE hard disk drive; Model: HTS726060M9AT00," September 2003.

[Hitachi, 2003b] "Hard Disk Drive Specification; Hitachi Ultrastar 15K73-73/36," June 2003.

[IBM, 2001] "IBM's 'Pixie Dust' Breakthrough to Quadruple Disk Drive Density," May 2001, Available at: http://www.research.ibm.com/resources/news/20010518_pixie_dust.shtml.

[Kawamoto *et al.*, 2000] Kaoru Kawamoto, Jonathan G. Koomey, Bruce Nordman, Richard E. Brown, Mary Ann Piette, and Alan K. Meier, "Electricity Used by Office Equipment and Network Equipment in the U.S.," In *Proc. of the 2000 ACEEE Summer Study on Energy Efficiency in Buildings*, August 2000.

[Laroia and Condon, 2003] Rocky Laroia and Rich Condon, "White Paper: Adaptive Formatting in Hitachi Drives," September 2003.

[Li *et al.*, 1994] Kester Li, Roger Kumpf, Paul Horton, and Thomas Anderson, "Quantitative Analysis of Disk Drive Power Management in Portable Computers," In *Proc. of the 1994 Winter USENIX Conference*, pages 279–291, January 1994.

[Maximum Throughput Inc, 2002] "Power, Heat and Sledgehammer," April 2002, Available: http://www.max-t.com/downloads/whitepapers/SledgehammerPowerHead20411.pdf.

[Noble *et al.*, 1997] Brian Noble, Mahadev Satyanarayanan, Dushyanth Narayanan, James Eric Tilton, Jason Flinn, and Kevin R. Walker, "Agile Application-Aware Adaptation for Mobility," In *Proc. of the 16th ACM Symposium on Operating Systems Principles*, October 1997.

[Papathanasiou and Scott, 2003] Athanasios E. Papathanasiou and Michael L. Scott, "Energy Efficiency Through Burstiness," In *Proc. of the 5th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA'03)*, pages 44–53, October 2003.

[Papathanasiou and Scott, 2004] Athanasios E. Papathanasiou and Michael L. Scott, "Energy Efficient Prefetching and Caching," In *Proc. of the USENIX 2004 Annual Technical Conference*, June 2004.

[Patterson *et al.*, 1988] David A. Patterson, Garth Gibson, and Randy H. Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)," In *Proc. of the 1988 ACM SIGMOD International Conference on Management of Data (SIGMOD'88)*, pages 109–116. ACM Press, 1988.

[Pering *et al.*, 2000] Trevor Pering, Tom Burd, and Robert Brodersen, "Voltage Scheduling in the lpARM Microprocessor System," In *Proc. of the 2000 International Symposium on Low Power Electronics and Design (ISLPED'00)*, pages 96–101, July 2000.

[Pinheiro *et al.*, 2001] Eduardo Pinheiro, Ricardo Bianchini, Enrique V. Carrera, and Taliver Heath, "Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems," In *Workshop on Compilers and Operating Systems for Low Power. In conjuction with PACT'01.*, September 2001.

[Seagate, 2003] "Seagate Cheetah 15K.3 SCSI Disc Drive: ST373453LW/LC, ST336753LW/LC, ST318453LW/LC. Product Manual, Volume 1," March 2003.

[Weiser *et al.*, 1994] Mark Weiser, Brent Welch, Alan Demers, and Scott Shenker, "Scheduling for Reduced CPU Energy," In *Proc. of the 1st USENIX Symposium on Operating Systems Design and Implementation (OSDI'94)*, November 1994.

[Weissel *et al.*, 2002] Andreas Weissel, Bjorn Beutel, and Frank Bellosa, "Cooperative I/O: A Novel I/O Semantics for Energy-Aware Applications," In *Proc. of the 5th USENIX Symposium on Operating Systems Design and Implementation (OSDI'02)*, December 2002.

[Zedlewski *et al.*, 2003] John Zedlewski, Sumeet Sobti, Nitin Garg, Fengzhou Zheng, Arvind Krishnamurthy, and Randolph Wang, "Modeling Hard-Disk Power Consumption," In *Proc. of the 2nd USENIX Conference on File and Storage Technologies (FAST'03)*, pages 217–230, March 2003.

[Zeng *et al.*, 2002] Heng Zeng, Xiaobo Fan, Carla S. Ellis, Alvin R. Lebeck, and Amin Vahdat, "ECOSystem: Managing Energy as a First Class Operating System Resource," In *Proc. of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'02)*, October 2002.

[Zhu *et al.*, 2004] Qingbo Zhu, Francis David, Yuanyuan Zhou, Cristo Devaraj, and Pei Cao, "Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management," In *Proc. of the 10th International Symposium on High Performance Computer Architecture (HPCA-10)*, February 2004.