

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

Athanasios E. Papathanasiou and Michael L. Scott

University of Rochester

{papathan,scott}@cs.rochester.edu

<http://www.cs.rochester.edu/~papathan,~scott>

June 2004

Background

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 [5]. 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 [1, 2]. Increased data center power consumption translates directly into higher total cost of ownership, attributable to operating power, cooling, and decreased reliability.

The disk array of a server-class system can account for a significant portion of the server's total power budget. A recent white paper suggests that disk drives in a data center can account for 27% of total electric consumption [5]. In some configurations the fraction can be significantly higher. A Dell PowerEdge 6650 [3], 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. In such a configuration the hard disks consume 15 times more power than the processors.

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. Server-class disks are characterized by higher power mode transition costs, both in terms of power and latency, while server workloads are significantly more data intensive, leading to very short idle periods that cannot be exploited efficiently by non-operational low power modes.

Adaptive Throughput

To improve the energy efficiency of server-class storage systems Gurumurthi et al. have suggested the use of DRPM [4], 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 [1], 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 of mirrored, energy-efficient laptop-class disks. 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. By keeping activated 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.

Current technological design points suggest replacing server-class disks with laptop-class disks at a ratio of one to three [7]. The principal disadvantage of a "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.

Research Challenges

The use of laptop disk arrays in large scale storage systems raises several research challenges and a large design space that we explore in our work:

Disk selection policy. 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. A more power-friendly approach, which we explore in our work, 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 a certain minimum quality of service.

Handling of write activity. Unfortunately, while reads can be spread across disks in a mirrored disk array, 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.

Power management policy. Disk power mode transitions require significant amounts of time, on the order of a few seconds for the lowest power modes. Such delays are usually not acceptable for server-class storage systems. Allowing several disks to enter a low power mode at the same time has the risk of increasing request response times significantly during sudden workload increases. A power management policy for a laptop disk array should take into account the penalties of disk reactivation and attempt to minimize or hide their effect on request response time. Possible solutions to the problem include conservative strategies that keep more disks activated than required to sustain the current workload intensity or predictive strategies that preactivate powered-down disks in advance of an anticipated increase in workload intensity.

Two-dimensional disk arrays. 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. As an example, striped RAID systems have performance advantages over mirrored systems for workloads with large request sizes. Moreover, parity stripes provide enough reliability to make unnecessary the immediate execution of write operations on all

copies of the mirrored disk array and, hence, can lead to reduced power consumption during write bursts (see *Handling of write activity*).

Current Status

We have completed construction of a detailed multi-disk laptop array simulator. We also have available, from previous work [6], a lab-bench system capable of high precision measurements of power consumption in individual disks, which we may extend to a disk array.

Back-of-the-envelope calculations (confirmed by simulations) indicate that our three-for-one mirrored array proposal can achieve a baseline power savings of 50% when all mobile disks are active. Simulations also 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 [7].

References

- [1] BOHRER, P., ELNOZAHY, E. N., KELLER, T., KISTLER, M., LEFURGY, C., MCDOWELL, C., AND RAJAMONY, R. The Case for Power Management in Web Servers. In *Power Aware Computing*. Kluwer Academic Publishers, 2002, pp. 261–289.
- [2] CHASE, J. S., AND DOYLE, R. P. Balance of Power: Energy Management for Server Clusters. In *Proc. of the 8th Workshop on Hot Topics in Operating Systems (HotOS VIII)* (May 2001).
- [3] Dell PowerEdge 6650 Executive Summary, Jan. 2003. Available at: http://www.tpc.org/results/individual_results/Dell/dell_6650_010603_es.pdf.
- [4] GURUMURTHI, S., SIVASUBRAMANIAM, A., KANDEMIR, M., AND FRANKE, H. DRPM: Dynamic Speed Control for Power Management in Server Class Disks. In *Proc. of the 30th International Symposium on Computer Architecture (ISCA'03)* (June 2003), ACM Press, pp. 169–181.
- [5] Power, Heat and Sledgehammer, Apr. 2002. Available at: <http://www.max-t.com/downloads/whitepapers/SledgehammerPowerHead20411.pdf>.
- [6] PAPATHANASIOU, A. E., AND SCOTT, M. L. Energy Efficient Prefetching and Caching. In *Proc. of the USENIX 2004 Annual Technical Conference* (June 2004).
- [7] PAPATHANASIOU, A. E., AND SCOTT, M. L. Power-efficient Server-class Performance from Arrays of Laptop Disks. Tech. Rep. 837, Computer Science Department, University of Rochester, Apr. 2004.