H-0283 (H1005-008) May 15, 2010
Computer Science

IBM Research Report

Disk Storage Power Management

Doron Chen, George Goldberg, Roger Kahn, Ronen I. Kat, Kalman Meth, Dimitry Sotnikov
IBM Research Division
Haifa Research Laboratory
Mt. Carmel 31905
Haifa, Israel

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Disk Storage Power Management

Doron Chen, George Goldberg,,
Roger Kahn, Ronen I. Kat, Kalman Meth, and Dmitry Sotnikov
{cdoron,georgeg,rogerk,ronenkat,meth,dimitrys}@il.ibm.com
IBM Research - Haifa, Israel

***Abstract*—Data center power management has become increasingly important in recent years. In particular, the need to understand and manage storage power consumption has arisen. We developed a framework for estimating the power consumed by the storage components of a data center under varying workloads. Such a framework is useful for capacity planning tools, for enabling estimation of future performance and power consumption, and for online storage systems providing power estimation per disk, per array, and per volume. In addition, we present a technique for controlling the power consumed by disk drives that support *acoustic modes*. This technique reduces instantaneous power consumption but sacrifices performance.**

1 Introduction

Data center power considerations play an increasingly important role in the operation of data centers. This trend only increases with the growing demand for storage [22]. Since storage accounts for 13-20% of the cost of powering and cooling a data center [11], [23], [25] understanding and controlling storage power consumption is becoming increasingly important.

The power consumption of disk drives consists of two parts. The fixed portion, or *static power*, is the power consumed when the disk is in the idle state. The static power is the result of the disk spindle motor that spins the platters, and the onboard disk electronics. The variable portion, or *dynamic power*, is the power that is affected by the I/O workload. The factors which contribute to the dynamic power are the data transfer to/from the disk and the power required to move the disk head during a seek. The total power consumed by a disk is the sum of the dynamic and static power. The dynamic power of the disk can be as much as a third of the total disk power consumption. The power consumption of the disk can be further divided into mechanical power (using a $12V$ power source) for the disk spindle and seek head, and electronics power (using a $5V$ power source) for the disk electronics and data transfer operations.

Detailed understanding of storage power consumption is critical to data center management. Proper management of power consumption, in accordance with realistic workloads, can prevent over-provisioning for power and cooling.

Our contribution. We present two innovations: one for power modeling and estimation of storage, and the other for controlling and budgeting the power consumption of disks. Our modeling and estimation framework provides workload-aware power estimations for disks, disk arrays and storage controllers. It translates system-level or RAID-level operations to disk-level activities, such as disk seek and data transfers. Once the disk-level activities are determined, the workload-dependent dynamic power can be estimated.

We use *Acoustic Management*, the ability to reduce the acoustic noise of a disk drive when performing a seek operation, for controlling and budgeting disk power and energy consumption. While acoustic modes were designed to reduce the noise of the disk during a seek operation, they also reduce the instantaneous power consumption and often the energy consumption of the disk during I/O operations. Disks which support acoustic modes are in accordance with the ATA/ATAPI-6 specification [3] which defines automatic acoustic management (AAM). In this paper we emphasize the difference between power and energy. *Power* is an instantaneous measurement while *energy* is the overall power consumed over a given interval.

2 Related Work

There are several recent works on power reduction in storage systems utilizing ideas such as spinning down disks during idle time and taking advantage of caching for the purpose of increasing idle time [9], [10], [13], [16], [17], [18], [26].

A large body of research deals with multiple speed disks, also known as dynamic RPM (DRPM). While acoustic modes affect the disk-head speed, DRPM deals with the disk's rotational speed. In notable contrast to acoustic modes, there are currently no available disks that support DRPM. The works of [7], [12], [16], [18], [25] show that adapting the disk rotational speed to the required performance level can reduce power consumption.

A power simulator called *Dempsey* [24] reads I/O traces and interprets them for power and performance using DiskSim [6]. Dempsey was tested on mobile disk drives and does not take into account the effect of disk arrays. Since it requires exact traces, it cannot be used as a predictive tool. Another simulator was presented in [19], whose goal was investigating disk design optimizations for power, performance, and capacity. Stoess et al. [20] model power consumption based on disk utilization. Their model takes into account disk transfer rates and response times, but ignores the effect of seek operations on power consumption. Recently, Hylick et al. [14], [15] studied disk drive power dissipation, but they did not address storage arrays.

3 Storage Power Modeling

Our power modeling framework computes the power consumption of each storage I/O path component as it handles an

I/O operation from the time the I/o request is received and until the time the request processing is completed. The framework takes into account workloads, power states, and configurations. In addition to modeling power consumption of the I/O path, the model also takes into account power consumed by the storage components while idle. The method can be applied to single disk drives or to a storage array (e.g., a RAID array). We use the term storage controller when referring to a storage array.

3.1 The Model

It is common practice for storage controllers to report statistical performance counters (information) for each type of I/O workload operation. These operations are: sequential read, sequential write, random read, and random write. Performance counters typically include the rate of each type of operation, the transfer sizes, response time, and other statistical information. Our framework uses performance counters and differentiates between the I/O workload at the frontend of the storage controller and at the backend of the controller. The *frontend workload* refers to the I/O operations arriving from the host. The *backend workload* refers to the actual I/O operations performed by the disks. It is the backend workload which determines the power consumption of the disks.

The backend workload is affected by the read and write caching activities, by virtualization layers, and by resiliency (e.g., RAID) mechanisms. Caching activities include caching read data, performing read ahead during sequential data access (pre-fetching), and delayed (cached) writes. Caching leads to less disk activity and therefore, lower power consumption. The virtualization and resiliency layers influence the backend workload, as data can be organized into stripes across the disks in the array and write operations are translated into write transactions. For example, in a RAID 10 array, two copies of the data must be updated. Therefore, computing the backend workload from the frontend workload requires taking into account the type of operation, transfer size, data organization across stripes, etc. In order to estimate the power cost of disk I/O operations, our model calculates how many backend disk operations are needed for each type of workload. We then estimate the dynamic power cost of those backend I/O operations based on the estimated number of seeks and on the amount of transferred data.

Our framework uses a small dataset of power consumption tables, one for each storage component, to compute the dynamic power consumption. The dataset consists of power consumption values for various amounts of backend operations. For example, the dataset includes power consumption data for various amounts of data transferred and various seek rates. Building the dataset is a one-time process for each type of storage array.

The framework consists of: (i) translating frontend workloads to backend workloads; and (ii) using interpolation to estimate the power consumption of each activity, based on the pre-computed dataset. Additional details on the process can be found in [5].

3.2 Validation

We performed extensive validation runs over several types of disks and RAID configurations, using a variety of I/O access patterns and disk utilization levels. We ran various micro-benchmarks, using Iometer [1] and an industry standard SPC-1-like workload [2] to examine the accuracy of the power modeling estimations.

Disk drive results. When comparing our modeling power estimation with the actual power measured for a single disk we have observed an average modeling error of less than 3% and maximal error of 6.5% for a $15K$ $300GB$ disk. For a $10K$ $300GB$ we have observed an average modeling error of less than 5.2% and maximal error of 10%.

Disk array results. We validated our results on a RAID 5 array in a mid-range enterprise controller populated with 16 146GB 10K enterprise disks. For random read (write) workloads with transfer sizes ranging from $4K$ to $512K$ (up to $64K$, respectively) we observe a modeling estimation error of less than 5%. For larger random write transfer sizes, $128K$ to $512K$, we observe a modeling estimation error of up to 10%. We have also run SPC-1-like workloads showing a maximal power estimation error of 2.5%.

4 Power Management Using Acoustic Modes

For our investigation of acoustic modes we measured the performance and power consumption of disk drives. We use a custom-made LabVIEW [4] application for measuring the power consumption of the disk drives. Vdbench [21], a Java-based open-source tool, was used for running I/O workloads. We ran random access micro-benchmarks using Vdbench to observe the effect of the differences in seek operations in normal and in quiet acoustics.

In addition, in order to understand the effect on real-world workloads, we ran an industry standard SPC-1 workload [2]. The SPC-1 workload is a synthetic, yet sophisticated and fairly realistic, online transaction processing (OLTP) workload.

We studied a high capacity Hitachi HUA721010KLA330 1TB 3.5" disk drive, which supports acoustic modes. A full and detailed report on how acoustic modes affect the power and energy profile can be found in [8].

4.1 Power Capping

We analyzed the behavior of a seek operation in normal and quiet modes by reviewing the power profile of a single seek operation. We sampled the power dissipation of a single *long-range* seek, from one end of the platter to another, at a rate of 50K samples per second. We observed the power dissipation for the different phases of a seek operation:

Acceleration: During this phase the seek head accelerates to its maximum speed. In quiet mode the acceleration is slower, so the power dissipated at any given time throughout this phase is less than in normal mode. Only the 12V power dissipation is impacted here.

Coast: In this phase the disk head remains at its maximum speed (the maximum speed of quiet mode is slower than that for normal mode). The power dissipated at any time throughout this phase is about the same in both normal and

quiet modes. This phase lasts longer in quiet mode, causing more overall energy to be consumed per seek.

Deceleration: During this phase the disk head is slowed down by reversing the current direction of the voice coil motor (VCM). The power dissipation is generally similar to that of the acceleration phase. In quiet mode less power is dissipated at any given time.

Data transfer: During this phase, the disk head is at a complete standstill, and the data is being transferred to and from the disk. The 5V power dissipation increases, but the behavior is the same in both normal and quiet mode.

Although the power in quiet mode can be capped at 73% of the power dissipated in normal mode, the overall energy consumed by a single long-range seek operation is greater in quiet mode. For example, our analysis of 12V and 5V power consumption shows that overall energy (both 12V and 5V components) consumed by a single seek operation is 17% greater for quiet mode than for normal mode. This is due to: *i*) the fact that the duration of the long-range seek is longer, since the head moves the same distance but at a lower velocity; and *ii*) the 12V power decreases only during acceleration and deceleration (and not during coast), while the 5V power is not affected by the acoustic mode. This leads to a 5V energy consumption increase of 49% and 12V energy consumption increase of about 3%.

4.2 Energy Reduction

We now investigate the energy consumption of various workloads when using quiet mode. One effect of running in quiet mode is that moving the disk head takes longer - that is, the seek time increases. Since the disk power consumption has a static component, a seek operation that takes longer may consume more energy, depending on the balance between the saved energy of the slower acceleration and deceleration and the added energy for longer seek time.

We simulated a real-world online transaction processing workload by running SPC-1 workloads. SPC-1 is a concurrent workload composed of random reads, random writes, and sequential access across various parts of the disk drive. We generated an SPC-1 I/O trace and replayed the I/O trace in normal and quiet modes. We ran the benchmark at three I/O rates of 10, 25 and 50 I/O's per second. We measured the power consumption and computed the energy in Joules of each of the three runs. In all cases both the power consumption and the total energy consumed was lower for the quiet mode. The energy saving was between 2.2% for 10 I/O's per second and 12.54% for 50 I/O's per second. We executed I/O's at the same rate for both normal and quiet modes. At low I/O rates, 10 and 25 I/O's per second, the response time increased slightly. In these cases, when running in normal mode, the disk is in fact idle in between some I/O operations. In quiet mode, the seeks take longer, and the disk has less or no idle time. In this case, we exchange wasted disk idle time, when power is also consumed, with a longer and slower seek. Running at 50 I/O's per second results in little or no idle time, even in normal mode. The I/O requests are generated at the same rate both in normal and quiet modes. However, in quiet mode, the disk serves these requests at a slower rate, which may cause a longer queue of I/O's to form based on the fact that the response time doubles.

There are examples of workloads for which the use of quiet mode leads to an *increase* in the overall (total) energy consumption. We generated and executed a trace of $30,000$ random-read I/O's using $1, 2$, and 4 concurrent I/O threads in both normal and quiet modes. Each thread executed the I/O's synchronously without delay. We measured the power consumption and computed the energy consumption of each execution. When using less than 4 concurrent threads, the energy consumption is notably higher in quiet mode due to longer seek times. Longer seek times, in turn, lead to a longer run time. When the number of I/O threads increased to 4 we achieved a reduction in total energy. Using 4 concurrent threads we achieved an energy savings of over 2%, but when using 1 or 2 concurrent threads the energy consumption increased by nearly 6%.

4.3 Application Performance

We analyzed the impact of acoustic modes on application performance. Runing an online-generated SPC-1-like workload in both normal and quiet modes shows that in quiet mode we are able to achieve only up to 55 I/O's per second, while in normal mode we can reach more than 70 I/O's per second. At I/O rates up to 20 I/O's per second the response time in quiet mode is only slightly higher than in normal mode. However, beyond 20 I/O's per second the response time in quiet mode increases significantly and reaches, at 55 I/O's per second, almost double the response time as in normal mode.

5 Concluding Remarks

Power modeling and estimation. Our power modeling and estimation methods are based purely on performance information. Therefore, any inaccuracy in the performance data leads to an estimation inaccuracy as well. We have encountered cases where the controller failed to correctly identify the workload pattern. For example, incorrectly reporting a sequential stream as a random stream introduces errors to the estimations. Another possible source of inaccuracy is lack of information regarding background tasks (e.g., bit scrubbing, battery maintenance); better reporting of background activity will lead to improved accuracy.

Our power modeling can be used in a power-aware capacity planning tool predicting the power consumption based on the given configuration and workloads. Our modeling can also provide online power estimations, per disk array and disk volume, for storage systems.

Power and energy management using acoustic modes. We have explored the effects of acoustic management on performance and power consumption. While acoustic management can in some cases be applicable for energy savings, it is always effective for power capping (or budgeting).

Quiet acoustic modes change the way disks perform seek operations, so there is no power reduction when no seeks are performed, for example, during idle time or during sequential access. Since only seeks are affected, the power for

the electronics remains the same. This limits the ability of acoustic modes to save power. For random-read workloads the power reduction is at most 23%, depending on the actual I/O workload.

Quiet mode causes an increase in response time. This prevents the use of quiet mode for mission-critical applications that are sensitive to I/O response time. Single-threaded applications that require high throughput will suffer a 25% reduction in I/O throughput. Moreover, they will consume more energy in quiet mode than in normal mode, but will benefit from a lower peak power consumption. Multi-threaded applications with a mixed workload of read and write operations, both random and sequential, will be able to sustain the same I/O throughput, but with longer response time. Such applications may need to use a larger number of threads while using quiet mode, in order to sustain the same I/O throughput as in normal mode.

We have found that in some cases, seek operations consume more overall energy in quiet mode than in normal mode, though they consume less instantaneous power. We have also encountered workloads for which quiet mode leads to energy savings. The SPC-1 workload tests clearly demonstrate that OLTP applications are good candidates for energy savings, when they can tolerate a degradation in response time.

REFERENCES

[1] "Iometer, performance analysis tool." http://www.iometer.org/.

[2] "Storage performance council," http://www.storageperformance.org/.

[3] "INCITS 361-2002 (1410D): AT attachment - 6 with packet interface (ATA/ATAPI - 6)," 2002.

[4] "LabVIEW release notes," 2009, http://www.ni.com/pdf/manuals/371778e.pdf.

[5] M. Allalouf, Y. Arbitman, M. Factor, R. Kat, K. Meth, and D. Naor, "Storage modeling for power estimation," in *Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference*, 2009.

[6] J. S. Bucy, J. Schindler, S. W. Schlosser, G. R. Ganger, and Contributors, "The DiskSim simulation environment - version 4.0 reference manual," May 2008.

[7] E. V. Carrera, E. Pinheiro, and R. Bianchini, "Conserving disk energy in network servers," in *Proceedings of the 17th Annual International Conference on Supercomputing*, June 2003, pp. 86–97.

[8] D. Chen, G. Goldberg, R. Kahn, R. I. Kat, K. Meth, and D. Sotnikov, "Leveraging disk drive acoustic modes for power management," in *MSST'10 Research Track: Proceedings of the 26th IEEE Conference on Mass Storage Systems and Technologies (MSST2010): Research Track*, 2010.

[9] D. Colarelli and D. Grunwald, "Massive arrays of idle disks for storage archives," in *Proceedings of the 2002 ACM/IEEE conference on High Performance Networking and Computing*, November 2002, pp. 1–11.

[10] F. Douglis, P. Krishnan, and B. Bershad, "Adaptive disk spin-down policies for mobile computers," in *Proceedings of the 2nd USENIX Symposium on Mobile and Location-Independent Computing*, April 1995, pp. 121–137.

[11] EPA, "Epa report to congress on server and data center energy efficiency," *Public Law 109-431*, 2007.

[12] 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*, June 2003, pp. 169–181.

[13] 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*, March 2003, pp. 123–132.

[14] A. Hylick, A. Rice, B. Jones, and R. Sohan, "Hard drive power consumption uncovered," *SIGMETRICS Performance Evaluation Review*, vol. 35, no. 3, pp. 54–55, 2007.

[15] A. Hylick, R. Sohan, A. Rice, and B. Jones, "An analysis of hard drive energy consumption," in *MASCOTS*, 2008, pp. 103–112.

[16] X. Li, Z. Li, F. David, P. Zhou, Y. Zhou, S. Adve, and S. Kumar, "Performance directed energy management for main memory and disks," in *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM Press, 2004, pp. 271–283.

[17] D. Peek and J. Flinn, "Drive-thru: fast, accurate evaluation of storage power management," in *ATEC '05: Proceedings of the annual conference on USENIX Annual Technical Conference*. Berkeley, CA, USA: USENIX Association, 2005, pp. 30–30.

[18] E. Pinheiro and R. Bianchini, "Energy conservation techniques for disk array-based servers," in *Proceedings of the 18th Annual International Conference on Supercomputing*, June 2004, pp. 68–78.

[19] S. Sankar, Y. Zhang, S. Gurumurthi, and M. R. Stan, "Sensitivity-based optimization of disk architecture," *IEEE Trans. Comput.*, vol. 58, no. 1, pp. 69–81, 2009.

[20] J. Stoess, C. Lang, and F. Bellosa, "Energy management for hypervisor-based virtual machines," in *ATC'07: 2007 USENIX Annual Technical Conference on Proceedings of the USENIX Annual Technical Conference*. Berkeley, CA, USA: USENIX Association, 2007, pp. 1–14.

[21] H. Vandenbergh, "Vdbench 5.00 users guide," 2008, http://garr.dl.sourceforge.net/project/vdbench/vdbench/Vdbench%205.00/vdbench.pdf.

[22] R. Villars, "Three keys for storage success: Content, architecture and getting personal. IDC," July 2007.

[23] S. W. Worth, "SNIA green storage tutorial," 2007, http://www.snia.org/forums/green/programs/SWWorth_Green_Storage.pdf.

[24] J. Zedlewski, S. Sobti, N. Garg, F. Zheng, A. Krishnamurthy, and R. Wang, "Modeling hard-disk power consumption," in *FAST '03: Proceedings of the 2nd USENIX Conference on File and Storage Technologies*. Berkeley, CA, USA: USENIX Association, 2003, pp. 217–230.

[25] Q. Zhu, Z. Chen, L. Tan, Y. Zhou, K. Keeton, and J. Wilkes, "Hibernator: helping disk arrays sleep through the winter," in *Proceedings of the Symposium on Operating Systems Principles (SOSP)*, October 2005, pp. 177–190.

[26] Q. Zhu, F. M. David, C. F. Devaraj, Z. Li, Y. Zhou, and P. Cao, "Reducing energy consumption of disk storage using power-aware cache management," in *HPCA '04: Proceedings of the 10th International Symposium on High Performance Computer Architecture*. Washington, DC, USA: IEEE Computer Society, 2004, p. 118.