

Should Disks be Speed Demons or Brainiacs?

Sudhanva Gurusurthi
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
gurusurthi@cs.virginia.edu

ABSTRACT

Disk drives play a critical role on the performance of I/O intensive applications. Over the years, disk drive performance has grown as a result of advances in magnetic recording density and faster rotational speeds. In essence, the performance driver in disks has been the data rate. In this paper, we show that data rate is going to be increasingly difficult to optimize, due to power/thermal constraints. We argue that disk drive designers should instead focus their efforts on providing more computational capabilities that data intensive applications could leverage in order to boost performance. We also discuss the scope for provisioning powerful processors inside disk drives to provide these computational capabilities.

1. INTRODUCTION

The storage system is critical to the performance of a wide range of server applications, from transaction processing and web services to scientific computing. These applications process large volumes of data, and have to do so with a low turnaround time. For example, many businesses post advertisements for their products via online portals. These hosting companies earn revenue based on the number of users who click on these advertisements. Various statistics are gathered on each individual click, which can then be studied to infer trends about user preferences. This database of user access information can be several tens to hundreds of Terabytes in size. To maximize revenue, the business analyst often needs to interactively query this database whose contents might be continuously changing, say, due to round-the-clock customer transactions over the Internet. Each query to such a database can take several tens of hours, even on a high-performance clustered database server [23].

In recent years, data intensive applications have started becoming more and more prevalent in personal computing as well. For example, with the advent of digital photography and disk drives that can store hundreds of gigabytes of data, we all accumulate a large collection of images on our per-

sonal computers. After some time, we might want to search through this collection for specific content, say, a particular type of scenery or a prominent landmark. However, most users are not assiduous enough to provide keywords that indicate the content of each image to facilitate a text-based search, and therefore the actual content of the images would need to be analyzed. This search procedure, which is known as Content-Based Image Retrieval (CBIR) [22], consists of a complex processing pipeline and is very I/O intensive. Running CBIR applications efficiently requires considerable processing resources, as well a high performance I/O system.

The I/O performance is largely determined by the disk drives. Disk drive performance depends on several factors, including, the linear density, rotational speed (RPM), seek time, and the size of the on-board cache. Over the past two decades, the performance of disk drives (i.e., the data rate) has been growing at 40% per year. The main drivers for this growth have been density and higher rotational speeds, along with some reductions in the seek time [13, 7]. Although this scaling of the data rate has been instrumental in improving I/O performance, there are a number of physical constraints that pose a big problem to continued scaling at this 40% rate [11].

In this paper, we argue that instead of focusing solely on disk access time, it might be more worthwhile to invest effort into providing *processing power* at the disks. In other words, we should transform disk drives into *data-processing devices*, rather than merely use them for data storage and transfer. Such “active disks” have been proposed in the past [1, 16, 18] and their benefits have been documented in the literature. However, except in specific circumstances (e.g., encryption [21]), there are no commercial disk drives that provide this feature. A major reason for this is power. Disk drives have relatively low power budgets and the bulk of the power is consumed by the electro-mechanical parts, such as, the spindle motor and the arm actuator. With emphasis placed solely on optimizing the data rate, disk drive designers have focussed on making the data transfer system fast, within the device power constraints, leaving little to no room for optimizing the disk processor capabilities. Instead, storage-side processing is provided via high-end storage clusters, whose hardware, power, and cooling costs are significantly higher than those of disk drives.

The EMC Centra [5], Netezza Performance Server [24], and the Exegy TextMiner [6] are three examples of storage so-

lutions that implement data processing operations close to the disks. These products use high-end CPUs, sometimes in conjunction with FPGAs [24, 6], to provide the functionalities of active disks. The EMC Centera, which is a content-addressable object-based storage system that provides features for data audit and compliance management, is a clustered storage system, with anywhere from 4 to 128 nodes. Each Centera node consists of a high-end Pentium 4 type CPU and multiple SATA disk drives [19]. Even the low-end Pentium 4 processors consume over 50 Watts of power, which is an order of magnitude higher than a typical disk processor. The Netezza Performance Server, which is designed for data warehousing applications (e.g., the business analytics application described earlier) and consists of a single high-performance host, which runs a SQL database, communicating with several Snippet Processing Units (SPU). A SPU is the equivalent of an active disk and consists of a 500 MHz PowerPC processor, a FPGA, and an ATA disk drive [23]. The Exegy TextMiner appliance is again intended for accelerating searches on large unstructured/unindexed datasets. TextMiner uses two dual-core AMD Opteron processors, reconfigurable logic, and over 16 GB of RAM to process the data on its disks.

In this paper, we argue that instead of going in for high-end power-hungry solutions, we can build computationally powerful active storage systems *within the power budgets of the disk drives*. Building such active storage systems requires a holistic view of disk power, factoring in the power of the electro-mechanical and electronic parts. This is achievable by reducing the power consumption of the electro-mechanical parts, notably the spindle motor, and diverting this “surplus power” towards the processing subsystem. Borrowing terminology from processor architecture [12], we have the opportunity to transform disk drives from mere data rate “speed demons” to data processing “brainiac” devices.

The next section describes the power/thermal problems facing performance scaling in disk drives. Section 3 shows the scope for computationally powerful active disks and Section 4 concludes this paper.

2. THE DISK DRIVE "THERMAL WALL"

A hard disk drive is composed of electro-mechanical and electronic parts. The electro-mechanical parts include the spindle motor (SPM) and the platters, and the arm actuator motor (also called the Voice Coil Motor or VCM) and the disk arms. The electronic parts include the disk controller, which is an ARM-type core running at around 200 MHz [2], and 8-32 MB of DRAM memory. The data channel is used to transport the bits between the platter media and the electronic buffers in the hard drive. Among these different components, the SPM tends to dissipate the highest amount of power, primarily due to viscous dissipation at high RPMs [10].

Designing disk drives involves tradeoffs between capacity, performance, and power. The capacity of a disk drive can be increased through the use of larger and/or multiple platters. The number of platters and their size affects the heat that is generated inside the disk drive, due to viscous dissipation, by a linear factor and by nearly the fifth power respectively [15]. The data rate of the disk drive can be increased through

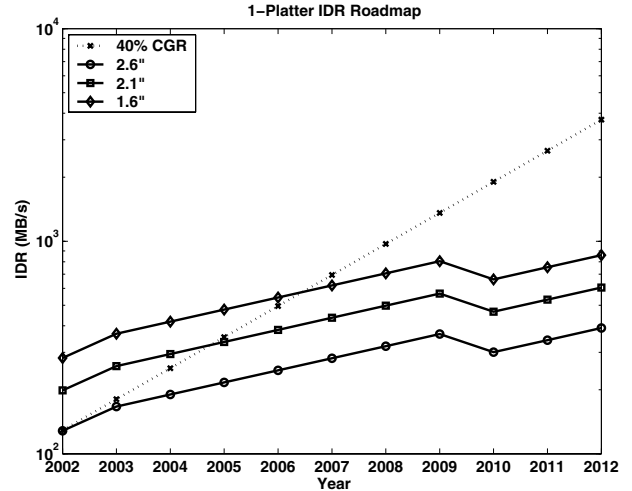


Figure 1: Disk Drive Roadmap. Each solid curve gives the maximum attainable Internal Data Rate (IDR) within the thermal envelope. The dotted line indicates the 40% IDR growth that we have enjoyed for the past two decades.

improvements in the linear density and higher disk RPMs. The latter causes the heat that is generated to increase by nearly a cubic factor. One of the requirements in disk drive design, intended for reliability, is to always keep the operating temperature below a particular threshold, known as the *thermal envelope*. Therefore, for a given maximum external ambient temperature, every disk is designed such that, even under worst-case operating conditions, its temperature does not exceed this threshold. The performance improvements within this thermal-constrained design space are obtained through a combination of density increases and structural changes to the disk drive. The structural modifications involve shrinking the platters, which has the fifth power impact, and exploiting this slack to ramp up the RPM. Moreover, in a rack-based server system, the disks may be in close proximity to the processors and memory cards, which themselves have thermal constraints. Excess heat from the disks can preheat the air around these other components and vice-versa. Given the high costs associated with cooling modern electronic systems, it is important that disk drives do not increase this burden even further. Looking at this thermal constraint in another way, disks need to operate within a constant maximum power.

For nearly the past two decades, this thermal-constrained design methodology has allowed the data rate of disks to grow briskly from year to year, and has been a leading driver of I/O performance. However, this approach is facing serious scalability issues. In [11], it was shown that the data rate would grow at the rate of 14% per year, down from the 40% rate. This trend is shown in Figure 1. These scalability issues are due to a combination of hitting certain fundamental limits in scaling the size of magnetic bit cells [3] and thermal limitations on how aggressively we can scale up the drive RPM to compensate for these density problems. There is already some evidence of these problems, with newly introduced disks running noticeably hotter than

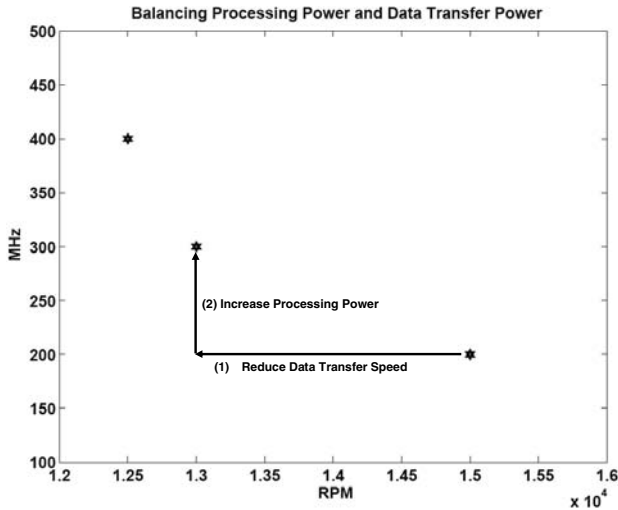


Figure 2: Trading off data transfer speed for higher disk processor clock frequency. The arrows show the steps involved in doing the tradeoff. All three points on the graph have nearly equivalent power consumption.

previous generations of the same product [20] and a slowdown in the data rate growth [8]. Unfortunately, temperature mitigation techniques for disk drives could lead to severe performance degradation [9]. All these factors point to the end of the road for RPM scaling as the primary performance enhancing technique.

3. THE SCOPE FOR HIGH-PERFORMANCE ACTIVE DISKS

A key advantage of active disks is that we can filter data that needs to be transported to the host by processing them at the disk. For example, if we wish to search the data stored on disk for specific content, we could perform the search operation locally at the disk controller and return only the result to the host [1, 18]. If there are multiple disks in the storage system (which is quite common in servers), the parallel processing of the data can provide significant performance boosts. However, based on the nature of the computation that we wish to do, we may require powerful processors at the disk, which would need to be accommodated within the power budget of the disk drive. Another benefit of active storage is that, since data does not have to move all the way from the disk to the host, we could ease the requirements on the data rate. We can reduce the data rate of the disk by lowering the RPM and scaling down the frequency of the data channel commensurately. This data rate reduction would also lower the disk power consumption. One straightforward benefit of this power reduction is that we could reduce the overall energy consumption of the storage system. Another possible benefit is that the power could be “re-purposed” to boost the performance of the embedded processor inside the disk. The availability of more computing power would provide more flexibility (and motivation) to implement active disks. We now show that it is indeed feasible to provision more computational capabilities inside disk drives without exceeding their power budgets.

Hardware Device	Power (Watts)
Spindle Motor @ 15,000 RPM	6.36
Spindle Motor @ 13,000 RPM	5.88
Spindle Motor @ 12,500 RPM	5.7
Windage Loss @ 15,000 RPM	0.90
Windage Loss @ 13,000 RPM	0.60
Windage Loss @ 12,500 RPM	0.54
Data Channel	2.00
200 MHz Disk Processor	1.64
300 MHz Disk Processor	2.06
400 MHz Disk Processor	2.6
Disk Main-Memory	0.523

Table 1: Component-Wise Breakdown of Disk Power Consumption.

The performance of an active disk is a function of the characteristics of the disk processor and the underlying electro-mechanical data transfer system. Consider a typical SCSI disk drive. Assume that it is a 15,000 RPM disk drive with a single 2.6” platter, has a 200 MHz ARM-based disk processor and a 16 MB disk cache. For such a disk, the power drawn by the SPM assembly is 7.257 Watts (calculated using the device measurements given in [4] and detailed thermal modeling tools [11]). The disk processor is assumed to be an Intel XScale-based PXA 255 processor [14], which supports the ARM ISA and is available in 200, 300, and 400 MHz clock frequencies. The processor datasheet gives the maximum power consumption at these three frequencies to be 1.64 Watts, 2.06 Watts, and 2.6 Watts respectively. The disk cache is assumed to be a single Micron 128 Mb mobile SDRAM part and its power consumption is 0.523 Watts [17]. By interacting with engineers at Seagate, it was found that the data channel of such a disk consumes 2 Watts of power. Let us assume that the disk is in the active mode, where it is transferring data to or from the platters. In this mode, the disk processor, disk cache, and data channel are powered, the platters are spinning, and the disk arm is stationary (and therefore does not consume any power). The disk consumes 11.42 Watts of power when it is in this mode. Let us designate this as the baseline configuration.

The graph in Figure 2 shows three points that consume the same amount of power as the baseline. The component-wise breakdown of power for these three configurations is given in Table 1. When the RPM of the disk drive is reduced from 15,000 to 13,000 (as shown by the horizontal arrow in Figure 2), the disk power drops to 10.64 Watts. This reduction in power allows us to increase the disk processor frequency from 200 to 300 MHz, as shown by the vertical arrow, thereby providing us with a faster processor for the same power consumption. When the RPM is further lowered to 12,500, the clock frequency can be increased to 400 MHz. Since the SPM accounts for a large fraction of the disk power and has a cubic relation to the RPM, even small reductions in the rotational speed are sufficient to attain significant boosts in the disk processor clock frequency.

4. CONCLUDING REMARKS

Active disks provide the opportunity to perform general purpose computation close to the data. Given the problems associated with the scaling up of data rates and the gap between the power consumption of the electro-mechanical

parts and the embedded processor within disks, we are at the juncture where we should consider “activeness” as the key driver for improving disk performance.

5. ACKNOWLEDGMENTS

This research has been funded in part by NSF grants 0551630 and 0627527. We would like to thank Shahrukh Tarapore for his valuable feedback on this paper.

6. REFERENCES

- [1] A. Acharya, M. Uysal, and J. Saltz. Active Disks: Programming Model, Algorithms and Evaluation. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 81–91, October 1998.
- [2] ARM Collaborates With Seagate For Hard Disc Drive Control, June 2002. ARM Press Release.
- [3] S. Charrap, P. Lu, and Y. He. Thermal Stability of Recorded Information at High Densities. *IEEE Transactions on Magnetics*, 33(1):978–983, January 1997.
- [4] S. Chen, Q. Zhang, H. Chong, T. Komatsu, and C. Kang. Some Design and Prototyping Issues on a 20000 RPM HDD Spindle Motor with a Ferro-Fluid Bearing System. *IEEE Transactions on Magnetics*, 37(2):805–809, March 2001.
- [5] EMC Centera, 2005.
<http://www.emc.com/products/systems/centera.jsp>.
- [6] Exegy TextMiner (Whitepaper).
<http://www.exegy.com>.
- [7] E. Grochowski and R. Halem. Technological Impact of Magnetic Hard Disk Drives on Storage Systems. *IBM Systems Journal*, 42(2), 2003.
- [8] W. Gruener. Seagate Ships First Perpendicular 3.5” Hard Drive. *TG Daily*, April 2006.
http://www.tgdaily.com/2006/04/17/seagate_announces_cheetah_perpendicular/.
- [9] S. Gurumurthi. The Need for Temperature-Aware Storage Systems. In *Proceedings of the Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM)*, pages 387–394, May 2006.
- [10] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke. DRPM: Dynamic Speed Control for Power Management in Server Class Disks. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*, pages 169–179, June 2003.
- [11] S. Gurumurthi, A. Sivasubramaniam, and V. Natarajan. Disk Drive Roadmap from the Thermal Perspective: A Case for Dynamic Thermal Management. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*, pages 38–49, June 2005.
- [12] L. Gwennap. Comparing RISC Microprocessors. In *Proceedings of the Microprocessor Forum*, October 1994.
- [13] Hitachi Global Storage Technologies - HDD Technology Overview Charts, 2003.
<http://www.hitachigst.com/hdd/technology/overview/storagetechart.html>.
- [14] Intel PXA255 Processor - Electrical, Mechanical, and Thermal Specification, February 2004.
<http://www.intel.com/design/pca/applicationsprocessors/manuals/278780.htm>.
- [15] I.Sato, K. Otani, M. Mizukami, S. Oguchi, K. Hoshiya, and K.-I. Shimokura. Characteristics of Heat Transfer in Small Disk Enclosures at High Rotation Speeds. *IEEE Transactions on Components, Packaging, and Manufacturing Technology*, 13(4):1006–1011, December 1990.
- [16] K. Keeton, D. Patterson, and J. Hellerstein. The Case for Intelligent Disks (IDISKS). *SIGMOD Record*, 27(3):42–52, September 1998.
- [17] Micron Mobile SDRAM Products.
<http://www.micron.com/products/mobileddram/sdram/>.
- [18] E. Riedel, G. Gibson, and C. Faloutsos. Active Storage for Large-Scale Data Mining and Multimedia. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 62–73, August 1998.
- [19] J. Schindler. EMC Corporation, May 2006. Private Correspondence.
- [20] P. Schmid and A. Roos. Hitachi’s 500GB DeskStar Monster. *Tom’s Hardware*, July 2005.
<http://www.tomshardware.com/2005/07/14/hitachi/page7.html>.
- [21] Seagate Introduces World’s First 2.5-Inch Perpendicular Recording Hard Drive; First Major HDD Maker to Deliver Notebook PC Drive With Hardware-Based Full Disc Encryption Security, June 2005. Seagate Press Release.
- [22] A. Smeulders, M. Worrington, S. Santini, A. Gupta, and R. Jain. Content-based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.
- [23] H. Smith. Delivering on the Promise of Business Analytics, Technology Roundtable IV Talk, November 2005.
- [24] The Netezza Performance Server 8000 Series - Whitepaper, 2005.
<http://www.netezza.com/products/products.cfm>.