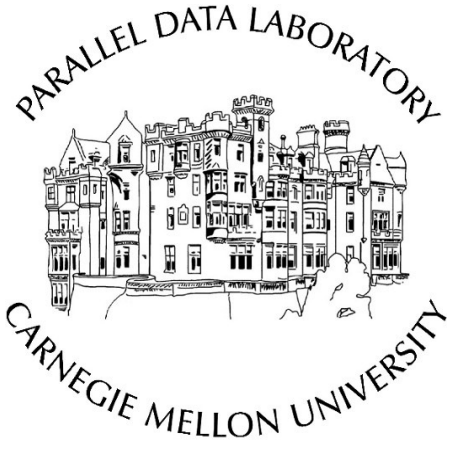


Disks Are Like Snowflakes: No Two Are Alike

Elie Krevat*, Joseph Tucek†, and Gregory R. Ganger*
*Carnegie Mellon University †HP Labs

CMU-PDL-11-102
Feb 2011

Parallel Data Laboratory
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Abstract

Gone are the days of homogeneous sets of disks. Even disks of a given batch, of the same make and model, will have significantly different bandwidths. This paper describes the disk technology trends responsible for the now-inherent heterogeneity of multi-disk systems and disk-based clusters, provides measurements quantifying it, and discusses its implications for system designers.

Acknowledgements: We thank the members and companies of the PDL Consortium (including APC, EMC, Facebook, Google, Hewlett-Packard Labs, Hitachi, IBM, Intel, LSI, Microsoft Research, NEC Laboratories, NetApp, Oracle, Riverbed, Samsung, Seagate, STEC, Symantec, VMWare, and Yahoo! Labs) for their interest, insights, feedback, and support. This research was sponsored in part by an HP Innovation Research Award and by CyLab at Carnegie Mellon University under grant DAAD19–02–1–0389 from the Army Research Office.

Keywords: adaptive zoning, disk performance, linear density, track density, variable recording

1 Introduction

Many systems are designed and built assuming uniformity of performance. People buy identical hardware, configure them the same, and expect to achieve uniform performance across them all. Assuming homogeneity simplifies load balancing, allows for easier distribution of work when parallelizing tasks (e.g., disk striping), and facilitates effective performance tuning and debugging.

Until recently, this assumption worked quite well for disk drives and the systems that depend on them. When a particular disk drive didn't perform the same way as others of the same model, it was usually a faulty disk. Now, every disk has, by design, unique performance characteristics individually determined according to the capabilities of its physical components; for a given system setup and workload, and even for the same physical region on disk, some disks are slower, some disks are faster, and no two disks are alike.

In fact, disk performance varies in new ways both within a disk and across same-model disks. For years, disk speed has varied across "zones", which are groups of co-located tracks used to pack more sectors onto the longer, outer tracks [19]. Until recently, the zone arrangements (e.g., sectors per track, tracks per zone) were the same for every surface of every disk of a given model. Now, they aren't; in modern disks, the density of each surface is unique. As a result, under normal operation, disk bandwidth to/from corresponding regions of a set of disks can be expected to vary by 20% or more from the fastest to the slowest.

This paper explains the source and characteristics of this new non-uniformity of disk drives, and discusses its implications. Briefly, the root cause is manufacturing variations, especially of the disk head electronics, that were previously masked and are now being exploited. Like CPUs that are binned by clock frequency, different disk heads can store and read data at different maximum densities. Instead of only using each head at pre-specified densities, wasting the extra capabilities of most, manufacturers now configure per-head zone arrangements, running each head as densely as possible. We refer to this approach as *adaptive zoning*. The upside is bigger, cheaper, and faster disks. The downside is the much more varied and non-homogeneous bandwidths on which this paper focuses, since disk bandwidth is directly proportional to per-track storage density.

Despite relative quiet regarding this new feature, we have found evidence of adaptive zoning being used by all major disk manufacturers, ranging from patent applications to measurements to informal conversations with employees. We have experimentally confirmed adaptive zoning being used in a number of disk makes and models, and we report example data in this paper. In a sample of identically labeled disks of the same model, we have measured bandwidths that range from 5.8% faster to 14.5% slower than the average across the disks. Furthermore, this range seems to be growing over generations of disk drives. Similar bandwidth variation is also visible between adjacent blocks that cross over to different surfaces in each disk, since each head and surface combination provides a distinct bandwidth.

Many systems assume homogeneity and, in its absence, will be inherently inefficient. For example, RAID systems [13] and high performance computing file systems that stripe data across many disks [7] will operate at the speed of the slowest disk. We first perceived this issue of disk non-uniformity while partitioning work for a prototype parallel dataflow system across a set of identical nodes and observing delays due to slower disks. In general, any system that assumes the same performance from "equal" disks will waste resources waiting for the slowest across the sizable range of their speeds.

With the changes in modern disks, heterogeneity now has to be expected in all distributed systems that rely on disks. Other work has effectively argued that performance assumptions need to be avoided in scale-out distributed systems, that hardware heterogeneity is non-trivial to control, and that programs should respond to system behavior dynamically to optimize performance [1, 2]. Even if the hardware and software performed homogeneously, there are many subtle sources of performance variation, such as room temperature affecting CPU clock speeds [11]. These are all compelling arguments. However, many have disregarded this advice in the past and relied on careful control of the hardware, software, and computing environment to make efficient use of their resources. If disks are involved, this is no longer an option.

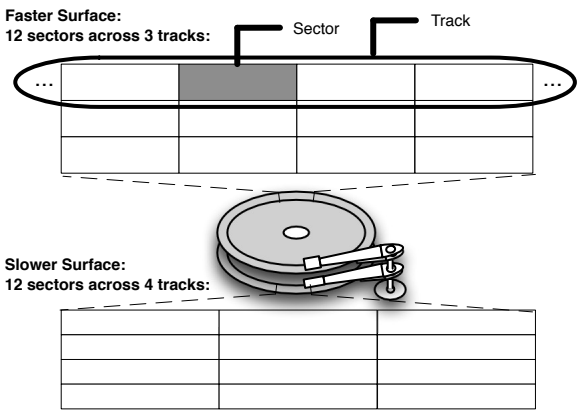


Figure 1: **Adaptive zoning of disk drives.** The same area of different disk surfaces within a drive are formatted according to the physical properties of the disk head. Each of the surfaces pictured have equivalent areal density, although the top surface is faster because it accommodates more sectors per track over fewer tracks, while the bottom surface has a narrower disk head that requires fewer sectors per track but allows for more tracks.

2 Advances in Disk Technology

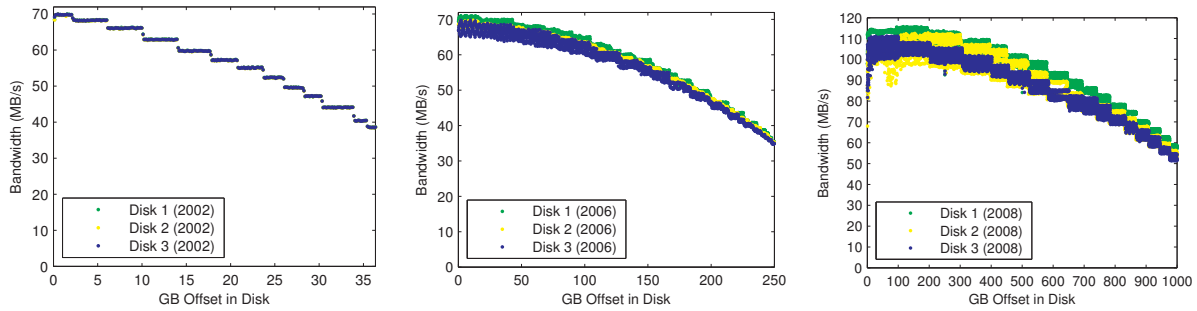
Magnetic disk drives have come a long way since their 1950s debut, constantly being refined while maintaining the same basic design mechanisms: rotating platters coated with magnetic material, and on each surface of a platter, a moveable head that induces a magnetic field to read and write data. The smallest unit on a disk is a sector of 512 bytes. Most disks spin at fixed rates, typically 5400, 7200, 10K, or 15K RPM. There have been many improvements in disk technology, with the goal of increasing capacity, reliability, and speed, while reducing size, cost, and power. These include faster spinning disks, quicker servo seek and head settle times, and better track-following systems that use positioning information on the disk [14].[1] However, the bread and butter of technological advancement in disk drives is increasing areal density [8].

Kryder's Law [20] states that areal density of magnetic disks will double every year, a rate of increase which puts Moore's law to shame. Areal density is defined as the product of a disk's *linear density* in bits per inch per track (BPI), and the disk's *track density* in tracks per inch (TPI). Since the outer tracks of a disk have more linear space, manufacturers pack more sectors into the longer outer tracks than the shorter inner tracks, a data layout technique called zoning [19]. Zoning schemes increase the capacity of the disk and, for a given disk rotation speed, allows for faster maximum transfer speeds. On the other hand, if bits are packed too closely together, it causes interference.

One of the most crucial components that determines the possible proximity of sectors on a disk is the capability of the disk head to read and write a fine-grained area. Disk head accuracy has improved over time with lower electrical resistance and tinier head sizes in the tens of nanometers [18]. Modern disk heads are mass produced with thin film and photolithographic processes [5], much like with CPUs and other integrated circuits. As with CPUs, disk heads have process variation—they operate at different signal-to-noise ratios, depending on the manufactured widths of the read sensor and write pole tip.

In the past, the linear density of bits varied only according to different zones, and bit densities were conservatively drawn so that most disk heads could read data error-free. The classic approach predefines

[1] For power savings, some disks sacrifice performance and run at variable spindle speeds [9]. Other hybrid disks use expensive non-volatile memory (such as MEMS [17] or flash) to complement the cheaper magnetic platters. However, we do not include either of these types of disks in our analysis, focusing instead on the more common consumer and enterprise magnetic drives.



(a) 2002 model disks (b) 2006 model disks (c) 2008 model disks

Figure 2: **Evolution of disk behavior over time.** Comparing 128 MB block read bandwidth for a representative sample of 3 sets of identical-model disks, from 2002, 2006, and 2008, demonstrates both a large increase in capacity and a growing trend of heterogeneity within each set of disks. Results for disks within a set are plotted atop one another.

the zones for a particular disk model before manufacturing. To deal with process variation, a trade-off is made between the aggressiveness of the predefined density and the number of disk heads that are discarded because they can't meet the full operating requirements.

To reduce costs and improve component utilization in the face of increasing process variation, new manufacturing techniques determine the capability of a disk head post-production and use that information to optimize the sector layout on the platter surface. Referring to Figure 1, the same target densities can be achieved in many ways, by varying the number of sectors per track and the number of tracks per disk. However, since bandwidth for a fixed rotational speed depends only on the linear density of sectors per track, some disk heads and platter combinations will transfer data faster than others. The practice is now common across the major disk storage vendors, although each vendor has a different name for it and has additional trade secrets for implementing it. For simplicity, we refer to the general practice of adjusting densities according to the capabilities of the particular disk surface and head combination as *adaptive zoning*.

Unlike other new technologies in disk drives, manufacturers have been mostly silent about their use of adaptive zoning. Because of the secrecy surrounding each vendor's approach, very little has been published about it, even though these practices have been going on for a number of years. A Hitachi technical brief is the only documentation that we found [6], where it is referred to as *adaptive formatting*. The best references available for current practices are patent applications, where we have found evidence of this going on at all the largest disk manufacturers, including Toshiba/Fujitsu [12][2], Hitachi/IBM [4], Samsung [21], Seagate [10], and Western Digital [3]. While patents alone don't necessarily mean that the technology has been incorporated into actual products, we have also confirmed that this is happening with sources at these vendors who wish to remain anonymous. Furthermore, measurements of adaptive zoning on modern disks is presented in the next section, confirming high variability of transfer speeds within each individual disk and across a cluster of identical model disks.

Disk drive manufacturers have already solved many issues surrounding adaptive zoning (e.g., how to hide different capacities of surfaces within a drive), however, the focus of this paper is on the visible effects of adaptive zoning to the overarching system. Foremost among these effects is that the same range of block addresses will transfer at variable rates on different disk drives of the same make and model. Traditionally, the same logical address would map to equivalent disk surfaces and approximate locations on different disks, and two adjacent data blocks would transfer at the same rate unless they crossed over a regular zone

[2]Toshiba purchased Fujitsu's disk business in October 2009, and Hitachi purchased IBM's disk business in January 2003.

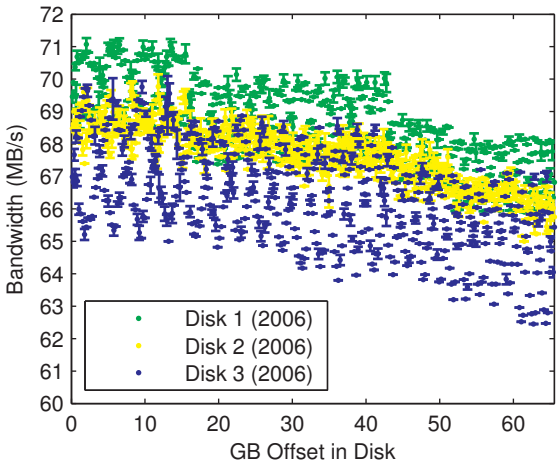


Figure 3: **A closer look at inter-disk behavior with adaptive zoning.** The first 64 GB of the 2006-era disks show that the same logical blocks have consistently different bandwidths across disks (error bars are shown with standard deviation).

boundary. That is no longer the case, as some disks may be able to transfer data at higher speeds than others, and adjacent data blocks that cross over surface boundaries may also exhibit a sawtooth bandwidth pattern of increased variability.

3 Measuring Changes to Disks

The effects of modern disk manufacturing techniques can be seen through bandwidth measurements. Our measurements happen to come from disk drives manufactured by Seagate and Western Digital, but these results and trends are applicable to all the major disk storage companies. The oldest set of drives measured consists of nine Cheetah 10K.6 SAS drives from 2002, each of 36 GB capacity. The next set of drives consists of nine Barracuda 7200.9 SATA drives from 2006, each of 250 GB capacity. The next set consists of 25 Barracuda ES.2 SATA drives from 2008, each of 1 TB capacity. The last set consists of 25 WD RE3 SATA drives from 2009, each of 1 TB capacity.

The evolution of disk behavior is illustrated in Figure 2, revealing a trend of both increasing capacity and heterogeneity over time. This figure plots the results of 128 MB block reads from the raw device for a representative sample of the first three sets of identical-model disks (from 2002, 2006, and 2008). When running 10 trials per disk, where each trial makes a full sweep through the disk, each 128 MB byte range usually obtains similar bandwidth across trials with a standard deviation less than 1 MB/s (error bars not shown). The downward-trending staircase of bandwidth for all the disks is expected because of zoned recording. However, a comparison of these disks from different years shows increasingly varying behavior. The oldest disks, from 2002, all produce roughly the same bandwidth for every block, creating the appearance of one line when there are actually three plotted atop one another. The two more modern drives in (b) and (c) are faster and hold greater capacity, but they also exhibit a range of performance variation within and across disks.

Figure 3 zooms in on the first 64 GB of three representative 2006-era disks, to see the relative performance across disks at a finer granularity. Each disk consistently operates between a different range of throughputs, and the same blocks (i.e., the same logical addresses) achieve different bandwidths across disks. Aliasing effects are present because the 128 B block size always spans more than one surface, creating the

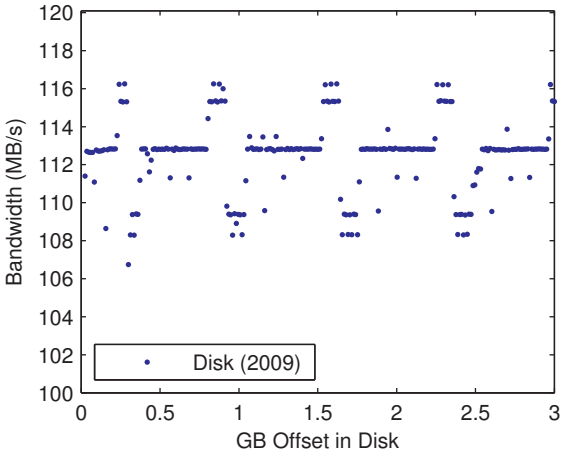


Figure 4: **A closer look at intra-disk behavior with adaptive zoning.** A smaller block size of 12 MB for the 2009-era disks clearly distinguishes between bandwidth differences across disk heads and surfaces.

appearance of diamond-like patterns.

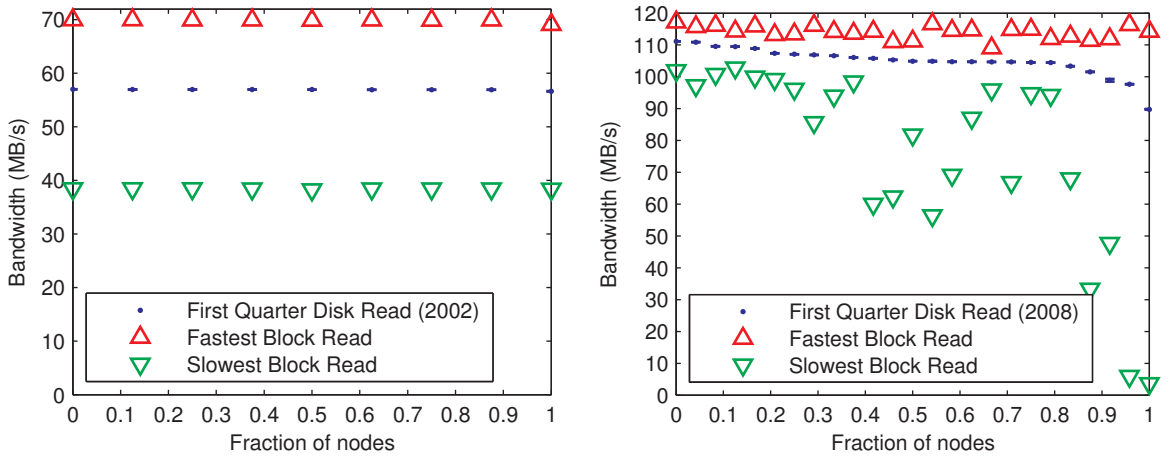
To see the effects of adaptive zoning with less intra-disk aliasing, Figure 4 plots results for the streaming read benchmark on the 2009-era disks using 12 MB blocks, instead of 128 MB, for just the first 3 GB. This is a 3-platter disk, so there are 6 heads. The drive switches heads approximately every 120 MB, so the pattern of switching platters is more visible; the fastest of these heads is capable of 116 MB/s, the slowest head 109 MB/s, and four heads can achieve 113 MB/s. The densities of each head appear to be quantized by the manufacturer. While not shown in this figure, among a sample of 25 of these drives, the average drive performance was fairly similar. However, this is likely because it is a more expensive enterprise-class drive, where consistent performance is a more important QoS metric, and it was built with only top-performing disk heads.

To further illustrate the cross-node performance statistics, Figure 5 provides the average bandwidth for the first quarter of each 2002-era drive (9 GB) and the first quarter of each 2008-era drive (250 GB). The first quarter of the drive provides a large enough sample to compare total performance across nodes, and it also tends to cross over just a couple traditional zoned recording regions (3 zones for both disk types, in this case), so the effects of larger performance variations isn't obscured by zoned recording. As expected, each of the 2002-era disks perform at the same average bandwidth, 67.8 MB/s with a 0.2 MB/s standard deviation. The behavior of the 2008-era drives is much more interesting: the streaming read benchmark performs on average at 105.0 MB/s across disks with a 4.4 MB/s standard deviation. The actual distribution of disk averages falls into a 21 MB/s range up to 14.5% slower than the mean or up to 5.8% faster. Furthermore, the fastest and slowest average block bandwidths during the benchmark reveal great variation for individual (128 MB) byte ranges. A few outliers appear to have some areas of very poor average block bandwidth, possibly due to other defects, although running a SMART disk test completed successfully without any failures.

4 Implications for System Design

There are many implications of adaptive zoning schemes on the design of systems that depend on fast and consistent storage access times.

Homogeneous disk-based clusters no longer exist: The linear and track densities of each surface in



(a) 2002 model disks

(b) 2008 model disks

Figure 5: **Disk drive performance statistics.** Reading the first quarter (9 GB) of an older 2002-era drive that doesn't implement adaptive zoning shows identical behavior across 9 disks. However, reading the first quarter (250 GB) of a modern 2008-era drive with adaptive zoning produces a 21 MB/s spread of average bandwidth, and the minimum and maximum values for each 128 MB block also have considerably more variation across otherwise identical-model disks.

a cluster vary according to the capabilities of its manufactured parts. Variations in disk performance are not indicitive of a fault [2], but are instead to be expected.

Equal work partitioning schemes are inefficient: Dynamic scheduling of tasks (e.g. as in [1]) is even more important for good overall utilization, even in tightly controlled environments.

Striping in disk arrays wastes bandwidth: Instead of achieving the sum of the disk bandwidths, striped disk transfer raquests will instead receive N times the bandwidth of the slowest disk.

Spindle synchronization is useless for RAID arrays: Spindle synchronization is a way to make the position times, including seek and rotational delay, for all N disks be equal. However, since sectors will not be located in the same place across disks, it can't work.

Techniques that require low level disk layouts are harder: Techniques like traxtents [15], or Atropos [16], which rely on the details of track layout, will have to measure each disk individually. Accurately modeling disk performance [14] also becomes more difficult.

Accurate experiments are even harder to achieve: Which disk you happen to get can be added to the long list of things, like your user name [11], that can impact the validity of your experiments.

5 Summary

Recent changes in the fundamental performance characteristics of disk drives, caused by the modern practice of adaptive zoning, make homogeneous sets of disks a thing of the past. Disk performance over the same logical byte range now varies by 20% or more across different disks of equivalent make and model, while blocks within a disk but accessed with different disk heads and surface densities experience similar variations. When building distributed systems with storage that exhibit these new performance characteristics, it is important to recognize that limitations of the storage system may be to blame for distributed systems that perform inefficiently. From here on forward, performance-sensitive disk-dependent systems have no choice but to use more dynamic and sophisticated methods for balancing work.

References

[1] Remzi H. Arpaci-Dusseau, Eric Anderson, Noah Treuhaft, David E. Culler, Joseph M. Hellerstein, David Patterson, and Kathy Yelick, *Cluster I/O with River: making the fast case common*, Workshop on I/O in parallel and distributed systems, 1999.

[2] Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau, *Fail-stutter fault tolerance*, Proceedings of the Eighth Workshop on Hot Topics in Operating Systems (Washington, DC, USA), IEEE Computer Society, 2001.

[3] Raffi Codilian, William D. Johns, Charles A. Park, David D. Nguyen, and Jack M Chue, *Method of manufacturing and disk drive produced by measuring the read and write widths and varying the track pitch in the servo-writer.*, U.S. Patent 6,885,514, April 2005.

[4] Steven R. Hetzler, Prakash Kasiraj, and Richard M. H. New, *Method for adaptive formatting and track traversal in data storage devices.*, U.S. Patent 6,137,644, October 2000.

[5] Hitachi Global Storage Technologies, *Recording head/advanced recording head processing*, https://www1.hitachigst.com/hdd/research/recording_head/ headprocessing/index.html.

[6] Rocky Laroia and Rich Condon, *Adaptive formatting in hitachi drives*, Hitachi Technical Note (2003).

[7] W. B. Ligon, III and R. B. Ross, *Implementation and performance of a parallel file system for high performance distributed applications*, Proceedings of the 5th IEEE International Symposium on High Performance Distributed Computing, August 1996.

[8] Paul Massiglia, *Digital large system mass storage handbook*, Chapter 2.

[9] Chris Mellor, *Western Digital launches power-efficient disk drives*, November 2007, http://news.techworld.com/green-it/10711/western-digital-launches-power-efficient-disk-drives.

[10] Forrest C. Meyer and Tong Shi, *Method and apparatus for utilizing variable tracks per inch to reduce bits per inch for a head.*, U.S. Patent 7,046,471, May 2006.

[11] Todd Mytkowicz, Amer Diwan, Matthias Hauswirth, and Peter F. Sweeney, *Producing wrong data without doing anything obviously wrong*, In Proc. of Intl Conf. on Architectural Support for Programming Languages and Operating Systems, ACM, 2009, pp. 265–276.

[12] Yoshihiko Nakamura and Takeshi Hara, *Disc device, disk formatting method, and disk formatting apparatus.*, U.S. Patent 7,355,809, April 2008.

[13] David A. Patterson, Garth Gibson, and Randy H. Katz, *A case for redundant arrays of inexpensive disks (RAID)*, ACM SIGMOD International Conference on Management of Data, 1988.

[14] Chris Ruemmler and John Wilkes, *An introduction to disk drive modeling*, IEEE Computer **27** (1994), 17–28.

[15] Jiri Schindler, John Linwood Griffin, Christopher R. Lumb, and Gregory R. Ganger, *Track-aligned Extents: Matching Access Patterns to Disk Drive Characteristics*, USENIX Symposium on File and Storage Technologies, 2002.

[16] Jiri Schindler, Steven W. Schlosser, Minglong Shao, Anastassia Ailamaki, and Gregory R. Ganger, *Atropos: A disk array volume manager for orchestrated use of disks*, Proceedings of the 3rd USENIX Conference on File and Storage Technologies, 2004.

[17] Steven W. Schlosser, John Linwood Griffin, David F. Nagle, and Gregory R. Ganger, *Designing computer systems with mems-based storage*, In Proceedings of the 9th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS, 2000.

[18] Tony Smith, *Hitachi halves hard drive head size*, 2007, http://www.reghardware.com/2007/10/15/ hitachi_hdd_head_size_breakthrough.

[19] Rodney Van Meter, *Observing the effects of multi-zone disks*, Proceedings of the annual conference on USENIX Annual Technical Conference, 1997.

[20] Chip Walter, *Kryder's law*, Scientific American (2005), http://www.scientificamerican.com/article.cfm?id=kryders-law.

[21] Jong Yun Yun, Hae Jung Lee, Ken Bovatsek, and Kieu Lien Dang, *Flexible bpi and tpi selection in disk drives.*, U.S. Patent 6,956,710, October 2005.