13th Workshop on Hot Topics in Operating Systems (HotOS 2011), Napa Valley, CA. May 2011.

Disks Are Like Snowflakes: No Two Are Alike

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

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.

1 Introduction

Many multi-component systems are designed and built assuming uniformity of performance. People buy identical hardware, use the same configuration, and expect to achieve similar performance. 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 for the same corresponding physical regions across disks, 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," groups of co-located tracks that allow for more sectors on the longer, outer rings [19]. Until recently, zone arrangements (i.e., sectors per track, tracks per zone) were the same for every surface of every disk of a given model. Now, modern disks each have a unique layout of surface density. 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, characteristics, and implications of this new non-uniformity of disk drives. 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 linear 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 more varied and non-homogeneous bandwidths on which this paper focuses, since disk bandwidth is directly proportional to per-track linear storage density.

Despite relative quiet regarding this new feature, we have found evidence of adaptive zoning being used by all major disk manufacturers, coming from patent applications, third-party performance measurements, and 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 5.9% 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 (by LBN) on 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 inefficient. For example, RAID systems [14] and high-performance parallel file systems that stripe data across many disks [8] may operate at the speed of the slowest disk. We first perceived this issue of disk non-uniformity while observing the delays of slower disks on otherwise identical nodes configured for a prototype parallel dataflow system [6]. In general, any system that depends on 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 [12]. 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.

2 Advances in Disk Technology

Magnetic disk drives have come a long way since their 1950s debut, regularly being refined while maintaining the same basic design mechanisms: rotating platters coated with magnetic material and coupled with moveable heads that induce a magnetic field to read and write data. 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 [15].

The bread and butter of disk drive advancement is increasing areal density [9], which has been achieved at tremendous rates. In 2005, Kryder's Law [20] stated that the areal density of magnetic disks was doubling every year, a rate of increase that put Moore's law to shame. Areal density is 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). For a given disk assembly and data encoding technology, if bits are packed too closely together, then the magnetic signal quality can suffer from interference. Because the outer tracks of a disk have more linear space, manufacturers increase average linear density by fitting more sectors onto them than the shorter, inner tracks, a data layout technique called zoning [19]. Most disks also spin at fixed rates, typically 5400, 7200, 10K, or 15K RPM.[1] Zoning schemes increase the capacity of the disk and, for a given disk rotation speed, allow for faster maximum transfer speeds.

One of the most crucial factors that determines areal density is the capability of the disk head to read and write a fine-grained area. The accuracy of disk head components has improved over time with lower electrical resistance and tinier head sizes in the tens of nanometers [18]. Modern disk head components are mass-produced with thin film and photolithographic processes [5]. As with CPUs and other integrated circuits, 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 set conservatively so that most disk heads could process data error-free. The classic approach predefines the zones for a particular disk model before manufacturing. To deal with process variation, a trade-off was made between the aggressiveness of the predefined density and the number of disk heads discarded because they didn't meet the full operating requirements.

[1]For power savings, some disks sacrifice performance and run at variable spindle speeds [10]. However, we focus instead on the more common consumer and enterprise magnetic drives.

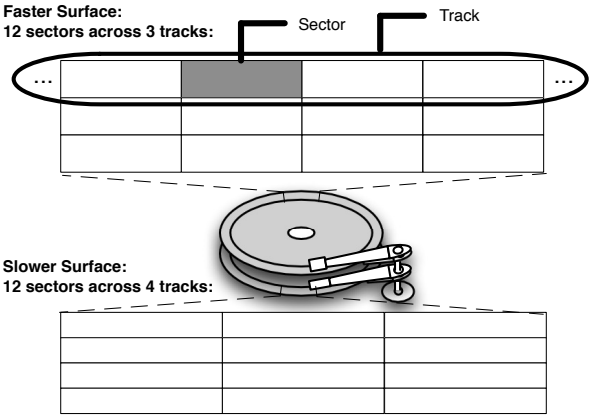
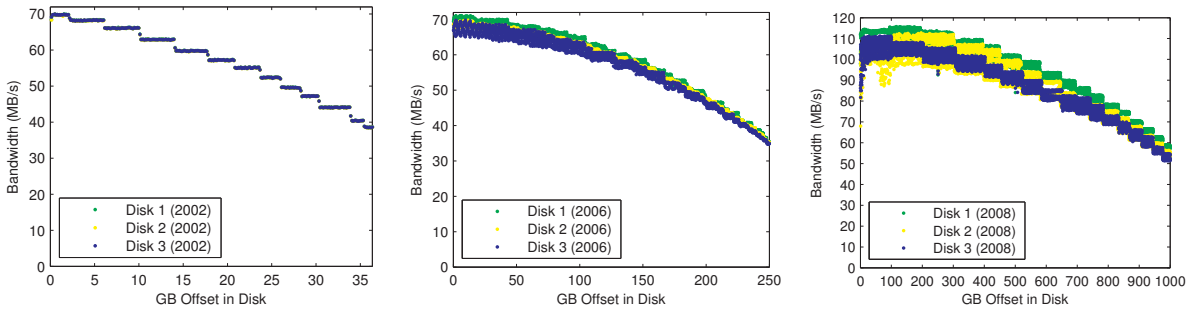


Figure 1: Adaptive zoning of disk drives. The same area of different surfaces within a drive is formatted according to the physical properties of the disk head. In this example, each surface has equivalent areal density, although the top surface is faster because its disk head accommodates more sectors per track over fewer tracks, while the bottom surface's disk head requires fewer sectors per track but allows for more tracks.

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 or 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 head and platter combinations will transfer data faster than others. We refer to the general practice of adjusting densities according to the capabilities of the particular disk surface and head combination as *adaptive zoning*. This practice is now common across the major disk storage vendors, although each vendor may refer to it with a different name or implement it with additional trade secrets.

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 [7] is the only documentation that we found, where it is referred to as *adaptive formatting*. The best references available for current practices are patent applications, where we have found evidence of adaptive zoning at all the largest disk manufacturers, including Toshiba/Fujitsu [13][2], Hi-

[2]Hitachi purchased IBM's disk business in January 2003, Toshiba purchased Fujitsu's disk business in October 2009, and WD announced an agreement to purchase Hitachi's disk business in March 2011.



(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 three 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.

tachi/IBM [4], Samsung [21], Seagate [11], 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 are 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). The focus of this paper is on the visible effects of adaptive zoning on the overlying system. Foremost among these effects is that the same range of block addresses will transfer at different 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 boundary. That is no longer the case.

3 Measuring Changes to Disks

The effects of modern disk manufacturing techniques can be seen through bandwidth measurements. Our disk drives are manufactured by Seagate and Western Digital, but the results and trends are applicable to all the major disk storage companies. The oldest set of measured drives consists of nine Seagate Cheetah 10K.6 SAS drives from 2002, each of 36 GB capacity. The next set of drives consists of nine Seagate Barracuda 7200.9 SATA drives from 2006, each of 250 GB capacity. The third set consists of twenty-five Barracuda ES.2 SATA drives from 2008, each of 1 TB capacity. The last set consists of twenty-five 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 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 2002, 2006, and 2008 sets of identical-model disks. 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 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 at the same address, creating the appearance of one line when there are actually three plotted atop one another. The more modern drives in (b) and (c) are faster with larger capacity, but they also exhibit a range of performance differences 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 MB block size always spans more than one surface.

To see the effects of adaptive zoning with less intra-disk aliasing, Figure 4 plots the first 3 GB of a streaming read benchmark on the 2009-era disks using smaller 12 MB blocks, so the pattern of switching heads is more visible. This is a three-platter disk (i.e., six heads), and the drive switches heads approximately every 120 MB; the fastest of these heads is capable of 116 MB/s, the slowest 109 MB/s, and four heads can achieve 113 MB/s. The densities of each head/platter combination appear to

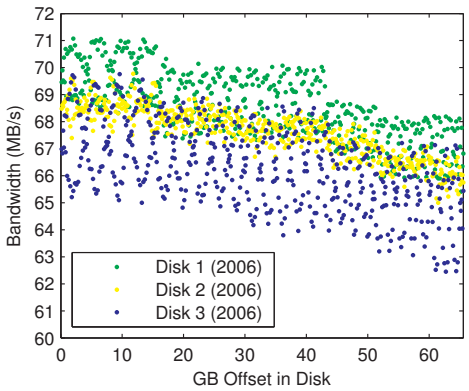


Figure 3: A closer look at inter-disk behavior with adaptive zoning. The first 64 GB of the 2006-era disks show that the same logical blocks have consistently different bandwidths across disks.

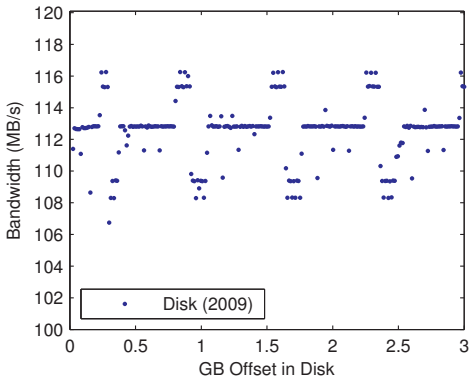


Figure 4: A closer look at intra-disk behavior with adaptive zoning. A smaller block size of 12 MB for the 2009-era disks clearly distinguishes between bandwidth differences across disk heads and surfaces.

be quantized by the manufacturer. While not shown in this figure, among a sample of twenty-five drives, average performance was similar. However, this is likely because these are more expensive enterprise-class drives, built with top-performing disk heads to meet stricter performance guarantees.

To further illustrate cross-node performance statistics, Figure 5 provides the average bandwidths 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 (three zones for both disk types, in this case), so the effects of larger performance variations aren'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 2008-era drives, on the other hand, perform a streaming read benchmark at an average of 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 that varies as much as 5.9% faster or 14.5% slower than the mean. Furthermore, the fastest and slowest average block read bandwidths during the benchmark reveal considerable differences for individual 128 MB blocks. A few disks also have some areas of very poor average block bandwidths, possibly due to other defects or bad sector remappings, although a SMART disk test didn't find any issues.

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

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

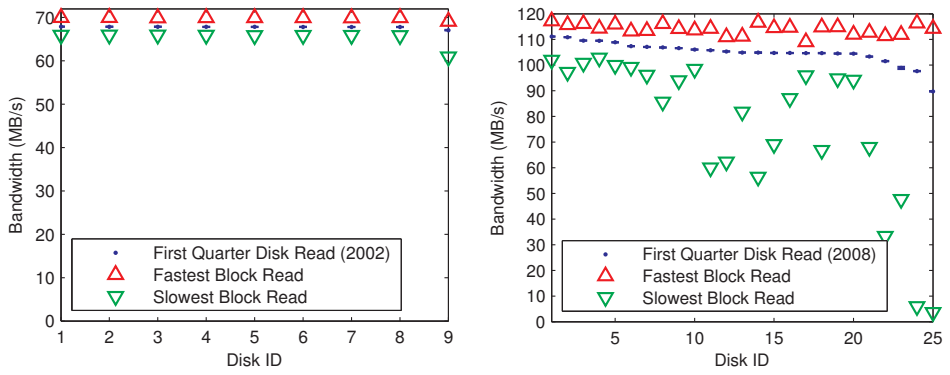
Striping in disk arrays wastes bandwidth: Instead of achieving the sum of the disk bandwidths for larger transfers, striped disk transfer requests will receive N times the bandwidth of the slowest disk.

Spindle synchronization is useless for RAID arrays: Spindle synchronization attempts to make the positioning 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 more difficult: Techniques like traxtents [16] or Atropos [17], which rely on the details of track layouts, will have to measure each disk individually. Correct modeling of disk performance [15] 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, as extreme as your user name [12], that can impact the validity of your experiments.

More solutions are required to manage disk heterogeneity: Some variability across disks (e.g., from unsynchronized spindles) has always been an issue, but could be mitigated through larger per-disk transfers, deeper queues, or more asynchronous I/O. Now that performance differences are persistent over ranges of 10s or 100s of megabytes, simply extending the same ideas may not address the problem.



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

Figure 5: Disk performance statistics. Reading the first quarter (9 GB) of 2002-era drives has similar behavior across disks. However, reading the first quarter (250 GB) of modern 2008-era drives with adaptive zoning shows a large 21 MB/s spread of average bandwidths across otherwise identical-model disks (with more differences for the fastest and slowest average block reads).

5 Summary

Changes in the fundamental performance characteristics of disk drives, caused by 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 disks of equivalent make and model, and blocks on different surfaces within a disk experience similar differences. When building distributed systems with storage that exhibits these new characteristics, inherent heterogeneity of the storage system may be to blame for inefficiencies. From now on, performance-sensitive disk-dependent systems must use more dynamic and sophisticated methods to balance work.

Acknowledgements

We thank Michael Stroucken and Mitch Franzos for assistance in configuring hardware, and Raja Sambasivan, Matthew Wachs, Jay Wylie, Garth Gibson, David Andersen, the anonymous vendor sources, the reviewers, and the members and companies of the PDL Consortium for their feedback and support. This research was sponsored in part by an HP Innovation Research Award, by CyLab at Carnegie Mellon University under grant #DAAD19–02–1–0389 from the Army Research Office, and by an NDSEG Fellowship from the Department of Defense.

References

[1] R. H. Arpaci-Dusseau, et al. Cluster I/O with River: Making the Fast Case Common. IOPADS, 1999.

[2] R. H. Arpaci-Dusseau and A. C. Arpaci-Dusseau. Fail-Stutter Fault Tolerance. HotOS, 2001.

[3] R. Codilian, et al. Method of manufacturing and disk drive produced by measuring the read and write widths and varying the track pitch in the servo-writer, 2005. U.S. Patent 6,885,514.

[4] S. R. Hetzler, et al. Method for adaptive formatting and track traversal in data storage devices, 2000. U.S. Patent 6,137,644.

[5] Hitachi Global Storage Technologies. Recording Head Processing. https://www1.hitachigst.com/hdd/research/recording_head/headprocessing/index.html.

[6] E. Krevat, et al. *Applying Simple Performance Models to Understand Inefficiencies in Data-Intensive Computing*. Technical report. 2011. CMU-PDL-11-103.

[7] R. Laroia and R. Condon. Adaptive Formatting in Hitachi Drives. *Hitachi Technical Note*, 2003.

[8] W. B. Ligon, III and R. B. Ross. Implementation and Performance of a Parallel File System for High Performance Distributed Applications. HPDC, 1996.

[9] P. Massiglia. Digital Large System Mass Storage Handbook, 1986. Chapter 2.

[10] C. Mellor. Western Digital launches power-efficient disk drives. http://news.techworld.com/green-it/10711/western-digital-launches-power-efficient-disk-drives.

[11] F. C. Meyer and T. Shi. Method and apparatus for utilizing variable tracks per inch to reduce bits per inch for a head, 2006. U.S. Patent 7,046,471.

[12] T. Mytkowicz, et al. Producing Wrong Data Without Doing Anything Obviously Wrong. ASPLOS, 2009.

[13] Y. Nakamura and T. Hara. Disc device, disk formatting method, and disk formatting apparatus, 2008. U.S. Patent 7,355,809.

[14] D. A. Patterson, et al. A Case for Redundant Arrays of Inexpensive Disks (RAID). SIGMOD, 1988.

[15] C. Ruemmler and J. Wilkes. An introduction to disk drive modeling. *IEEE Computer*, **27**:17–28, 1994.

[16] J. Schindler, et al. Track-aligned Extents: Matching Access Patterns to Disk Drive Characteristics. FAST, 2002.

[17] J. Schindler, et al. Atropos: A Disk Array Volume Manager for Orchestrated Use of Disks. FAST, 2004.

[18] T. Smith. Hitachi halves HD head size. http://www.reghardware.com/2007/10/15/hitachi_hdd_head_size_breakthrough.

[19] R. Van Meter. Observing the Effects of Multi-Zone Disks. USENIX ATC, 1997.

[20] C. Walter. Kryder's Law. *Scientific American*, 2005.

[21] J. Y. Yun, et al. Flexible BPI and TPI selection in disk drives, 2005. U.S. Patent 6,956,710.