

Disk Drive Roadmap from the Thermal Perspective: A Case for Dynamic Thermal Management

Sudhanva Gurumurthi Anand Sivasubramaniam Vivek K. Natarajan

Department of Computer Science and Engineering
The Pennsylvania State University, University Park, PA 16802
{gurumurt,anand,vnataraj}@cse.psu.edu

Abstract

The importance of pushing the performance envelope of disk drives continues to grow, not just in the server market but also in numerous consumer electronics products. One of the most fundamental factors impacting disk drive design is the heat dissipation and its effect on drive reliability, since high temperatures can cause off-track errors, or even head crashes. Until now, drive manufacturers have continued to meet the 40% annual growth target of the internal data rates (IDR) by increasing RPMs, and shrinking platter sizes, both of which have counter-acting effects on the heat dissipation within a drive. As this paper will show, we are getting to a point where it is becoming very difficult to stay on this roadmap.

This paper presents an integrated disk drive model that captures the close relationships between capacity, performance and thermal characteristics over time. Using this model, we quantify the drop off in IDR growth rates over the next decade if we are to adhere to the thermal envelope of drive design. We present two mechanisms for buying back some of this IDR loss with Dynamic Thermal Management (DTM). The first DTM technique exploits any available thermal slack, between what the drive was intended to support and the currently lower operating temperature, to ramp up the RPM. The second DTM technique assumes that the drive is only designed for average case behavior, thus allowing higher RPMs than the thermal envelope, and employs dynamic throttling of disk drive activities to remain within this envelope.

Keywords: Disk Drives, Thermal Management, Technology Scaling.

1. Introduction

The importance of I/O performance, disk drives in particular, continues to grow over the years, with the widen-

ing disparity in speeds between semiconductor and electro-mechanical components. There are numerous applications in both the commercial and scientific domains that are critically dependent on I/O performance. Data-centric services such as file and e-mail servers and transaction processing within an enterprise, together with Internet based services such as Google, stock trading, etc., rely heavily on disks for their storage needs. Even in the scientific domain, the working sets of applications continue to out-pace memory system capacities (requiring out-of-core I/O support), and data sets are being shared across several applications (requiring explicit file I/O), often putting the I/O subsystem under the spotlight in limiting their scalability.

There have been several improvements over the years to address the I/O bottleneck, including better caching/buffer management [35], parallelism in the form of RAID [34], and high bandwidth interconnects such as SAN. However, at the core of these extensive I/O subsystems lie the disk drives, whose performance advances have woefully lagged behind the rest of the system components. As Amdahl's law dictates, a single such laggard can eventually limit overall system performance. Further, it is quite possible that a faster drive can actually alleviate the need for going to expensive storage area networks and higher levels of parallelism when deploying balanced systems. This becomes particularly important for their usage in home gaming, DVRs, and other low cost markets. Consequently, it is always important to understand and explore how much we can push the envelope of disk drive performance.

Disk drive vendors express performance in terms of Internal Data Rates (IDR) and seek times. When we examine the mechanics of drives, it may intuitively appear that we can easily push the envelope by spinning disks faster, making the platters smaller (to reduce seek distances), and avail of better recording technologies whenever possible. However, one of the most fundamental factors affecting disk drive design is the heat generated by some of these actions and its effect on reliable operation. High temperatures can cause off-track errors due to thermal tilt of the disk stack and actuator, or even cause head crashes due to the out-gassing of spindle and voice-coil motor lubricants [20]. Even a fifteen degree Celsius rise from the ambient temperature can double the failure rate of a disk drive [2], and it is important to provision some design margin within the disk to accommodate slight variations in the external temperature. This makes it imperative for the disk to operate below a thermal envelope.

The disk drive industry has tried to chart out a 40% IDR annual growth rate, and has used different techniques until now towards meeting it, while remaining within the ther-

mal envelope. Increasing the RPM alone to meet the target IDR can sometimes push the drive beyond the envelope since heat dissipation is proportional to nearly the cubic power (2.8-th power to be precise) of RPM. In conjunction with such RPM increases, the trick is to decrease the platter size, since that affects the heat dissipation in nearly the fifth power (4.6-th power to be precise). When doing so, one needs to rely on the increases in recording densities or add more platters to ensure the capacity does not diminish.

While these techniques have been successful in improving IDR until now, without exceeding the thermal envelope, this is getting to be more difficult because (i) further density improvements are not as easy since they require high error/noise tolerance and very complex head designs, (ii) stronger error-correcting codes to tolerate errors can consume a significant portion of disk capacity and reduce the effective user data rate, (iii) going to smaller enclosures decreases the amount of heat that can be drained to the outside air and (iv) external ambient temperatures are becoming more difficult to contain since the heat dissipated by other system components is increasing, and extensive cooling systems are very expensive. In order to continue to make innovations, it is important to understand all these trade-offs and how they impact the disk drive roadmap over the next decade, which is an important goal of this paper.

Until now, industry has been rather conservative in producing designs that assume worst-case operating conditions from the thermal perspective. As our results will show, the thermal envelope puts a severe restriction on what IDR growth (much less than 40%) we can get in the future if we continue along the same philosophy. Rather than under-provision the drive for performance at manufacturing time (which restricts the benefit to applications) assuming the worst-case scenarios for temperature, we point out that one may want to design the drives for more average-case behavior. This can let applications push the envelope of performance, while still dynamically controlling the disk activities (seeks and rotational speeds) in order to ensure that it operates within the thermal envelope. In fact, similar techniques are being increasingly suggested [4, 41] in the design and usage of semiconductor components as well, where thermal issues are starting to seriously restrict design. In addition to allowing applications to benefit from higher performance on disks designed with average-case temperature behavior, Dynamic Thermal Management (DTM) techniques can also be used on today's disks (that are more restrictive in performance) to further lower their operating temperatures. As earlier studies have shown [2], operating at lower temperatures can enhance long term drive reliability.

Understanding and studying all these issues mandates comprehensive models of disk drives. Most of the previous published work has been mainly on performance models [12], with a few isolated studies on modeling temperature within a drive [11]. Rather than study these issues in isolation, for the proposed research we need to understand (i) how different drive parameters impact the three pronged metrics of interest, namely, capacity, performance and heat, (ii) how these metrics are inter-related, and (iii) how these interactions vary over time. To our knowledge, this is the first paper to do so, and it makes the following contributions towards the important goal of temperature-aware disk drive design:

- We present models to study the capacity, performance and thermal characteristics of disk drives in a unified way to evaluate the different trade-offs over time. We validate the values given by these models with data from thirteen real drives from four different manufacturers over four calendar years.
- We compute a disk drive roadmap based on our models and expected trends in technology scaling of the fundamental parameters. We show that while we have been able to meet the IDR growth rates until now, the thermal design envelope severely restricts data rates and capacity growth in the future. This problem is further accentuated with higher costs of cooling the external ambient air.
- In order to buy back some of the performance/capacity loss imposed by the thermal envelope, we present Dynamic Thermal Management (DTM) of disks as an option. Specifically, we present two ways of performing DTM: (i) ramping up rotational speed when detecting a thermal slack between current temperature and what the disk was designed for, and (ii) designing a disk for the average case and dynamically employing throttling to control disk activities when reaching close to the thermal envelope. Though the availability of multi-speed disks can allow more scope for DTM, and there are disks [24] available today which allow dynamic modulation between two RPMs, throttling can be applied even on existing disks. We also show that the data rate gain which one can obtain with DTM can considerably ease the response times for I/O intensive workloads (providing 30-60% improvement in response times with a 10K increase in RPM).

The next section reviews related work. The inter-related models for capacity, performance and thermal characteristics are given in section 3. Section 4 charts out the roadmap when designing disks for a thermal envelope. The DTM possibilities are discussed in section 5. Finally, section 6 summarizes the contributions and identifies research directions for the future.

2. Related Work

The importance of the I/O subsystem on the performance of server applications has resulted in fairly detailed performance models of the storage hierarchy (e.g. [13, 14]). The Disksim simulator [13], that we use in this paper, is one such publicly distributed tool that models the performance characteristics of the disk drive, controllers, caches and interconnects. The importance of modeling power has gained attention over the last decade, primarily in the context of conserving battery energy for drives in laptops [48, 25]. Temperature-aware design is becoming increasingly important [41] as well. In the context of disk drives, [8] describes a model of the thermal behavior of a drive based on parameters such as the dimensions, number of platters in the disk stack, their size, and properties of the constituent materials. We adapt this model to study the thermal ramifications of current and future hard disk drives. There has also been some recent work on modeling and designing disk arrays in a temperature-aware manner [27].

Power management for hard disks is a well-researched area in the context of single-user systems such as laptops and desktops [10, 30, 33]. Here, energy savings are obtained by spinning the disk down (by turning off the spindle-motor that rotates the platters) during periods when no requests come to it. There has also been a study that has proposed replacing a laptop disk with an array of smaller form-factor disks [47]. More recent studies [19, 5, 17] have started to look at disk power issues in the context of servers. Conventional spin-down-based power management techniques are challenging to apply in server systems, due to the relatively smaller durations of the idle periods and also due to the mechanical characteristics of server-class disks [19]. To address this issue, the use of multi-speed disks [17, 5] has

been proposed. There has also been work on cache management for optimizing disk power consumption [49]. In [9], the authors propose replacing a tape backup system with an array of disks, which are kept in a low-power mode as long as possible.

The historical evolution of different aspects of hard-disk drive design along with projections for their future trends are given in a set of papers published by industry [2, 16, 22, 37]. There have also been studies on the characteristics of designing future high density disk-drives. These papers cover issues such as the impact of bit-aspect ratios [7] and error-correcting code overheads [45] in such drives. There are several proposals [45, 31, 43] on how to build Terabit areal density drives, covering magnetic recording physics issues and also engineering considerations.

3. Modeling the Capacity, Performance and Thermal Characteristics of Disk drives

This section describes the models that we use in our study for capturing capacity, data rates and thermal characteristics of disk drives. The metrics that we use are standard ones that are reported by the disk drive manufacturers for their products.

3.1. Modeling the Capacity

The model begins with an abstraction of the fundamental properties of recording technologies via two quantities, namely, the linear bit-density given in Bits-per-Inch (*BPI*) for a track, and the radial track-density, which is expressed in Tracks-per-Inch (*TPI*). *BPI* improvements are a result of technological advances in read/write head design and recording medium materials. *TPI* is improved by advances in the servo design, track misregistration reduction techniques, and more sophisticated heads [3]. The product of *BPI* and *TPI* is known as the *areal density* and is one of the most fundamental determinants of both drive speed and capacity. The ratio $\frac{BPI}{TPI}$ is known as the bit aspect-ratio (*BAR*) and will be used later in this study to set up the technology scaling predictive model. Another metric of interest to disk-drive designers is the Internal Data Rate (*IDR*), which is expressed in MB/s. The *IDR* is the actual speed at which data can be read from or written into the physical media. The *IDR* is affected by the *BPI*, platter size, and disk RPM.

Let us assume that we know the outer radius, r_o , of the disk drive. We set the inner radius to be half that of the outer radius, i.e., $r_i = \frac{r_o}{2}$. Although this rule of thumb was common in the past, modern disks may not necessarily follow this rule [3]. As the exact inner radius tends to be manufacturer specific and even varies across a single manufacturer's products, we still use this rule in our evaluations.

Let n_{surf} denote the number of surfaces in the drive - this value is twice the number of platters. Then, the number of tracks on the disk surface, which is also denoted as the number of *cylinders* in the disk, n_{cylin} , is given by $n_{cylin} = \eta(r_o - r_i)TPI$, where η is the stroke efficiency, which measures the fraction of the total platter surface that is user accessible. If $\eta = 1.0$, then the equation gives the number of tracks that can be laid out in the area between the innermost to the outermost edge of the platter. However, in practice, η is much lesser than 1 since portions of this real estate are dedicated for recalibration tracks, manufacturer reserved tracks, spare tracks (to recover from defects), landing zone for the head slider, and other manufacturing tolerances. An accepted stroke efficiency used by

practitioners is around $\frac{2}{3}$ [28], which is the value that we use. From these, we can calculate the raw capacity (in bits), C_{max} , of the disk drive as

$$C_{max} = \eta \times n_{surf} \times \pi(r_o^2 - r_i^2)(BPI \times TPI)$$

In reality, even this C_{max} is not completely usable because: (i) Outer tracks can hold more sectors due to their longer perimeters. However, allocating storage on a per-track basis would require complex channel electronics to accommodate different data rates for each track [3]. Instead, *Zoned Bit Recording (ZBR)* or *Multi-Band Recording* is used, which can lead to some capacity loss; (ii) in addition to user data, each sector needs additional storage for servo patterns and Error Correcting Codes (ECC) leading to a further reduction in capacity. These components are modeled as follows.

Capacity Adjustment due to Zoned Bit Recording (ZBR): ZBR is a coarse-grained way of accommodating variable sized tracks, where the tracks are grouped into zones, with each track in a zone having the same number of sectors. Such grouping can provide good trade-offs between capacity and complexity of the electronics. ZBR can allow more data to reside on outer tracks to benefit from higher data rate due to a constant angular velocity, without extensively complicating the electronics.

Each track, j , has a raw bit capacity C_{t_j} , which is given by $C_{t_j} = 2\pi r_j BPI$, where r_j is the radius of track j . Let $j = 0, 1, \dots, n_{cylin} - 1$, where 0 is the outermost track and $n_{cylin} - 1$ is the innermost. Then, for any two tracks m and n such that $m < n$, $C_{t_m} > C_{t_n}$ since $r_m > r_n$. Since the recordable area is within $r_o - r_i$, and we have n_{cylin} cylinders that are assumed to be uniformly spaced out, the perimeter of any given track j , denoted as P_{t_j} is given by

$$P_{t_j} = 2\pi[r_i + (\frac{r_o - r_i}{n_{cylin} - 1})(n_{cylin} - j - 1)] \quad (1)$$

Assuming that each zone has an equal number of tracks, the number of tracks per zone, n_{tz} , is $n_{tz} = \frac{n_{cylin}}{n_{zones}}$, where n_{zones} is the desired number of zones, which is around 30 for modern disk-drives. Therefore, zone 0 would be composed of tracks 0 to $\frac{n_{cylin}}{n_{zones}} - 1$, zone 1 would have tracks $\frac{n_{cylin}}{n_{zones}}$ to $(\frac{2n_{cylin}}{n_{zones}} - 1)$ and so on. For each zone z , let the bit capacity of its smallest perimeter track in the zone be denoted as $C_{tz_{min}}$. In our ZBR model, we allocate, for every track in zone z , $C_{tz_{min}}$ bits (or $(\frac{C_{tz_{min}}}{4096})$ sectors). Thus each zone has a capacity of $n_{tz}(\frac{C_{tz_{min}}}{4096})$ sectors, making the total disk capacity (in 512-byte sectors), with losses due to ZBR taken into account, as $C_{ZBR} = n_{surf} \sum_{z=0}^{n_{zones}-1} n_{tz}(\frac{C_{tz_{min}}}{4096})$.

Capacity Adjustments due to Servo Information: Servo are special patterns that are recorded on the platter surface to correctly position the head above the center of a track. In older drives, an entire surface (and head) used to be dedicated for servo information, leading to considerable loss of usable capacity. To mitigate this, modern drives make use of a technique known as *embedded servo*, where the servo patterns are stored along with each sector. There are no special servo surfaces and the read/write heads that are used for user data access are also used to read the servo information.

We model the storage-overheads for servo by considering the number of bits required to encode the track-identifier

information for each servo-sector. Other fields in the servo information such as those for write-recovery (which signals the beginning of a servo pattern) and for generating the Position Error Signal (which indicates the position of the actuator with respect to a particular track) are not modeled due to the lack of information about their implementation in real disk drives. We model the servo based on the information given in [32]. The track-id information is encoded as a Gray Code, such that the fields for any two adjacent tracks differ only by a single bit. This enables fast and accurate seeks to be performed. Thus, the number of bits needed in the code to encode a track on a surface is $\log_2(n_{cylin})$. As the servo information is embedded with each sector, the total number of bits used for storing servo in each sector, C_{servo} is given by

$$C_{servo} = \lceil \log_2(n_{cylin}) \rceil \quad (2)$$

Capacity Adjustments due to Error-Correcting Codes: Each bit cell in a track is composed of multiple magnetic grains (typically 50-100 grains/cell [23]). A bit storing a digital ‘one’ is composed of a region of grains that are uniformly polarized, and a region where there is a transition in the magnetic polarity represents a ‘zero’. When a write is performed on a bit cell, all the grains in the region have their magnetic polarity altered by the write head. To achieve higher areal density, the size of the bit cell needs to be reduced, and this typically involves a reduction in the grain size. However, the superparamagnetic limit [6] imposes a minimum grain size so that the signal energy stored in the grain does not drop below the ambient thermal energy. Otherwise, the magnetic grains would become thermally unstable and would flip their polarity within a short time span (effectively rendering disk drives as volatile storage!). One way to overcome this limit is to use a recording medium that is more coercive, and thus requiring a stronger field to change the state of the bits. Designing write heads to achieve this is quite challenging [28]. Therefore, in order to continue achieving areal density growth beyond this point, it would be necessary to reduce the number of grains per bit as well. The use of fewer grains in the bit cell leads to lower Signal-to-Noise Ratios (SNR). In order to accommodate such noisy conditions, Error-Correcting Codes (ECC) are required, and most modern disks use Reed-Solomon codes [36]. It has been shown that, for current disks, the ECC storage requirement is about 10% of the available capacity and would increase to 35% for disks whose areal densities are in the Terabit range [45]. Thus, in our model, the total capacity used by ECC (in bits), C_{ECC} , is 416 bits/sector for drives whose areal densities are less than 1 Tb/in², whereas those in the terabit range use 1440 bits/sector.

Derated Capacity Equation: From our above discussions on ZBR, Servo and ECC costs, we can calculate their total space overhead (in bits/sector) as

$$\alpha = n_{tz} \left(\frac{C_{tzmin}}{4096} \right) (C_{servo} + C_{ECC})$$

Therefore, the estimated capacity of the disk in terms of 512 byte sectors is given by

$$C_{actual} = n_{surf} \sum_{z=0}^{n_{zones}-1} n_{tz} \left(\frac{n_{tz} C_{tzmin} - \alpha}{4096} \right) \quad (3)$$

Validation: To verify the accuracy of this derived model, we calculated the capacity estimated by our model for a

set of server disks of different configurations, manufacturers, and from different years (obtained from [2]). The numerical validation results are given in [18]. For most disks, the difference between the actual and estimated capacities is within 12%. The errors are primarily due to some of the assumptions made along the way, and also because we assume 30 zones for each disk (which is optimistic for many of the older disks in the table which used only around 10-15 zones).

3.2. Modeling the Performance

As mentioned in Section 2, there have been several earlier studies on modeling and simulation of disk drives. Rather than come up with an entirely new model, our goal here is to merely leverage from prior work in modeling two main performance related drive parameters that we need for our later studies, namely the *seek time* and the *internal data rate*, as explained below.

Seek Time: The seek time depends on two factors, namely, the inertial power of the actuator voice-coil motor and the radial length of the data band on the platter [16]. We use a similar model as [46] which uses three parameters, namely, the track-to-track, full-stroke, and average seek time values, which are usually specified in manufacturer datasheets. The track-to-track seek time is the time taken for the actuator to move to an adjacent track. The full-stroke seek time is the time taken for the actuator to move from one end of the data band to another. It has been observed that, except for very short seeks (less than 10 cylinders), a linear interpolation based on the above three parameters can accurately capture the seek time for a given seek distance [46]. To determine these values for hard disk drives of the future that we will be evaluating later, we used a linear interpolation of data from actual devices of different platter sizes.

Calculating Internal Data Rate (IDR): The maximum IDR would be experienced by tracks in the outermost zone (zone 0) of the disk drive, since there are more bits stored there while the angular velocity is the same across the tracks. Consequently, we can express the maximum IDR (in MB/sec) of the disk as:

$$IDR = \left(\frac{rpm}{60} \right) \left(\frac{n_{tz0} \times 512}{1024 \times 1024} \right) \quad (4)$$

where n_{tz0} is the number of sectors/track in zone 0, and rpm is the angular velocity expressed as rotations-per-minute.

Validation: The seek time models have already been verified in earlier work [46]. To validate the IDR model, we computed the IDR from the specifications for the disks used in the capacity validation and compared it against the manufacturer supplied IDR value. The resulting data is presented in [18]. Again, we assume that each of the disks uses ZBR with 30 zones. In general, we observe that for most of the disks, the IDR predicted by our model and the actual IDR are within 15%.

3.3. Modeling the Thermal Behavior

Even though there are disks in the market today that are equipped with temperature sensors, it is necessary to develop a thermal model for disk drives since we are intending to investigate designs that are yet to appear. Such a model can also help us analyze the heat transfer between different components more closely. The thermal model that we employ is based on the one developed by Clauss and Eibeck [8]. This model evaluates the temperature distribution of the

drive by calculating the amount of heat generated by components such as the spindle motor (SPM) and the voice-coil motor (VCM), the conduction of heat along the solid components and the convection of heat to the air. It is assumed that the drive is completely enclosed and the only interaction with the external air is by the conduction of heat through the base and the cover and convection with the outside air. The outside air is assumed to be maintained at a constant temperature by some cooling mechanism. This is true in most modern systems where air flow is provided, typically using fans, to maintain a constant external temperature [40].

The model divides the hard disk into four components, namely, (i) the internal drive air, (ii) the SPM assembly that consists of the motor hub and the platters, (iii) the base and cover, and (iv) the VCM and the disk arms. The heat transfer rate over a time interval t , $\frac{dQ}{dt}$ (in Watts), through a cross-sectional area A is given by Newton's Law of Cooling as

$$\frac{dQ}{dt} = hA\Delta T$$

where h is the heat-transfer coefficient and ΔT is the temperature difference between the two entities. For solids, where heat is transferred via conduction, the heat transfer coefficient h depends upon the thermal conductivity k and the thickness of the material and is given by $\frac{k}{Thickness}$.

Between solids and fluids, the heat exchange takes place via convection, where the heat transfer coefficient depends on whether the fluid flow is laminar or turbulent, and also on the exact geometry of the solid components. The model makes use of empirical correlations to calculate the heat transfer coefficient of the different solid components of the disk drive. The heat of the internal drive air is calculated as the sum of the heat energy convected to it by each of the solid components and the viscous dissipation (internal friction) in the air itself minus the heat that is lost through the cover to the outside. The viscous dissipation is related linearly to the number of platters in the disk stack, 2.8th power of with the disk RPM and to the 4.6th power of the platter diameter [8, 39].

To solve the heat equations for each component, the model uses the finite difference method [29]. At each time step, the temperatures of all the components and the air are calculated, and this is iteratively revised at each subsequent time step until it converges to a steady state temperature. The air temperature is assumed to be uniform over the entire drive at each time step. The accuracy of the model depends upon the size of the time steps [8]. Using a coarse-grained time step provides a faster model (in terms of computation time), but the solution may not be accurate. On the other hand, an extremely fine-grained time step can provide an accurate solution at the expense of a high computation time. We experimented with a wide range of different sizes and found that a value of 600 steps per minute gave a solution very close to that of the finer-grained ones.

There are a number of input parameters to the thermal model. The first set of parameters relate to the disk geometry, such as the inner and outer radii of a platter, the enclosure form-factor dimensions, and the length of the disk arms. Another set of parameters pertain to the properties of the materials, such as the thermal conductivity and density. There are also operational parameters such as the number of platters, the RPM, the temperature of the outside air and the VCM power.

With regard to the materials, the platters on most current disk drives are typically made of an Aluminum-Magnesium alloy and the base/cover castings are Aluminum [21]. As the exact alloy that is employed tends to be proprietary information, we assumed that the platters, together with the disk

arm and spindle hub, are made of Aluminum. With regard to the operational parameters, we set the external ambient temperature to 28 C, which is the maximum operating wet-bulb temperature. The wet-bulb temperature measures the temperature with the humidity of the air taken into account. Many disks, including some of the thirteen that we have examined, specify a maximum wet-bulb external temperature of 28-29.4 C. When calculating the power of the VCM, which is dependent on platter dimensions, we made use of previously published data [42]. This earlier work shows that the VCM power is roughly twice for a 95 mm (3.7") platter compared to that for a 65 mm (2.5") one, and nearly four times that for the 47 mm (1.8") size.

Validation and Setting a Thermal Envelope The thermal model proposed earlier [11] was validated with disk drives that are over 15 years old. In addition to making sure that the model is still applicable for modern drives, we also need to define what should be the thermal envelope (i.e. the maximum operating temperature) for drive design when charting out our roadmap.

We modeled the Seagate Cheetah 15K.3 ST318453 SCSI disk drive [40] in detail. This disk-drive is composed of a single 2.6" platter (but a 3.5" form-factor enclosure) and rotates at 15K RPM. We took the disk apart and studied its geometry in detail. This allowed us to determine how the components are internally laid out and create geometry models parameterized for the platter-size and count. We also measured the physical drive parameters such as the length of the disk-arm, thickness of the platter, base, and cover etc., which are not considered by the capacity and performance models, using Vernier calipers. The VCM power of this disk is determined to be 3.9 W. The disk specifications (in their data sheets) typically include the maximum operating temperature which is the temperature that should not be exceeded even if the VCM and SPM are always on (note that in actual operation the temperature may be lower). In our validation experiment, we set the SPM and VCM always on, and measured the internal air temperature.

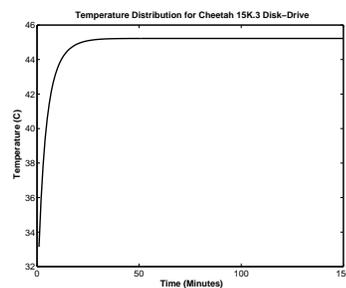


Figure 1. Temperature of the modeled Cheetah ST318453 disk over time starting from an external temperature of 28 C.

The temperature of the internal air over the duration of the experiment is shown in Figure 1. All components are initially at the temperature of the outside air, namely, 28 C. The temperature rises from 28 C to 33 C within the first minute and rapidly increases thereafter. It then stabilizes and reaches a steady state of 45.22 C after about 48 minutes. Note that in the model, we consider only the VCM and SPM heat sources, and do not consider the heat generated by on-board electronics, since our goal here is to focus on the primary drive parameters and their technological

impact over time on the heat dissipation. A detailed model of the electronic components using tools such as HotSpot [41] would be necessary to capture the heat dissipation due to drive electronics, and study their technological trends. Consequently, we discount the heat generated by these electronic components in all our results and reduce the thermal envelope of operation accordingly. In fact, earlier research [27] has shown that on-board electronics can add about 10 C to the temperature within the drive. If we consider this additive factor ($10 + 45.22 = 55.22$ C), the results presented here come very close to the rated maximum operating temperature of this drive (which is 55 C), verifying the validity of this model.

It is to be noted that the thermal envelope - the maximum temperature within a drive for reliable operation - itself has negligible variance over time. Consequently, we use the same *thermal envelope of 45.22 C* (obtained above without on-board electronics) when laying out the roadmap over time across disks of different platter sizes and counts.

4. Roadmap with Thermal Constraints

In the previous section, we developed three drive models for capacity, performance and thermal characteristics, which though explained independently are rather closely intertwined. This is because many of the parameters used by each model can depend on the results from another. For instance, the performance and thermal characteristics are closely dependent on the drive parameters (e.g. size and number of platters). The heat dissipation is closely dependent on the operating RPM. Finally, the capacity of drives is not only limited by recording technologies, but also by the thermal envelope (larger or more platters can lead to higher temperatures). It is important to ensure that we study all the inter-related factors together when charting out a disk drive technology roadmap, which is our intent in this section.

This roadmap is driven by two fundamental factors: (i) the innovations in magnetic technologies to increase recording densities (the BPI and TPI in particular), and (ii) the growing workload demands for high data transfer rates (the IDR). The trends in growth of BPI, TPI and IDR over the past decade have been made available by Hitachi [22] where the values for each year, together with their Compound annual Growth Rate (CGR), are given. For instance, for the year 1999, the values for BPI, TPI, and IDR were 270 KBPI, 20 KTPI, and 47 MB/s, and their CGRs have been 30%, 50%, and 40% respectively. This growth rate in the linear and track densities has resulted in an areal density CGR of 100%. These past/projected growth rates are the starting points for our roadmap. Even though we have benefited from these growth rates over the past decade, it is important to recognize some stumbling blocks which can affect their continued growth trends in the future:

- The growth in BPI is expected to slow down due to several factors. First, increases in linear density would require lower head fly heights. With current head fly heights already being only a few nanometers from the platter surface, it is very difficult to reduce this further. Second, increasing the BPI requires higher recording medium coercivity, for which, as we mentioned in Section 3.1, it is not feasible to design a write head with currently known materials. Finally, the smaller grains can lead to superparamagnetic effects.
- The CGR for TPI is also expected to decline [7]. Increasing the TPI requires that tracks be narrower, which makes them more susceptible to media noise. Further, more closely spaced tracks can lead to inter-track interference effects. Finally, the track edges are

noisier than the center region and the edge effects increase with narrower tracks.

As the BARs have also been dropping, there has been a larger slowdown in the BPI CGR than that for the TPI. It has been shown [7] that there exist optimal values for the BAR for a given areal density. The BAR is around 6-7 for disks today and is expected to drop to 4 or below in the future [22]. Furthermore, it is expected [15, 37] that the growth in areal density would slow down to 40-50%. Given this growth in areal density, the industry projections predict the availability of an areal density of 1 Tb/in² in the year 2010.

We studied a set of proposals for creating such a terabit density disk [45, 31, 43]. In particular, we are interested in the feasible values for BPI and TPI, given all the constraints related to the recording medium, head design, and noise margins, for constructing reliable terabit density disks. Among the proposals, we chose the one with more conservative assumptions about the BPI, which does not scale as well as TPI, to obtain values of 1.85 MBPI and 540 KTPI giving a BAR of 3.42 (which agrees with current expectations). We then adjusted the CGRs for the BPI and TPI to achieve this areal density in the year 2010, together with the expected BAR trends. This provides a BPI and TPI CGR of 14% and 28% respectively (down from the original values of 30% and 50%), to give an areal density CGR of about 46% per year.

Once we have the anticipated BPI and TPI for each year, we then generate a “roadmap”, starting from the year 2002, for a period of ten successive years, i.e., upto the year 2012. The basic premise when doing this is that we are trying to sustain the expected IDR growth rate of *at least* 40% per year over the 11 year period. The steps when generating the temperature-dictated disk drive technology roadmap are given below:

1. For each year, we first plug in the BPI and TPI from the above estimates into our capacity model calculated in section 3.1 for a given platter size and number of platters - carried over from the previous year. For the resulting disk configuration, we can calculate its IDR for a given RPM (which is again carried over from the previous year), by putting in the appropriate values for n_{tz0} in equation 4. If the resulting IDR meets the projected 40% growth rate, then the new configuration would remain within the tolerable thermal envelope (since the same platter size, number of platters and RPM yielded a disk within the thermal envelope in the previous year).
2. However, if the disk from step 1 does not meet the target IDR for that year, one option is to see whether increasing the RPM can get us to this IDR (by putting this value in the LHS of equation 4 and finding the rpm). We can then use the resulting disk configuration and RPM value in the thermal model of the disk in section 3.3 to see whether this remains within the thermal envelope. If it does, then we have achieved the target IDR using the same number of platters and platter sizes as the previous year by increasing the RPM.
3. If the necessary RPM from step 2 does not keep the new disk within the thermal envelope, then the other option for still meeting the IDR target is to shrink the platter sizes. Note that the viscous dissipation is proportional to the 4.6th of the platter size, and the 2.8th power of the RPM. Further, a smaller platter size implies shorter seeks, thus reducing VCM power as well. Consequently, it is possible to remain within the thermal envelope by shrinking the platter size (note that the resulting n_{tz0} in equation 4 decreases) and increasing the RPM proportionally to compensate for the drop in IDR.

4. Shrinking the platter size as in step 3 results in a drop in the overall capacity. To compensate for this reduction, it may become necessary to add platters, causing us to repeat all the steps enumerated above.

Thus, the roadmap is not a single disk drive design point but is a spectrum of different platter sizes (and their corresponding RPMs) that try to sustain the IDR growth rate from year to year. When generating this roadmap, we consider the initial platter size (in the year 2002) to be 2.6", with two subsequent shrinks of 2.1" and 1.6" for later years. We do not consider smaller platter sizes due to the unavailability of VCM power correlations, and disk enclosure design considerations at such small media sizes. For each platter size in a given year, we consider configurations with 1, 2, and 4 platters. These represent disks for the low, medium, and high capacity market segments for the same technology generation. Increasing the number of platters also increases the viscous dissipation. We take this into account and *provide different external cooling budgets for each of the three platter counts* in order to use the same thermal envelope (45.22 C) for these higher platter disks at the beginning of the roadmap. We assume that the cooling technology remains invariant over time, and the disks need to be designed for the thermal envelope solely based on internal choices. We have studied the ramifications of changes in the cooling system and the results are given in [18].

4.1. Results

As per our methodology, we would like to first find out what would be the disk speed required for a given platter size and its resultant thermal impact, when trying to meet the 40% IDR growth target. In the absence of any thermal constraints, if we are to meet the IDR target for a given year, we would use the largest platter size possible and merely modulate the RPM to reach the desired value (step 2 of the method). Table 1 gives the RPM that is required in each year for the three platter sizes that we consider and the steady state temperature that is reached for a one platter configuration. Trends for 2 and 4 platter configurations are similar, and are not explicitly shown here.

Let us analyze these results by first focusing on the 2.6" platter size. The IDR requirements (shown as $IDR_{Required}$ in the Table), from the year 2002 to 2012, increase nearly 29 times. A portion of the required increase is provided by the growth in the linear density, denoted in the Table as $IDR_{density}$. Any demands beyond that has to be provided by an increase in the RPM. For instance, the RPM requirements grow nearly 9.5 times from the year 2002 to 2012. For a better understanding, let us sub-divide the timeline into three regions, namely, the years before 2004, where the BPI and TPI/CGRs are 30% and 50% respectively, the years from 2004 to 2009, which are in the sub-terabit areal densities, and the region from 2010 to 2012. Recall that the growth rates in BPI and TPI slow down after 2003 to 14% and 28% respectively and the ECC overheads for terabit areal density disks would increase to 35%. The effects of these trends are shown in the Table, where there is only a 7.7% increase in the required RPM from 2002 to 2003, but the required RPM growth increases to about 23% per annum in the post-2003 region. During the terabit transition (from 2009 to 2010), a sudden 70% increase in RPM is required. This happens because of the way we model the impact of ECC, as a sudden increase from 10% to 35% when transitioning into the terabit region. Realistically, this transition would be more gradual. After this steep increase, the RPM growth rate again steadies out to 23% for the subsequent years.

When examining the thermal characteristics of the 2.6" drive, we find a similar trend for the three temporal regions of the roadmap. The heat due to viscous dissipation increases from 0.91 W in 2002 to 1.13 W in 2003. In the second region, due to the higher rotational speed growth (and its relationship in the cubic power), the viscous dissipation grows from 2 W in 2004 to over 35.55 W in 2009, causing a significant rise in temperature, well beyond the thermal envelope of 45.22 C. Therefore, all other things remaining constant, it is clear that future single platter disk drives would not be able to provide the desired IDRs at the 2.6" platter size. The viscous dissipation increases even further from year 2010 onwards and reaches a value of 499.73 W in 2012, causing the internal drive air temperature to reach as high as 602.98 C for this platter size.

The effect of shrinking the platter can be observed by examining the results for the 2.1" and 1.6" drives in Table 1. Even though a smaller platter size implies a higher RPM is needed to meet the required IDR (than for the 2.6" drive), we see that the higher RPMs can be somewhat offset by moving to the smaller sizes, helping us stay within the thermal envelope until around 2007. Beyond that, even the 1.6" size is too big to stay within the envelope.

Having seen that RPM increase is not always a viable option in drive design to achieve the target IDR, let us now analyze the impact of the thermal envelope in meeting the IDR requirements and the resulting capacity. Figure 2 shows the maximum achievable data rates (and the corresponding capacities) for the spectrum of disk designs where the points are all within the thermal envelope. For each experiment (with a given number of platters each of a given size), we find the maximum RPM that it can run at without exceeding the thermal envelope. This coupled with the density values for the corresponding year can be used to calculate the maximum IDR (and its capacity) that such a disk can sustain within the envelope. In addition to these lines, the IDR graphs also plot the 40% growth rate target (the dotted line). The IDR roadmap points which yield a value in any year larger than the corresponding value in the dotted line indicate that the corresponding configuration can yield a higher data rate than the target. Typically, in such years, the manufacturer of such a disk may opt to employ a lower RPM to just sustain the target IDR, rather than what is really possible. The more interesting points are where the roadmap lines intersect the target IDR line. Note that the y-axes of all IDR roadmap graphs are in log-scale.

Let us first consider the 1-platter roadmap. We can see that the 1.6" platter, and the 2.1" to a certain extent, are able to provide (and even exceed in the earlier years) the target IDR until about 2006. The 2.6" platter size, however, starts falling short of being able to meet the projections from 2003 onwards. The 2.1" and 1.6" sizes reach their maximum allowable RPMs in the 2004-2005 and 2006-2007 timeframes respectively, after which they fall short of the IDR requirements. At such points, the manufacturer is presented with three options:

- Sacrifice the data rate and retain capacity growth by maintaining the same platter size.
- Sacrifice capacity by reducing the platter size to achieve the higher data rate.
- Achieve the higher IDR by shrinking the platter but get the higher capacity by adding more platters.

For example, consider the year 2005. From Table 1, we notice that a speed of 30,367 RPM would be required to meet the IDR for the 2.1" size. However, this is 1,543 RPM in excess of what is required to be within the thermal envelope. If we shrink the platter to 1.6", we would be able to achieve this data rate with an RPM of 39,857. However, for the one platter device, the capacity drops from 61.13 GB to

Year	2.6"			2.1"			1.6"			IDR _{Required}
	IDR _{density}	RPM	Temp. (C)	IDR _{density}	RPM	Temp. (C)	IDR _{density}	RPM	Temp. (C)	
2002	128.14	15098	45.24	103.50	18692	43.56	78.86	24533	41.64	128.97
2003	166.53	16263	45.47	134.51	20135	43.69	102.51	26420	41.74	180.56
2004	189.85	19972	46.46	153.34	24728	44.37	116.83	32455	42.15	252.78
2005	216.37	24534	48.26	174.81	30367	45.61	133.19	39857	42.93	353.89
2006	246.66	30130	51.48	199.23	37303	47.85	151.83	48947	44.29	495.44
2007	281.19	37001	57.18	227.12	45811	51.81	173.04	60127	46.73	693.62
2008	320.47	45452	67.27	258.91	56259	58.81	197.27	73840	51.04	971.07
2009	365.34	55819	85.04	295.08	69109	71.17	224.88	90680	58.63	1359.5
2010	300.23	95094	223.01	242.49	117735	167.01	184.75	154527	117.61	1903.3
2011	342.13	116826	360.40	276.44	144586	262.19	210.62	189769	176.20	2664.61
2012	390.03	143470	602.98	315.02	177629	430.93	240.11	233050	279.75	3730.46

Table 1. The thermal profile of the RPM required to meet the IDR CGR of 40% for different platter-sizes. We assume a single-platter disk with $n_{zones} = 50$ and a 3.5" form-factor enclosure. The thermal envelope is 45.22 C.

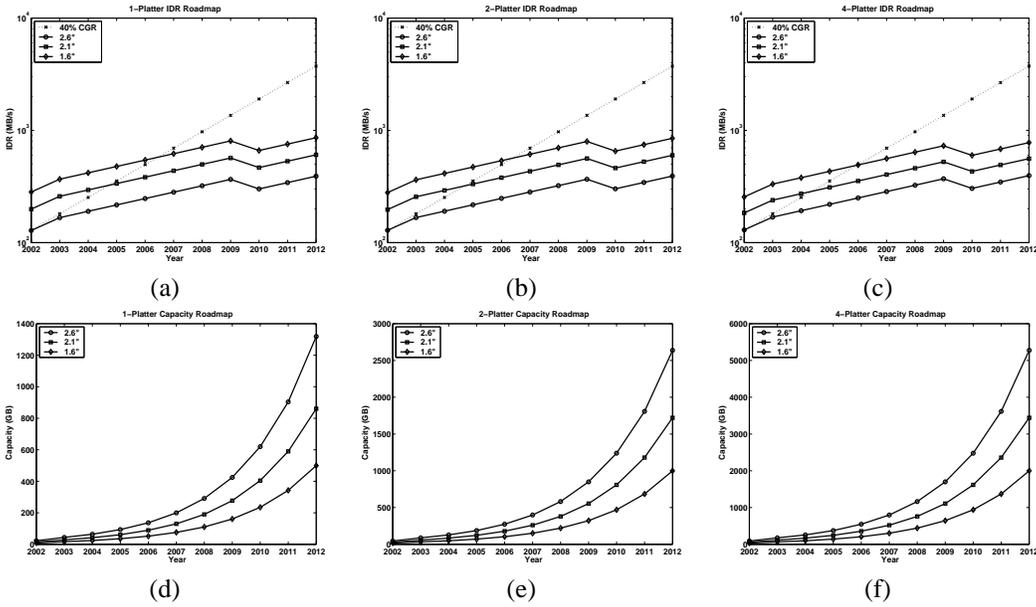


Figure 2. Disk Drive Roadmap. Each solid curve (for a given platter size) gives the maximum attainable IDR (in the top 3 graphs) with that configuration which is within the thermal envelope of 45.22 C, and the corresponding capacity (in the bottom 3 graphs), for a given year. The dotted line indicates the 40% target growth rate in IDR over time. Any curve which falls below this dotted line fails to meet the target for those years.

just 35.48 GB. If the manufacturer wishes to achieve a capacity that is closer to the 2.1" system, an additional platter may be added to push the capacity of the 1.6" drive to 70.97 GB. At this point, the roadmap would shift into the 2-platter system and consequently increase the cooling requirements for the product. In general, we find that the IDR growth of 40% can be sustained till the year 2006. The growth from 2006 to 2007, for the 1.6" platter-size, dips to 25% and to only 14% per-annum subsequently.

When transitioning to terabit areal densities in the year 2010, due to the large increase in the ECC overheads, which is not offset by the BPI growth, the IDR drops from 805.24

MB/s in 2009 to 661.39 MB/s in 2010. After this, the IDR growth of 14% is resumed. By the year 2012, there is over a 2,870 MB/s gap between the 40% CGR point and the best data rate achievable from our design space. Similar trends can be observed for the 2 and 4 platter roadmaps as well with the difference that the fall off from the roadmap is slightly steeper (despite conservatively assuming a higher cooling budget for them), since they incur higher viscous dissipation making RPM an even bigger issue in restricting their maximum data rates.

4.2. Impact of Other Technological Considerations

We have studied the influence of different drive parameters such as its form factor, aggressiveness of ZBR, etc., and other external conditions such as the effectiveness of the cooling system, on the above roadmap. Overall, the results are quite similar and the interested reader is referred to [18] for the details.

5. Dynamic Thermal Management (DTM)

We have seen that designing disk drives for the future, to provide the growth in data rates that we have enjoyed so far is going to be a challenge, both in terms of engineering complexity and cooling cost. In this Section, we present two possible remedies - Dynamic Thermal Management (DTM) mechanisms - for boosting performance while working under the thermal constraints:

1. Detecting thermal slack (difference between current temperature and the thermal envelope that the disk has been designed for), and exploit this slack to temporarily ramp-up RPM for better performance in multi-speed disks [17].
2. Deploying a disk that has been designed for the *average case* behavior to run it at a higher RPM than what the worst case would support most of the time, and use dynamic throttling techniques when getting close to thermal emergencies.

Although the first mechanism above would require multi-speed disks, such support is not necessary for the second mechanism (though it can be used for better throttling abilities). Note that in the following discussion, we merely point out the possibilities with these approaches together with identifying the trade-offs from a purely theoretical/synthetic perspective, rather than present detailed DTM algorithms/solutions (which is part of our future work). Before discussing these two possibilities, the question one may ask is whether such performance improvements are really needed from the application perspective. Consequently, we first (section 5.1) examine some commercial application traces to motivate the need for higher data rates. Subsequently, we discuss the above two mechanisms as possible ways of achieving such data rates in sections 5.2 and 5.3 respectively.

5.1. The Need for Faster Disks

Even though it is apparent that higher data rates would help bandwidth limited workloads, one is still interested in finding out how helpful this can be in the context of realistic workloads which may be intensive in seeks (that do not really benefit from a higher RPM). We conducted this evaluation using 5 commercial I/O traces given in Table 2. We used our model to capture some of the disk characteristics for the appropriate year (since this information was not always available). All the disks are assumed to have a 4 MB disk cache and ZBR with 30 zones/surface. For the RAID systems, RAID-5 was used with a stripe-size of 16 512-byte blocks. The performance of the disks was simulated using DiskSim [13] with the appropriate RPM. We conducted experiments by increasing RPM in steps of 5000 (without their thermal effects) to find the impact on response time. We summarize the results of this experiment below. The detailed results are given in [18].

Workload	Year	# Req.	RPM	Disk Cap. (GB)	# Disks	RAID?
Openmail	2000	3,053,745	10K	9.29	8	Yes
OLTP	1999	5,334,945	10K	19.07	24	No
Web	1999	4,579,809	10K	19.07	6	No
TPC-C	2002	6,155,547	10K	37.17	4	Yes
TPC-H	2002	4,228,725	7.2K	35.96	15	No

Table 2. Workloads Used. The Openmail workload was obtained from [1] and the OLTP and Web workloads from [44].

We find that a 5000 RPM increase from the baselines provides significant benefit in the I/O response time. Overall, the average response-times improved by 20.8%-52.5% for the 5000 RPM boost. These results suggest that these workloads would have benefited from a higher RPM even in those systems where they were initially run, though one may not have been able to get there because of the thermal envelope. This makes a strong case for continuing to support higher RPM in future disk drives, even those beyond thermal limits as long as we can provision dynamic thermal management techniques to avoid hitting those limits. As the subsequent two mechanisms illustrate, it would be possible to supply the additional 5-15K RPM, which provided us with the bulk of the performance benefits in these traces, with DTM.

5.2. Exploiting Thermal Slack

Note that the thermal envelope was previously defined based on the temperature attained with both the VCM and the SPM being on (i.e. the disk is constantly performing seeks). However, during idle periods (when not serving requests), the VCM is off, thus generating less heat. Further, there could be sequentiality in requests, reducing seek activities. This implies that there is a "thermal slack" to be exploited between the thermal envelope and the temperatures that would be attained if the VCM was off. However, the disk drive has been pre-set with a maximum RPM for a thermal limit based on the VCM being on constantly. If on the other hand, the disk provided multi-speed abilities, then we could temporarily push the RPM even higher during periods of no/few seeks without exceeding the thermal limits.

Note that a detailed study with workloads, and a realistic RPM modulation technique based on the seek characteristics and current temperatures is needed to evaluate the benefits of exploiting this slack in practice. Such a detailed study is beyond the scope of this paper, and is part of future work. Here we simply quantify the thermal slack of different designs and the higher RPMs that we may be able to drive it to when exploiting this slack.

Figure 3 (a) shows the RPM that we can drive the design to (for different platter sizes) when exploiting this slack, compared to the original maximum RPM we could support assuming the VCM was always on. We see that there is plenty of slack for the 2.6" platter size, allowing its speed to increase up to 26,750 RPM from the 15,020 RPM with the original thermal envelope. In terms of the data rate, this boost allows it to exceed the 40% CGR curve until the 2005-2006 time frame (Figure 3(b)). Even after this time frame, the data rates are around 5.6% higher than the non-slack based configuration. In fact, the slack for the 2.6" drive allows it to surpass a non-slack based 2.1" configuration, thus

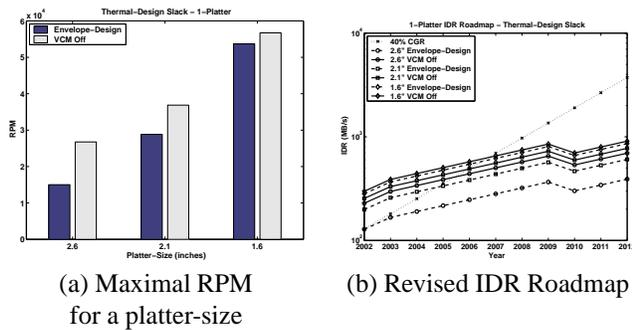


Figure 3. Exploiting the Thermal Slack. VCM-off corresponds to the RPM and IDR values attainable when the thermal slack is exploited. Envelope-Design corresponds to the RPM/IDR values when the VCM is assumed to be always on.

providing both better speed and higher capacity for the same thermal budget.

The amount of available slack decreases as the platter size is shrunk (see Figure 3(a)), since the VCM power is lower for smaller platter sizes (2.28 W for 2.1" vs. 0.618W for 1.6"). This makes the slack smaller to exploit in future designs with smaller platters. The next solution strategy can turn out to be more rewarding in such situations.

5.3. Dynamic Throttling

We consider two alternatives to exceed the thermal envelope RPM when designing disk drives, and dynamically modulating/throttling their behavior when we get close to the limits. These techniques are schematically shown in Figure 4. The basic idea is that by building a disk with higher RPM, we can benefit on performance in the average case, and throttle the system only when temperature limits are reached.

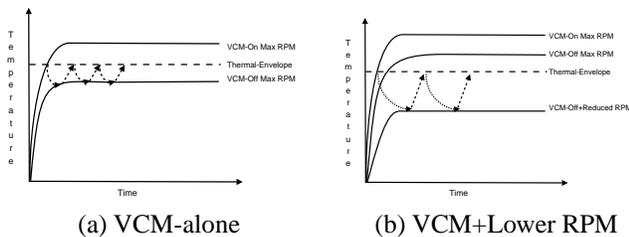


Figure 4. Dynamic Throttling Scenarios in the context of disks designed with average case behavior assumptions. In (a), only the VCM is off, with the disk continuing to spin at maximum RPM. In (b), the VCM is off and the disk is transitioned to a lower RPM.

Let us first consider the scenario in Figure 4(a). Here, if both the SPM and VCM are operating continuously, the temperature of the disk (depicted by the legend "VCM-On Max RPM") would violate the thermal envelope. In the absence of the VCM (either there are no requests issued to it, or the requests are sequential to avoid seeks), the temperature falls below the envelope (depicted by the legend "VCM-Off Max RPM"). The throttling mechanism can then operate as follows. Requests are sent to the disk, operating at this higher RPM (than permitted by the normal thermal envelope) until the temperature is close to the thermal limit. At that point, the requests are not issued to the disk for a while (t_{cool}), giving a thermal profile indicated by the downward-pointing dotted curve. After this period, requests (involving seeks) can be resumed and the disk would start heating up again (shown by the rising dotted curve), till it reaches close to the thermal envelope again in time t_{heat} .

Figure 4(b) shows a scenario for throttling with an even more aggressive (in terms of IDR) disk. In this disk, even turning off the VCM would not allow the disk to be within the thermal envelope since the RPM is so high. However, if the RPM was a lower value, then the temperature that would be reached with the VCM off for this lower RPM (depicted by the legend "VCM-Off+Reduced RPM") is lower than the thermal envelope. In this case, the throttling mechanism would not just stop issuing requests (to cut down VCM power) but would also pull down the RPM of this disk, when the temperature reaches close to the thermal envelope for a time t_{cool} as indicated in the Figure, and then let requests go on for time t_{heat} after bringing up the disk to full RPM. Note that in this case, even though we are performing RPM modulation, we only need a disk with 2 RPM levels, with the servicing of requests always being done only at the higher RPM. Such a disk is already available in the market today [24], since it only requires setting different SPM speeds, and does not need any further innovations in recording technologies. A full-fledged multi-speed disk [17] that services requests at different speeds, though not necessary for this throttling mechanism, can provide even finer granularity of temperature control.

The utility of both these techniques is largely dependent on the relative time given to cooling (t_{cool}) and the time it takes to get back from the lower temperature back to the thermal limits (t_{heat}). We call this ratio ($\frac{t_{heat}}{t_{cool}}$) as the *throttling-ratio*. In practice, we would like this ratio to be larger (greater than 1) since that allows for longer periods of operation of the disk compared to inoperation (i.e. its utilization is greater than 50%).

Let us consider a disk-drive that consists of a single 2.6" platter. The highest RPM that can be achieved by this disk, under the assumptions of our original roadmap, is 15020 RPM. Now let us suppose that we would like to be able to use the 2.6" size and be able to satisfy the 40% IDR CGR till the year 2005. From Table 1, we find that this needs an RPM of 24,534. Let us assume that we would like to build a disk which operates at this RPM even though in the worst case it would violate the thermal envelope and heat up to 48.26 C. We find that, if the VCM is turned off, the temperature of the drive is 44.07 C, which is within the design envelope and is thus a candidate for the first throttling approach. With such a disk (constant RPM of 24,534), we set the initial temperature to the thermal envelope. We then turn off the VCM for a specific period of time (t_{cool} in seconds) and then turn it back on again. We observe the time (t_{heat}) it takes for the disk temperature to reach the envelope. We repeat this experiment for various values of t_{cool} and the corresponding throttling ratios are plotted in Figure 5(a).

For the second throttling scenario, let us say that we would like to stretch the 2.6" roadmap to meet the CGR expectations till the year 2007, whereby a RPM of 37,001

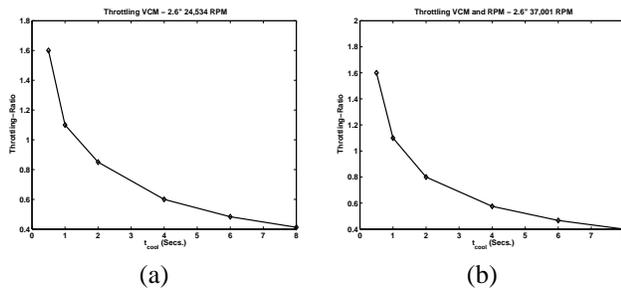


Figure 5. Throttling ratios with different t_{cool} for (a) VCM-alone and (b) VCM+Lower RPM

RPM would be required. The disk temperatures with and without the VCM turned on are 57.18 C and 53.04 C respectively, both of which are above thermal limits. We assumed that the disk drive is designed to operate at two RPMs, namely, the full-speed of 37,001 RPM and a lower-level of 22,001 RPM. We conduct a similar experiment as before, except that the RPM is also lowered by 15,000 in addition to turning off the VCM, when thermal limits are reached. The resulting throttling ratio graph for this scheme is shown in Figure 5(b).

Both these graphs give insight on how long we need to let the disk cool (throttle) between successive periods of activity. In both cases, we find that if we want to keep the active periods at least as long as the idle periods, throttling needs to be done at a relatively finer granularity (less than a second). The implications of these results, together with a discussion of possible techniques for throttling are given in the next section.

5.4. Discussion and Future Work

The results presented in this section indicate that there is some amount of thermal slack between when the VCM is on and off (section 5.2) to temporarily ramp up the RPM. Such techniques can buy us some IDR in the near future. However, as platter sizes continue to diminish, the benefits with this technique are likely to become less significant due to the drop in VCM power. This technique also requires multi-speed disks with advanced technologies to allow reading/writing at different speeds.

The more promising approach seems to be the dynamic throttling strategy where we can use a disk designed for the average case behavior and allow it to operate at its full performance abilities. Throttling is employed to reduce/avoid sending requests to the disk for a cooling down period, before resuming again. Here again, the second scenario where lowering the RPM (during t_{cool}) in addition to turning off the VCM, is more promising since just turning off the VCM may not buy too much slack in the future. Further, this only requires a couple of RPM levels, with the requests being always serviced at the highest level. Such a disk (from Hitachi) [24] is already available in the market today.

The attractiveness of implementing throttling even with existing disks warrants a closer investigation of techniques for achieving the intended goals with little performance loss. Our throttling ratio graphs indicate that keeping the disk utilization higher than 50% requires a finer granularity of throttling - possibly at the granularity of a few dozen requests. If the inter-arrival times of requests in a workload are appropriately spaced, then one could achieve such

throttling at little cost (or even for free). Even if the workload keeps issuing requests at a faster rate, there may be future work to be done in enhancing caching techniques to appropriately space out requests for allowing cooling, similar to how there has been recent work on caching for power management of disks [49]. Techniques for co-locating data items to reduce seek overheads (e.g. [38]) can reduce VCM power, and further enhance the potential of throttling. Finally, it is also possible to use mirrored disks (i.e. writes propagate to both) while reads are directed to one for a while, and then sent to another during the cool down period. The throttling ratio graphs give an indication of how many of these disks may need to be employed for a desired cool down period.

There are several other techniques to enhance IDR while remaining within thermal bounds. For instance, we could use two disks, each with a different platter size. The larger disk, due to its thermal limitations, would have a lower IDR than the smaller one, although the latter, assuming the platter counts to be the same, would have a lesser capacity. Such a configuration allows the smaller disk, which itself could have capacities in the order of several Gigabytes, to serve as a cache for the larger one. This is somewhat similar in philosophy to previously proposed cache-disks [26].

In this paper, our primary intention has been to briefly explore the potential offered by a few of these techniques in order to identify directions for further research. Estimating the benefits to be obtained with these techniques mandates a careful application-driven evaluation together with an exploration of the detailed design space of DTM control policies.

6. Concluding Remarks

This paper has presented an integrated framework for studying the inter-relationships between capacity, performance and thermal characteristics of disk drives, and how this can be used to chart out a roadmap based on a given thermal envelope. Though there could be minor variations in absolute values due to modeling inaccuracies, one cannot deny the sharp drop off in anticipated IDR growth rates as we move in to the future, because of the thermal envelope and emerging limitations in growth of areal densities, growing capacity needs of error correcting codes, together with the cooling costs.

We have presented Dynamic Thermal Management (DTM) as an option for buying back at least some of the loss in IDR growth rates for the future, by either exploiting the thermal slack, or by throttling disk activities. As we mentioned, these options are achievable even on existing disks. By employing these options, we find that there is around 5-10K RPM to gain in the near future. Even though this may not be enough to get us back all the way to the IDR roadmap, this gain still provides substantial improvements in response times for several server I/O traces. Even if one does not wish to consider DTM as a way of amplifying data rates, it is important to reiterate that temperatures have a considerable impact on long term drive reliability [2], and we can use DTM just to reduce the average operating temperature for enhancing reliability.

This paper identifies several future directions for research including those for developing DTM control algorithms. We also have plans in the future to release the code for the thermal model and its integration into the DiskSim infrastructure.

Acknowledgements: We would like to thank Steve Hetzler, Gordon Hughes, Ed Grochowski, and Erik Riedel for their

inputs on different aspects of disk drive design. This research has been supported in part by NSF grants 0429500, 0325056, 0130143, and 0103583.

References

- [1] G. Alvarez, K. Keeton, E. Riedel, and M. Uysal. Characterizing Data-Intensive Workloads on Modern Disk Arrays. In *Proceedings of the Workshop on Computer Architecture Evaluation Using Commercial Workloads*, January 2001.
- [2] D. Anderson, J. Dykes, and E. Riedel. More Than An Interface - SCSI vs. ATA. In *Proceedings of the Annual Conference on File and Storage Technology (FAST)*, March 2003.
- [3] K. Ashar. *Magnetic Disk Drive Technology: Heads, Media, Channel, Interfaces, and Integration*. IEEE Press, 1997.
- [4] D. Brooks and M. Martonosi. Dynamic Thermal Management for High-Performance Microprocessors. In *Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA)*, pages 171–182, January 2001.
- [5] E. Carrera, E. Pinheiro, and R. Bianchini. Conserving Disk Energy in Network Servers. In *Proceedings of the International Conference on Supercomputing (ICS)*, June 2003.
- [6] 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.
- [7] J. Chen and J. Moon. Detection Signal-to-Noise Ratio versus Bit Cell Aspect Ratio at High Areal Densities. *IEEE Transactions on Magnetics*, 37(3):1157–1167, May 2001.
- [8] N. Clauss. A Computational Model of the Thermal Expansion Within a Fixed Disk Drive Storage System. Master's thesis, University of California, Berkeley, 1988.
- [9] D. Colarelli and D. Grunwald. Massive Arrays of Idle Disks for Storage Archives. In *Proceedings of the Conference on Supercomputing*, pages 1–11, November 2002.
- [10] F. Douglass and P. Krishnan. Adaptive Disk Spin-Down Policies for Mobile Computers. *Computing Systems*, 8(4):381–413, 1995.
- [11] P. Eibeck and D. Cohen. Modeling Thermal Characteristics of a Fixed Disk Drive. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 11(4):566–570, December 1988.
- [12] G. Ganger. *System-Oriented Evaluation of I/O Subsystem Performance*. PhD thesis, The University of Michigan, June 1995.
- [13] G. Ganger, B. Worthington, and Y. Patt. *The DiskSim Simulation Environment Version 2.0 Reference Manual*. <http://www.ece.cmu.edu/ganger/disksim/>.
- [14] J. Griffin, J. Schindler, S. Schlosser, J. Bucy, and G. Ganger. Timing-Accurate Storage Emulation. In *Proceedings of the Annual Conference on File and Storage Technology (FAST)*, January 2002.
- [15] E. Grochowski. Hitachi GST, San Jose Research Center. Private Correspondence.
- [16] E. Grochowski and R. Halem. Technological Impact of Magnetic Hard Disk Drives on Storage Systems. *IBM Systems Journal*, 42(2):338–346, 2003.
- [17] 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.
- [18] S. Gurumurthi, A. Sivasubramaniam, and V. Natarajan. Disk Drive Roadmap from the Thermal Perspective: A Case for Dynamic Thermal Management. Technical Report CSE-05-001, The Pennsylvania State University, February 2005.
- [19] S. Gurumurthi, J. Zhang, A. Sivasubramaniam, M. Kandemir, H. Franke, N. Vijaykrishnan, and M. 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 (ISPASS)*, pages 123–132, March 2003.
- [20] G. Herbst. IBM's Drive Temperature Indicator Processor (Drive-TIP) Helps Ensure High Drive Reliability. In *IBM Whitepaper*, October 1997.
- [21] S. Hetzler. IBM Almaden Research Center. Private Correspondence.
- [22] Hitachi Global Storage Technologies - HDD Technology Overview Charts. <http://www.hitachigst.com/hdd/technology/overview/storagetechnology.html>.
- [23] Hitachi Global Storage Technologies - Research and Technology Overview. <http://www.hitachigst.com/hdd/research/storage/pm/>.
- [24] Hitachi Power and Acoustic Management - Quietly Cool. In *Hitachi Whitepaper*, March 2004. http://www.hitachigst.com/tech/techlib.nsf/productfamilies/White_Papers.
- [25] I. Hong and M. Potkonjak. Power Optimization in Disk-Based Real-Time Application Specific System s. In *Proceedings of the International Conference on Computer-Aided Design (ICCAD)*, pages 634–637, November 1996.
- [26] Y. Hu and Q. Yang. DCD Disk Caching Disk: A New Approach for Boosting I/O Performance. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*, pages 169–178, May 1996.
- [27] R. Huang and D. Chung. Thermal Design of a Disk-Array System. In *Proceedings of the InterSociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pages 106–112, May 2002.
- [28] G. Hughes. Center for Magnetic Recording Research, University of California, San Diego. Private Correspondence.
- [29] H. Levy and F. Lessman. *Finite Difference Equations*. Dover Publications, 1992.
- [30] Y.-H. Lu, E.-Y. Chung, T. Simunic, L. Benini, and G. Micheli. Quantitative Comparison of Power Management Algorithms. In *Proceedings of the Design Automation and Test in Europe (DATE)*, March 2000.
- [31] M. Mallary, A. Torabi, and M. Benakli. One Terabit Per Square Inch Perpendicular Recording Conceptual Design. *IEEE Transactions on Magnetics*, 38(4):1719–1724, July 2002.
- [32] H. Ottesen and G. Smith. Servo Format for Disk Drive Data Storage Devices. In *United States Patent 6,775,081*, August 2001.
- [33] A. E. Papatthanasou and M. L. Scott. Energy Efficient Prefetching and Caching. In *Proceedings of the USENIX Annual Technical Conference*, June 2004.
- [34] D. Patterson, G. Gibson, and R. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID). In *Proceedings of ACM SIGMOD Conference on the Management of Data*, pages 109–116, June 1988.
- [35] R. Patterson, G. Gibson, E. Ginting, D. Stodolsky, and J. Zelenka. Informed Prefetching and Caching. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, pages 79–95, December 1995.
- [36] I. Reed and G. Solomon. Polynomial Codes Over Certain Finite Fields. *Journal of the Society for Industrial and Applied Mathematics*, 8:300–304, June 1960.
- [37] E. Riedel. Device Trends - Where Disk Drives are Headed. In *Information Storage Industry Consortium (INSIC) Workshop on the Future of Data Storage Devices and Systems (DS2)*, April 2004.
- [38] C. Ruemmler and J. Wilkes. Disk Shuffling. Technical Report HPL-91-156, HP Laboratories, October 1991.
- [39] N. Schirle and D. Lieu. History and Trends in the Development of Motorized Spindles for Hard Disk Drives. *IEEE Transactions on Magnetics*, 32(3):1703–1708, May 1996.
- [40] Seagate Cheetah 15K.3 SCSI Disc Drive: ST3734553LW/LC Product Manual, Volume 1. <http://www.seagate.com/support/disc/manuals/scsi/100148123b.pdf>.
- [41] K. Skadron, M. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan. Temperature-Aware Microarchitecture. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*, pages 1–13, June 2003.
- [42] M. Sri-Jayantha. Trends in Mobile Storage Design. In *Proceedings of the International Symposium on Low Power Electronics*, pages 54–57, October 1995.
- [43] G. Tarnopolsky. Hard Disk Drive Capacity at High Magnetic Areal Density. *IEEE Transactions on Magnetics*, 40(1):301–306, January 2004.
- [44] UMass Trace Repository. <http://traces.cs.umass.edu>.
- [45] R. Wood. The Feasibility of Magnetic Recording at 1 Terabit per Square Inch. *IEEE Transactions on Magnetics*, 36(1):36–42, January 2000.
- [46] B. Worthington, G. Ganger, Y. Patt, and J. Wilkes. On-Line Extraction of SCSI Disk Drive Parameters. In *Proceedings of the ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 146–156, May 1995.
- [47] R. Youssef. RAID for Mobile Computers. Master's thesis, Carnegie Mellon University Information Networking Institute, August 1995.
- [48] J. Zedlewski, S. Sobti, N. Garg, F. Zheng, A. Krishnamurthy, and R. Wang. Modeling Hard-Disk Power Consumption. In *Proceedings of the Annual Conference on File and Storage Technology (FAST)*, March 2003.
- [49] Q. Zhu, F. David, C. Devraj, Z. Li, Y. Zhou, and P. Cao. Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management. In *Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA)*, February 2004.