Youngjae Kim
e-mail: youkim@cse.psu.edu

Jeonghwan Choi
e-mail: jechoi@cse.psu.edu

Department of Computer Science and Engineering,
The Pennsylvania State University,
University Park, PA 16802

Sudhanva Gurumurthi
Department of Computer Science,
University of Virginia,
Charlottesville, VA 22904
e-mail: gurumurthi@cs.virginia.edu

Anand Sivasubramaniam
Department of Computer Science and Engineering,
The Pennsylvania State University,
University Park, PA 16802
e-mail: anand@cse.psu.edu

Managing Thermal Emergencies in Disk-Based Storage Systems

Thermal-aware design of disk-drives is important because high temperatures can cause reliability problems. Dynamic thermal management (DTM) techniques have been proposed to operate the disk at the average case temperature, rather than at the worst case by modulating the activities to avoid thermal emergencies caused by unexpected events, such as fan-breaks, increased inlet air temperature, etc. A delay-based approach to adjust the disk seek activities is one such DTM solution for disk-drives. Even if such a DTM approach could overcome thermal emergencies without stopping disk activity, it suffers from long delays when servicing the requests. In this paper, we investigate the possibility of using a multispeed disk-drive (called dynamic rotations per minute (DRPM)), which dynamically modulates the rotational speed of the platter for implementing the DTM technique. Using a detailed performance and thermal simulator of a storage system, we evaluate two possible DTM policies—time-based and watermark-based—with a DRPM disk-drive and observe that dynamic RPM modulation is effective in avoiding thermal emergencies. However, we find that the time taken to transition between different rotational speeds of the disk is critical for the effectiveness of this DTM technique.
[DOI: 10.1115/1.2993152]

Keywords: storage system, disk-drives, temperature management

1 Introduction

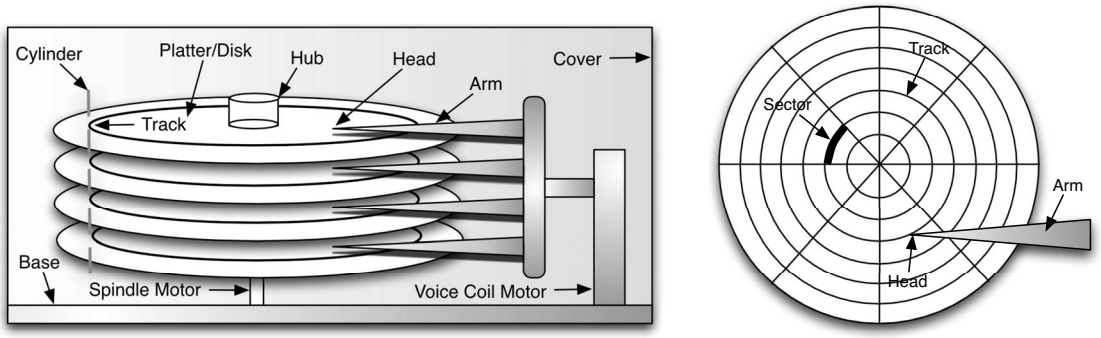
Thermal-awareness is becoming an integral aspect in the design of all computer system components, ranging from micro-architectural structures within processors to peripherals, server boxes, racks, and even entire machine rooms. This is increasingly important due to the growing power density at all the granularity of the system architecture. Deeper levels of integration, whether it be within a chip, or components within a server, or machines in a rack/room, cause a large amount of power to be dissipated in a much smaller footprint. Since the reliability of computing components is very sensitive to heat, it is crucial to drain away excess heat from this small footprint. At the same time, the design of cooling systems is becoming prohibitively expensive, especially for the commodity market [1,2]. Consequently, emerging technologies are attempting to instead build systems for the common case—which may not be subject to the peak power densities, and thereby operate at a lower cooling cost—and resort to dynamic thermal management (DTM) solutions when temperatures exceed safe operational values. This paper explores one such technique for implementing DTM for disk-drives.

Disk-drive performance is highly constrained by temperature. It can be improved by a combination of higher rotational speeds of the platters (called RPM), and higher recording densities. A higher RPM can provide a linear improvement in the data rate. However, the temperature rise in the drive enclosure can have nearly cubic relation to the RPM [3]. Such a rise in temperature can severely impact the reliable operation of the drive. Higher temperatures can cause instability in the recording media, thermal expansion of platters, and even outgassing of spindle and voice-coil motor lubricants, which can lead to head crashes [4]. One way of combating this generated heat is by reducing the platter sizes, which reduces the viscous dissipation by the fifth power. However, a smaller platter leads to a smaller disk capacity, unless more platters are added (in which case the viscous dissipation increases again by a linear factor). Moreover, a higher number of bits are necessary for storing error correcting codes to maintain acceptable error rates due to lower signal-to-noise ratios in future disk-drives. All these factors make it difficult to sustain the continued 40% annual growth that we have been enjoying in the data rates until now [1]. This makes a strong case for building drives for the common case, with solutions built-in for dynamic thermal management when the need arises. DTM has been already implemented in Seagate Barracuda ES drive in the industry [5].

There is one other important driving factor for DTM. It is not enough to consider individual components of a computing system in isolation any more. These components are typically put together in servers, which are themselves densely packed in racks in machine rooms. Provisioning a cooling system that can uniformly control the room so that all components are in an environment that matches the manufacturer specified "ambient" temperatures can be prohibitively expensive. With peak load surges on some components, parts of a room, etc., there could be localized thermal emergencies. Further, there could be events completely external to the computer systems—heating, ventilating, and air conditioning (HVAC)/fan breakdown, machine room door left open, etc.—which can create thermal emergencies. Under such conditions, today's disk-drives could overheat and fail, or some thermal monitor software could shut down the whole system. The disk is, thus, completely unavailable during those periods. The need to sustain 24/7 availability, and growing power densities lead to the increased likelihood of thermal emergencies. This makes it necessary to provide a "graceful" operation mode for disk-drives. During this graceful mode, even if the disk is not performing as well as it would have when there was no such emergency, it would still continue to service requests, albeit slowly. This graceful mode would essentially be a period during which certain dynamic thermal management actions are carried out in the background, while continuing to service foreground requests.

Multispeed disk operation [6,7] has been proposed as a solution to reduce disk-drive power, and can thus be a useful mechanism for thermal management as well. This mechanism is based on the observation that it is faster to change the rotational speed of a disk

2007 IEEE. Reprinted with permission from "Thermal Issues in Emerging Technologies: Theory and Application, 2007. THETA 2007. International Conference on 3–6 Jan. 2007." Manuscript received September 30, 2007; final manuscript received May 11, 2008; published online November 14, 2008. Assoc. Editor: Mohamed-Nabil Sabry.



(a) Vertical view (b) Horizontal view

Fig. 1 (*a*) shows a side view of the mechanical components of a disk-drive and (*b*) shows a view from the top

rather than spinning it all the way down/up. DRPM allows the disk to service requests at a slower rate even at the lower RPM. During a thermal emergency, we can not only reduce the speed to reduce temperature but also continue to service requests at the lower speed. A multispeed disk with two rotational speeds is commercially available [8]. Since the heat dissipated during the operation at a lower RPM is also much lower, the temperature within the drive can be lowered by employing this option during a thermal emergency. While a Hitachi's multispeed disk does provide a smaller window of time when the disk cannot service requests compared with a disk that only provides on-off modes, it still does not serve requests when it is at a lower RPM.

In this paper, we explore two options for temperature management during a thermal emergency. We first consider disks that are tuned for maximum performance with the ideal/constant ambient temperature. We then introduce thermal emergencies—by adjusting the external ambient temperature of the drive—which pushes the drive into the emergency regions. We then investigate these two multispeed drive options, and show that it is indeed possible for some regions of external ambient temperature variation to service disk requests even though such situations would have caused the drive to completely shut down in a nonmultispeed drive. As is to be expected, the performance during those periods is not as good as it would be when there are no emergencies. Between the two multispeed options, not servicing the requests at the lower RPM causes frequent switches between RPMs, thus not faring as good as the DRPM disk in its availability. All these experiments are conducted using a detailed performance and thermal simulator of storage system, called STEAM [9].

The organization of the rest of this paper is as follows. Section 2 describes basic disk-drive and its thermal modeling. Microbenchmark evaluations of a multispeed disk under external inlet ambient temperature variance are shown in Sec. 3. The effectiveness of the DTM technique for thermal emergencies with real server workloads are in Sec. 4. Finally Sec. 5 concludes this paper.

2 Disk-Drive and Thermal Modeling

2.1 Basic Disk-Drive Components and Behavior. Figure 1 shows the geometry of the mechanical components of a disk-drive from vertical and horizontal views. The disk platters are attached to the hub, which is empty inside. The hub connects to the base through a spindle motor. An arm actuator motor (voice-coil motor) is placed on the right end of the base. The arm-assembly is composed of one or more arms and each head for data read and write from/to the platter is placed at the end of each arm. The unit of data read-write is sector. The sector is represented in Fig. 1(*b*) as a thick curve. As shown in Fig. 1(*b*), a set of sectors composes a track and a disk-platter is composed of multiple tracks. A cylinder is a group of tracks that are vertically grouped as shown in the figure. All these components are enclosed by the disk-cover and are closed to the ambient air.

When an input/output (I/O) request is sent to the disk-drive from the host, disk operation behaves differently according to the request type (read or write). When the request is a read, the disk-controller first decodes the request. Then the controller disconnects from the bus and starts a seek operation that moves the arm-assembly toward the target track of data on the platter. The disk head at the end of the arm will properly settle down (head positioning time). Then, the data transfer occurs from the disk media to the host through a small computer system interface (SCSI) bus. Since reading the data of the disk media is slower than sending it over the bus, partially buffering is first required before sending it over the bus. Moreover, during this data transfer, head switch operation (to move the head to the next track) might be involved if necessary. When data transfer finishes, the complete status message is sent to the host. When the request is a write, data transfer to the disk's buffer is overlapped with head positioning time. If the head positioning time finishes, the data transfer to the disk media from the buffer. Also head switching could be involved on the needs as in read requests. If this data-recording on the media finishes, the complete status message is sent to the host.

2.2 Computational Model of Thermal Expansion in a Disk-Drive. The thermal simulation model is based on the one developed by Eibeck and Cohen [10]. The sources of heat within the drive include the power expended by the spindle motor (to rotate the platters) and the voice-coil motor/arm actuator motor (for moving the disk arms). The thermal model evaluates the temperature distribution within a disk-drive from these two sources by setting up the heat flow equations for different components of the drive such as the internal air, the spindle and voice-coil motor assemblies, and the drive base, and the cover as described in Fig. 1(*a*). The only interaction between the device components and the external environment is by conduction through the base and cover and subsequent convection to the ambient air. The finite difference method [11] is used to calculate the heat flow. It iteratively calculates the temperature of these components at each time step until it converges to a steady-state temperature. The accuracy of this model depends on the size of the time step. The finer the time step is, the more accurate the temperature distribution is over the disk-drive, but the simulation time is large [12].

2.3 Thermal Simulation Tool. Temperature-aware design has been explored for microprocessors [2], interconnection networks [13], storage systems [1], and even for the rack-mounted servers at machine rooms [14,15], because high temperature can lead to reliability problems and increase cooling costs. There have been various thermal simulation tools proposed to evaluate temperature-aware design. HotSpot [2] is a thermal simulator for

microprocessors using thermal resistance and capacitance derived from the layout of microprocessor architecture. STEAM [9] is a performance and thermal simulator for disk-drives that uses the finite difference method similar to that proposed by Eibeck Cohen [10] to calculate the heat flow and to capture the temperatures of different regions within the disk/storage system. MERCURY [16] is a software suite to emulate temperatures at specific points of a server by using a simple flow equation. In addition, THERMOSTAT [14] is a detailed simulation tool for rack-mounted servers based on computational fluid dynamics (CFD) [17], which provides more accurate thermal profiles by generating three-dimensional thermal profiles.

2.4 Dynamic Thermal Management. Dynamic thermal management has been adopted for individual components of the systems such as microprocessors [18,19] and disk-drives [12] or distributed environments such as distributed systems [20] and rack-mounted servers at data centers [14,15]. Of all these approaches, DTM for disk-drives has already been addressed [1].

A delay-based DTM has been applied to prevent this situation from happening [21]. When the temperature of the disk-drive reaches close to the thermal envelope, DTM is invoked by stopping all the requests issued; hence, all the seek activities stop and the service resumes only after the temperature is sufficiently reduced. However, even if this delay-based throttling (by controlling the seek activities) is feasible, many requests cannot be issued during thermal emergencies and thereby the performance is greatly affected by them. Today's Seagate's Barracuda ES drives have a similar DTM feature by adjusting the workloads for thermal management [5]. The other possible approach is to modulate the RPM speed in a multispeed disk. Since RPM has nearly cubic power relation to the viscous dissipation [3], it can be more effective to manage the temperature of the disk-drive. This technique of dynamic RPM modulation for thermal management is discussed in the rest of this paper.

3 Thermal Emergencies in Disk-Drives

Thermal emergencies are generally caused by unexpected events, such as fan-breaks, increased inlet air temperature, etc. These unexpected events threaten the reliability of the disk by causing data corruption on the disk. Unfortunately, predicting when such thermal emergencies happen in real time is a big challenge. In this section, we understand the impact of the external ambient temperature variation on the disk temperature with microbenchmark tests.

3.1 Disk-Drive Modeling for Simulation. In order to understand the heat distribution over all the components of a drive enclosure while it is in operation, we used STEAM to model an Ultrastar 146Z10 disk-drive [22] installed in a 42U computer system rack. The Ultrastar 146Z10 is composed of two 3.3″ platters and rotates at 10 K RPM. The power curve of the spindle motor (SPM) due to its rotational speed change is obtained using the equation describing the change in the rotation speed of the disk, which has a quadratic effect on its power consumption [7]. And the voice coil motor (VCM) power (which is dependent on the platter dimensions) is obtained by applying the power-scaling curve from Ref. [23]. The power values of SPM and VCM are set to be 10 W and 6.27 W, and all other required parameters such as disk geometry were supplied as inputs into the model.

3.2 Thermal Variation Over Disk-Drive. From Fig. 2, we see that the hottest component over the drive enclosure is the arm-assembly (which has heads at its end), whose temperature is around 68°C at Max (i.e., the disks are spinning and the arms are moving back and forth with VCM on all the time) and the lowest temperature is for the disk-cover surrounding the disk-drive (around 37°C at Max). This is because the heat is directly drained away to the ambient air through the convection process. When the disk-drive is Idle (the disks are spinning without any arm movement), they have a similar temperature distribution to when the disk-drive is Max (the disk-drive consumes maximum power while the disks are spinning and arms are always moving back and forth).

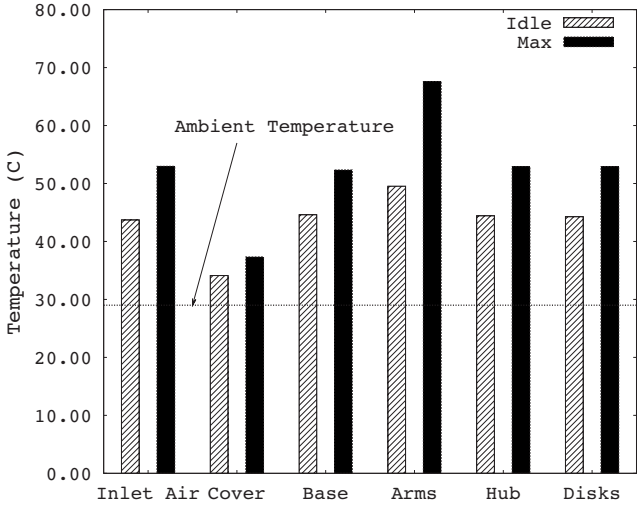


Fig. 2 Temperature distribution over IBM Ultrastar 146Z10. The dotted line denotes the ambient temperature (29°C). Each label in the *X*-axis indicates each component in the disk-drive. "Inlet Air" denotes external inlet air temperature. The description about all other labels can be found in Fig. 1(*a*).

3.3 Impact of Increased External Ambient Temperature on Disk-Drive's Temperature. We performed a microbenchmark evaluation to understand the impact of variation in ambient temperature on the disk temperature and the feasibility of dynamic throttling by RPM modulation in a multispeed disk. Figure 2 shows that there is high thermal variation in temperature over the disk-drive, which means that thermal sensor location is very critical in applying DTM to a multispeed disk. It is hard to decide the location that should be selected for detecting emergencies. The base temperature of the disk is chosen for DTM mechanism in this paper, because a thermal sensor is mounted on the back side of the electronics card close to the base of the actual disk-drive [4]. The highest RPM speed of a multispeed disk is restricted to 20 K and the baseline is 10 K RPM because 10 K RPM disk-drive is one of the most popular server disk-drives and 20 K RPM is known as the possible rotational speed of the disks for disk-drive's design until now. *Thermal slack* is defined as the temperature difference between current operating temperature and thermal envelope. We modeled two different disks for the experiments, where one is a 3.3″ one platter disk-drive used in HPL Openmail and the other is a disk-drive with 3.3″ four platters used for other workloads in Table 1.

We have measured the base temperature of the disk-drive with different ambient temperatures at the steady state in STEAM. We varied the ambient temperature from 29°C to 42°C for a 3.3 in. one platter disk and from 29°C to 33°C for the disk of 3.3 in. four platters. In the experiment, the thermal envelope was set to be 60°C because the possible operating temperature range of the disk-drive suggested in manuals is 5–60°C [22,26].

From Fig. 3(*a*), it is observed that thermal emergencies never happen with even 20 K rotational speed of platters and the VCM on all the time for 29°C ambient temperature. However, if the ambient temperature is increased further to 42°C, it could exceed the thermal envelope (60°C). For example, a multispeed disk operating at 20 K under 42°C is above the thermal envelope at both Idle and Max. However, if RPM drops down to 10 K, it comes below the thermal envelope at Idle, while it exceeds 60°C at Max. This result shows that thermal slack would be around 4°C at the

Table 1 Description of workloads and storage systems used and thermal emergency situations for real workloads. T_{init_amb} denotes the initial ambient air temperature. *Emg_amb* denotes the increased ambient air temperature due to thermal emergencies. *Emg_start* and *Emg_end*, respectively, denote the simulated thermal emergency starting and ending time.

Workload	Workload description and storage systems HPL Openmail[a]	OLTP application[b]	Search-Engine[b]	TPC-C
No. of requests	3,053,745	5,334,945	4,579,809	6,155,547
No. of disks	8	24	6	4
Per-disk capacity (Gbyte)	9.29	19.07	19.07	37.17
RPM	10,000	10,000	10,000	10,000
Platter diameter (in.)	3.3	3.3	3.3	3.3
No. of platters	1	4	4	4
	Thermal emergencies			
T_{init_amb} (°C)	29	29	29	29
T_{emg_amb} (°C)	42	33	33	33
Emg_start (s)	500.000	500.000	2,000.000	2,000.000
Emg_end (s)	2,500.000	30,000.000	12,000.000	10,000.000
Simulated time (s)	3,606.972	43,712.246	15,395.561	15,851.512

[a]Reference [24].
[b]Reference [25].

maximum. We also observe from Fig. 3(*b*) that, if the disk-drive has a larger number of platters in a similar disk geometry, it is more prone to thermal emergencies even for a small increase in ambient temperature because the heat dissipation inside the disk-drive is proportional to the number of platters [3]. Figure 3(*b*) shows that 33°C external ambient temperature introduces thermal emergency when it is operating at the higher speed. Thermal slack becomes larger around 10°C at the maximum. Moreover, even if the disk-drive is operating at the lower speed, thermal emergency could exceed the thermal envelope for even 29°C ambient temperature depending on the request patterns. Similarly, since the heat generated from the disk-drive is proportional to the 4.6th power of the disk-platter size [3,27], the disk-drive with the larger size of platters is more sensitive to variations in the ambient temperature.

Within these thermal slacks, a dynamic throttling mechanism can be applied to avoid thermal emergencies. It can be achieved by pulling down the rotational speed of the disks (when it reaches thermal emergencies). And then once the temperature is lower than a given thermal envelope, it brings up the disk to full rotational speed after the cooling period.

4 Designing Dynamic Thermal Management Technique for Disk-Drives

4.1 Multispeed Disk-Drive. In order to study the effect of a multispeed disk-drive when thermal emergencies happen, we have simulated the temperature behavior of RPM transitions in a multispeed disk-drive. We consider that the operable rotational speeds of the platters for this multispeed disk are 10 K and 20 K. From Fig. 3, we see how much the disk-drive's temperature can vary with different RPM speeds and external ambient temperatures. The maximum transition time taken between different rotational speeds of the disk is assumed to be 7 s (from the lower to the higher and vice versa as in the commercial multispeed disk-drive

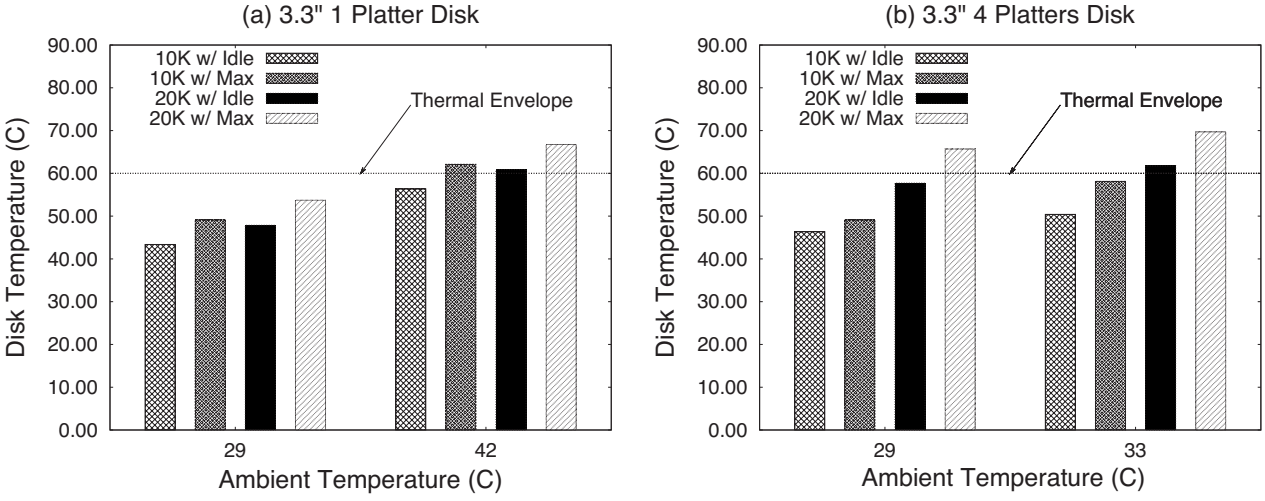


Fig. 3 Each bar denotes the average temperature of each component at the steady state for Max (where VCM is on all the times while the disk platters are spinning) and Idle (VCM just turns off). The horizontal line in each graph is the thermal envelope (60°C). (*a*) and (*b*) are the steady-state base temperatures of the disk for different disk dimensions (such as the size of platter and the rotational speed of platter) and different power consumption modes under various ambient temperatures. The horizontal line in each graph is the thermal envelope (60°C).

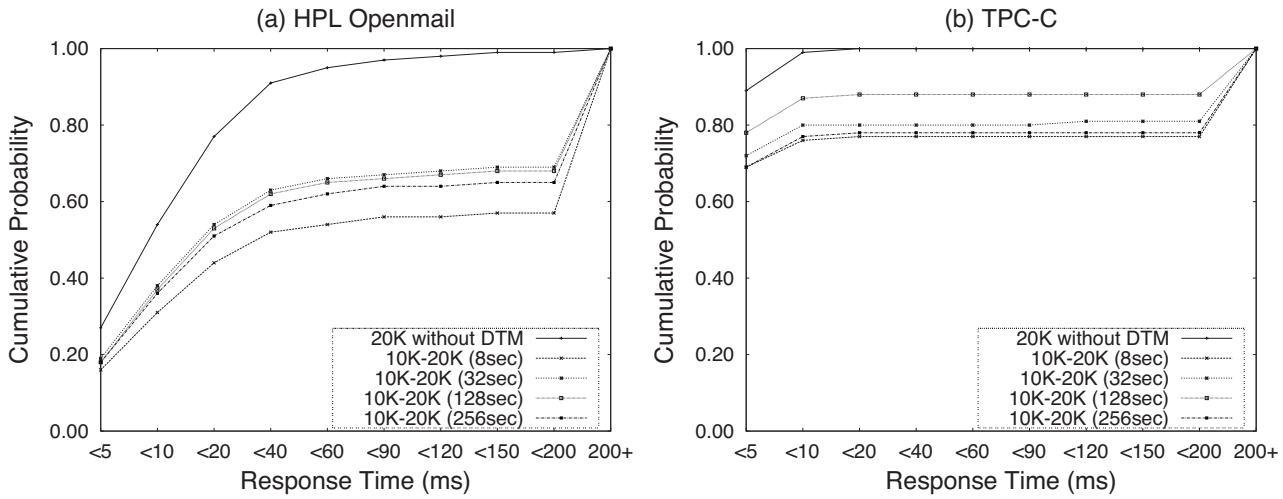


Fig. 4 Performance degradation of $DRPM_{simple}$ for the server workloads. The value in parentheses at each graph denotes a cooling unit time (which is given as a delay time, once it becomes close to the thermal envelope (60°C)).

of Hitachi [8]).

We used four commercial I/O traces for the experiment, whose characteristics are given in Table 1, and we consider two kinds of multispeed disk-drives as follows.

- $DRPM_{simple}$: This is the same approach as Hitachi's multi-speed disk, where the lower RPM is just used for cooling the hot disk, rather than servicing the requests.
- $DRPM_{opt}$: This is the technique that was proposed in Ref. [7], where the disk-drive still performs I/O at the lower RPM.

Note that the $DRPM_{opt}$ disk-drive is a disk-drive serviceable at any rotational speed of the platters while $DRPM_{simple}$ utilizes the lower rotational speed of the platters only for the disk-drive's

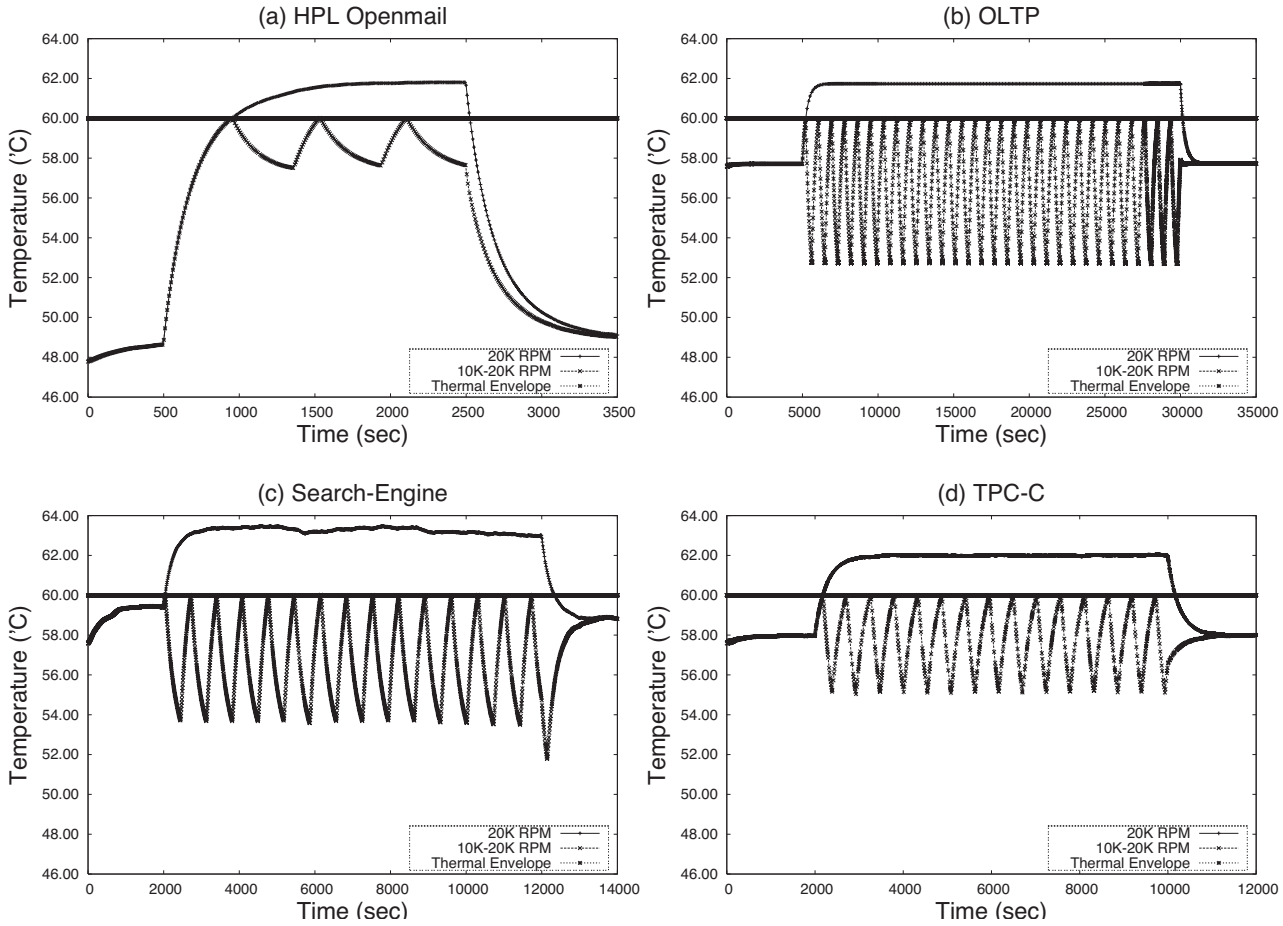


Fig. 5 Thermal profiles of the real workloads for $DRPM_{opt}$ under the scenarios of Table 1. They are all for the disk0 of disk arrays each of which is a 10–20 K multispeed disk with 7 s of RPM transition time and 400 s of a cooling unit time.

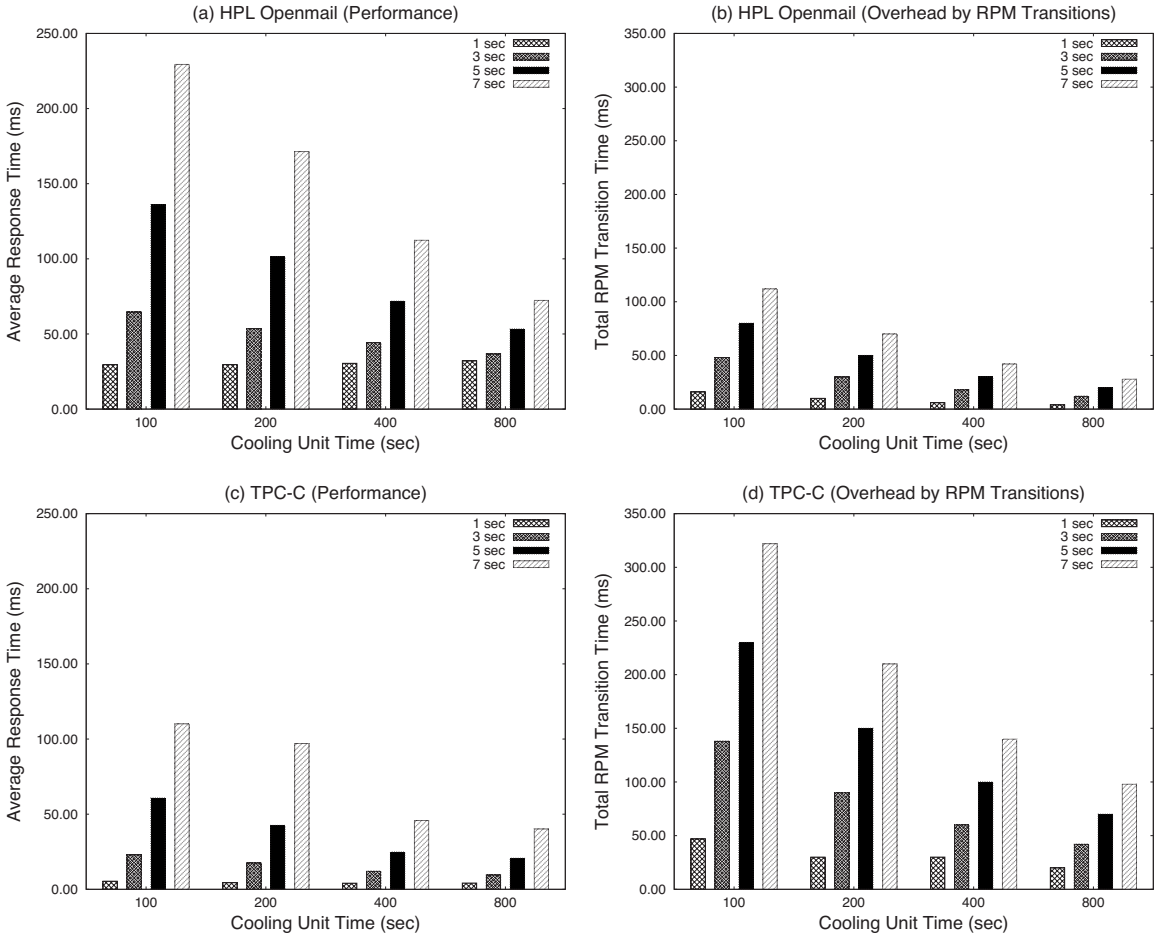


Fig. 6 Correlation between cooling unit time, RPM transition time, and performance (i.e., response time). Each bar denotes an average value across the disks at a disk array in the unit of millisecond.

cooling effect without servicing any request. We evaluate two possible DTM policies (called time-based and watermark-based) with these multispeed disk-drives.

4.2 Time-Based Throttling Policy. Time-based policy is based on a predefined period for cooling time before resuming to service the requests under thermal emergencies. The thermal sensor of the disk-drive periodically checks the temperature as a decent disk-drive does [26]. Once the disk's temperature reaches thermal emergency, the RPM drops down and the drive waits for a predefined period before resuming I/O operation by ramping up the RPM to full speed. Since $DRPM_{simple}$ is not available to service during the cooling and transition times, the performance is constrained by these two values. However, most server workloads generally have many requests issued with short interarrival times and they should be processed as quickly as possible. In addition, 7 s of delay for each RPM transition is not negligible to the performance of a multispeed disk.

Figure 4 shows the performance degradation by $DRPM_{simple}$, compared with the disk without any dynamic thermal management technique under thermal emergencies. Each graph shows the cumulative distribution function (CDF) of the average response time at an I/O driver across different disks. The response time is the time a disk-drive takes to finish a given input request. The solid curve in each graph shows the disk operating at the maximum speed of 20 K RPM without the DTM technique and others reflect a multispeed disk-drive with $DRPM_{simple}$. As is to be expected, many requests suffer from large delays (due to nonserviceable cooling time) more than 200 ms in the multispeed disk-drive of $DRPM_{simple}$. In Fig. 4(*a*), even 30–50% requests are serviced with their response times more than 200 ms while in Fig. 4(*b*) about 13–25% requests suffer from large response times more than 200 ms. Even if we varied cooling unit times to compensate for the performance degradation, none of them is effective for both workloads. $DRPM_{simple}$ might be desirable for a DTM solution, because such a straightforward policy does not only require significant additional complexity to the disk-controller design but also after reasonable delays, it could still overcome thermal emergencies by resuming the service below the thermal envelope.

$DRPM_{opt}$ has been designed to minimize the performance drawback caused by long delays of $DRPM_{simple}$ required for disk-drive's cooling effect. Figure 5 shows different thermal profiles for the workloads under thermal emergency situations described in Table 1. Since each disk-drive of the disk arrays of TPC-C, OLTP, and Search-Engine has the same disk dimension/characteristics and they have similar temperature profiles, we focus on the results for HPL Openmail and TPC-C. The upper curve for each graph is when DTM is not applied while operating at the maximum speed while the lower curve (going up and down) is the result from $DRPM_{opt}$ with 400 s of cooling unit time during which it operates at the lower rotational speed of the platters. As shown from Fig. 5, operating only at 20 K RPM exceeds the thermal envelope under thermal emergencies unless DTM is applied. However, $DRPM_{opt}$ avoids emergencies by dynamically modulating the rotational speed between high and low RPMs at need. However, even if $DRPM_{opt}$ could be available to service the requests during the cooling time, many RPM transitions (for example, as shown from many transitions for OLTP in Fig. 5) increase the overheads due to nonserviceable RPM transition time. Any arriving request during RPM transition should wait until it completes.

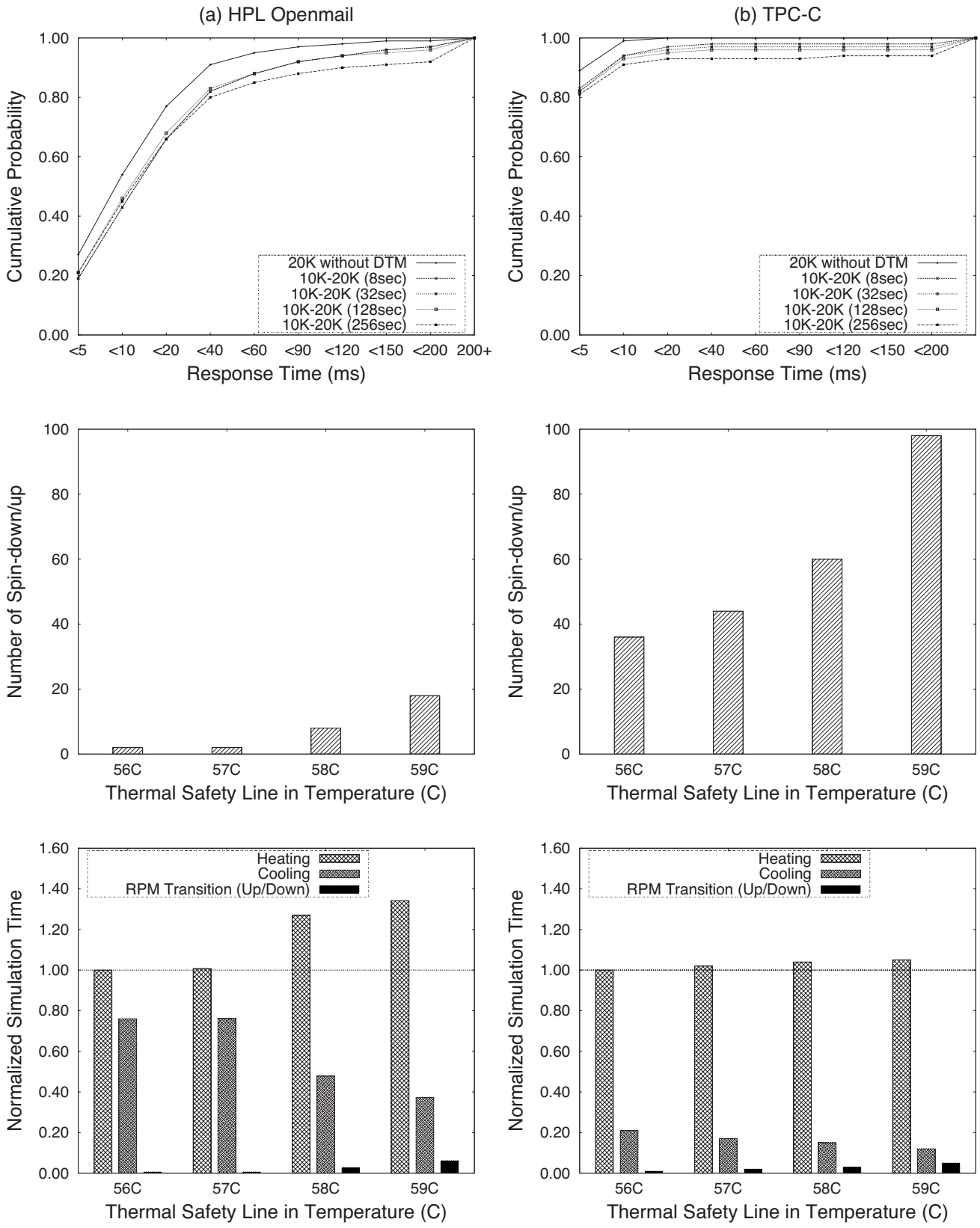


Fig. 7 Experimental results of $DRPM_{opt}$ using watermark-based policy for HPL Openmail and TPC-C. "Thermal Safety Line" in the graphs denotes temperature at which the disk-drive sufficiently cools down to operate.

The time taken for RPM transitions greatly affects the performance, and thus, to study the impact of non-negligible RPM transition on the performance, we have experimented $DRPM_{opt}$ for different RPM transitions ranging from 1 s to 7 s in steps of 2 s. In Fig. 6, cooling unit time is a predefined period during which the platters rotate at the lower rotational speed. The first column for each workload is to understand the correlation between cooling unit time and RPM transition time and the second column for each

workload shows the relationship between average total time taken for RPM transitions and cooling unit times. From Fig. 6(*a*) and 6(*c*), we see that a small RPM transition time shows better performance for a given constant cooling unit time. In addition, it is shown in Fig. 6(*b*) and 6(*d*) that high cooling times can hide the overhead of the RPM transitions by reducing the number of RPM transitions and sparing more time for I/O disk operations. In the DTM option of $DRPM_{opt}$, still servicing the requests at the lower rotational speed of the platters works positively in the performance improvement; however, it still has an upper bound in performance improvement. This is because more cooling implies that more requests should be serviced at the lowest speed of RPMs in a multispeed disk.

4.3 Watermark-Based Throttling Policy for $DRPM_{opt}$. *Watermark-based policy* uses two thresholds, T_{high} and T_{low}. As in time-based policy, the thermal sensor of the disk-drive periodically checks the temperature. If the thermal sensor detects that the temperature is close to thermal emergency (T_{high}), which is the temperature at which DTM is invoked, thermal management is applied to cool down the disk until the temperature gets down to the predetermined threshold (T_{low}). After this point, the disk-controller comes to know that the emergency has been resolved.

Figure 7 shows the experimental results of $DRPM_{opt}$ where T_{high} is 60°C (which is set to be the same as the thermal envelope in this experiment) and T_{low} is obtained by subtracting a few degree Celsius from T_{high}. The graphs in the first row of Fig. 7 show the CDF of the average response time across disk-drives. The solid curve in each graph represents the performance of the baseline system without any DTM and others are for DTM with different lower thresholds (T_{low}) (which is denoted by "Thermal Safety Line"). In this experiment, the RPM transition (up and down) time is assumed to 7 s. The lower thermal safety line of T_{low} helps the performance of a multispeed disk. This is because the lower value of T_{low} allows more relaxation in throttling and reduces the number of RPM transitions. As shown from the graphs in the second row of Fig. 7, the lower values of T_{low} result in a fewer number of RPM transitions. The graphs in the last row of Fig. 7 show the normalized simulation times to heating time at 56°C of thermal safety line. The larger fraction of heating time implies better performance because it implies that more requests could be serviced at the maximum speed. However, it is to be noted that it is not absolutely better than the small portion of heating time, because RPM transition time (in the order of seconds) offset this benefit.

5 Conclusions

This paper has presented graceful operation of a multispeed disk to handle thermal emergencies in large disk arrays. We studied several DTM policies (i.e., time-based and watermark-based) for different multispeed disk techniques (i.e., $DRPM_{simple}$ and $DRPM_{opt}$) executing real workloads and observed that the DRPM technique is one of the best solutions to avoid thermal emergencies.

$DRPM_{simple}$ technique overcomes thermal emergencies by dynamically modulating the rotational speed of disks and providing predefined delays. But such delays cause poor performance (such as response time), compared with a normal disk drive without any DTM technique. However, $DRPM_{opt}$ technique further improves the performance by continuously servicing the requests at the lower speed. time and watermark-based policies have been evaluated for thermal management and they showed that the time taken for RPM transition in a multispeed disk plays a critical role in the performance of thermal management.

Acknowledgment

This research has been funded in part by NSF Grant Nos. 0429500, 0325056, 0130143, 0509234, and 0103583, and an IBM Faculty Award.

References

[1] Gurumurthi, S., Sivasubramaniam, A., and Natarajan, V., 2005, "Disk Drive Roadmap From the Thermal Perspective: A Case for Dynamic Thermal Management," Proceedings of the International Symposium on Computer Architecture (ISCA), Jun., pp. 38–49.

[2] Skadron, K., Stan, M. R., Huang, W., Velusamy, S., Sankaranarayanan, K., and Tarjan, D., 2003, "Temperature-Aware Microarchitecture," Proceedings of the International Symposium on Computer Architecture (ISCA), Jun., pp. 1–13.

[3] Clauss, N. S., 1988, "A Computational Model of the Thermal Expansion Within a Fixed Disk Drive Storage System," M.S. thesis, University of California, Berkeley, CA.

[4] Herbst, G., 1997, "IBM's Drive Temperature Indicator Processor (Drive-TIP) Helps Ensure High Drive Reliability," IBM Whitepaper, Oct.

[5] Seagate Workload Management for Business-Critical Storage, http://www.seagate.com/docs/pdf/whitepaper/TP555_BarracudaES_Jun06.pdf.

[6] Carrera, E. V., Pinheiro, E., and Bianchini, R., 2003, "Conserving Disk Energy in Network Servers," Proceedings of the International Conference on Supercomputing (ICS), Jun.

[7] Gurumurthi, S., Sivasubramaniam, A., Kandemir, M., and Franke, H., 2003, "DRPM: Dynamic Speed Control for Power Management in Server Class Disks," Proceedings of the International Symposium on Computer Architecture (ISCA), Jun., pp. 169–179.

[8] 2004, "Hitachi Power and Acoustic Management: Quietly Cool," Hitachi Whitepaper, Mar., http://www.hitachigst.com/tech/techlib.nsf/productfamilies/White_Papers.

[9] Gurumurthi, S., Kim, Y., and Sivasubramaniam, A., 2006, "Thermal Simulation of Storage Systems Using STEAM," Proceedings of the IEEE Micro Special Issue on Computer Architecture Simulation and Modeling.

[10] Eibeck, P. A., and Cohen, D. J., 1988, "Modeling Thermal Characteristics of a Fixed Disk Drive," IEEE Trans. Compon., Hybrids, Manuf. Technol., **11**(4), pp. 566–570.

[11] Levy, H., and Lessman, F., 1992, *Finite Difference Equations*, Dover, New York.

[12] Kim, Y., Gurumurthi, S., and Sivasubramaniam, A., 2006, "Understanding the Performance-Temperature Interactions in Disk I/O of Server Workloads," Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA), Feb.

[13] Shang, L., Peh, L.-S., Kumar, A., and Jha, N. K., 2004,"Thermal Modeling, Characterization and Management of On-Chip Networks," Proceedings of the International Symposium on Microarchitecture (MICRO), Dec., pp. 67–78.

[14] Choi, J., Kim, Y., Sivasubramaniam, A., Srebric, J., Wang, Q., and Lee, J., 2007, "Modeling and Managing Thermal Profiles of Rack-Mounted Servers With ThermoStat," Proceedings of the International Symposium on High Performance Computer Architecture (HPCA), Feb., pp. 205–215.

[15] Sharma, R. K., Bash, C. E., Patel, C. D., Friedrich, R. J., and Chase, J. S., 2005, "Balance of Power: Dynamic Thermal Management for Internet Data Centers," IEEE Internet Comput., **9**(1), pp. 42–49.

[16] Heath, T., Centeno, A. P., George, P., Jaluria, Y., and Bianchini, R., 2006, "Mercury and Freon: Temperature Emulation and Management in Server Systems," Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems, Oct.

[17] Patterson, M. K., Wei, X., and Joshi, Y., 2005, "Use of Computational Fluid Dynamics in the Design and Optimization of Microchannel Heat Exchangers for Microelectronics Cooling," Proceedings of the ASME Summer Heat Transfer Conference.

[18] Brooks, D., and Martonosi, M., 2001, "Dynamic Thermal Management for High-Performance Microprocessors," Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA), Jan., pp. 171–182.

[19] Srinivasan, J., and Adve, S. V., 2003, "Predictive Dynamic Thermal Management for Multimedia Applications," Proceedings of the International Conference on Supercomputing (ICS), Jun., pp. 109–120.

[20] Weissel, A., and Bellosa, F., 2004, "Dynamic Thermal Management for Distributed Systems," Proceedings of the First Workshop on Temperature-Aware Computer Systems (TACS), Jun.

[21] Gurumurthi, S., 2006, "The Need for Temperature-Aware Storage Systems," Proceedings of the Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, May, pp. 387–394.

[22] Ultrastar 146Z10 hard disk drives specifications, http://www.hitachigst.com/hdd/support/146z10/146z10.htm.

[23] Sri-Jayantha, M., 1995,"Trends in Mobile Storage Design," Proceedings of the International Symposium on Low Power Electronics, Oct., pp. 54–57.

[24] The Openmail Trace, http://tesla.hpl.hp.com/private software/.

[25] UMass Trace Repository, http://traces.cs.umass.edu.

[26] Seagate Cheetah 15K.3 SCSI Disc Drive: ST3734553LW/LC Product Manual, Vol. 1, http://www.seagate.com/support/disc/manuals/scsi/100148123b.pdf.

[27] Schirle, N., and Lieu, D. F., 1996, "History and Trends in the Development of Motorized Spindles for Hard Disk Drives," IEEE Trans. Magn., **32**(3), pp. 1703–1708.