Computer Sciences Department

Robust, Portable I/O Scheduling with the Disk Mimic

Florentina I. Popovici
Andrea Arpaci-Dusseau
Remzi Arpaci-Dusseau

Technical Report #1457

November 2002

UNIVERSITY OF
WISCONSIN
MADISON

Robust, Portable I/O Scheduling with the Disk Mimic

Florentina I. Popovici, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau
Computer Sciences Department, University of Wisconsin, Madison

Abstract

We propose a new approach for I/O scheduling that performs on-line simulation of the underlying disk. When simulation is integrated within a system, three key challenges must be addressed: first, the simulator must be portable across the full range of devices; second, all configuration must be automatic; third, the computation and memory overheads must be low. Our simulator, the Disk Mimic achieves these goals by building a model of the disk as it observes the times for previous requests; each observed time is recorded in a table, indexed by a single parameter: the inter-request distance. We show that a shortest-mimic'ed-time-first (SMTF) scheduler, performs close to an approach with perfect knowledge of the underlying device and that it is superior to traditional scheduling algorithms such as C-LOOK and SSTF; our results hold as the seek and rotational characteristics of the disk are varied.

1 Introduction

High-performance disk schedulers explored in the research literature are becoming progressively more tuned to the performance characteristics of the underlying disks. Each generation of disk schedulers has accounted for more of the behavior of storage devices at the time. For example, disk schedulers analyzed in the 1970s and 1980s focused on minimizing seek time, given that seek time was often an order of magnitude greater than the expected rotational delay [8, 24, 28]. In the early 1990s, the focus of disk schedulers shifted to take rotational delay into account, as rotational delays and seek costs became more balanced [11, 19, 29].

At the next level of sophistication, a disk scheduler takes all aspects of the underlying disk into account: track and cylinder switch costs, cache replacement policies, mappings from logical block number to physical block number, and zero-latency writes. For example, Worthington *et al.* demonstrate that algorithms that effectively utilize a prefetching disk cache perform better than those that do not [29].

However, the more intricate the knowledge a scheduler has of the disk, the more barriers there are to its realization within operating system kernels. Specifically, there are three obstacles that must be overcome. First, the scheduler must discover the detailed knowledge of the underlying disk. Although a variety of tools have been described that automatically acquire portions of this knowledge [17, 23, 30], this knowledge must still be embedded into the disk model employed by the scheduler; the resulting scheduler is then configured to handle only a single disk with those specific characteristics. Second, the disk scheduler must also have knowledge of the current state of the disk, such as the exact position of the disk head. Given that head position is not exposed by current disk controllers and its position is not predictable due to low-level disk techniques such as wear leveling, predictive failure analysis, and log updates, the scheduler must control the current position using non-trivial techniques [31, 9]. Finally, the computational costs of a detailed modeling can be quite high [29]; it is not uncommon for the time to model request time to be larger than the time to service the request [3].

Due to these difficulties, few disk schedulers that leverage more than basic seek costs have been implemented for real disks. When considering rotational position, most previous work has been performed within simulation environments [11, 19, 21, 29]. The schedulers that have recently been implemented by researchers have either contained substantial simplifications [10] or have been painstakingly tuned for a small group of disks [9, 31]. Not surprisingly, the disk schedulers found in modern operating systems such as Linux, NetBSD, and Solaris, attempt to minimize only seek time.

1.1 A Different Approach

We believe that a promising alternative approach to embedding detailed knowledge of the disk into the scheduler is to embed an *on-line simulator* of the disk into the scheduler. An I/O scheduler is able to use on-line simulation of the underlying storage device to predict which request in its queue will have the shortest positioning time. Although a variety of disk simulators exist [3], most are targeted for performing traditional, off-line simulations, and unfortunately, the infrastructure for performing on-line simulation is fundamentally different.

In many respects, the requirements of an on-line simulator are more stringent than those of an off-line simulator. First, the on-line simulator must be *portable*; that is, the simulator must be able to model the behavior of any disk drive that could be used in practice. Second, the on-line simulator must have *automatic run-time configuration*, since one cannot know the precise characteristics of the underlying device when constructing the simulator; it is highly undesirable if a human administrator must interact with the simulator. Finally, the on-line simulator must have *low overhead*; the computation and memory overheads of an on-line simulator must be minimized such that the simulator does not adversely impact system performance.

In addition to the complexity it introduces, an on-line simulator also provides ample opportunities for simplification. First, the on-line simulator has the opportunity to observe the run-time behavior of the device; not only does this allow the simulator to configure itself on the fly, it also allows the simulator to adjust to changes in the behavior of the device over time. Second, the on-line simulator can be specialized for the problem domain in question. Finally, the on-line simulator does not need to be parameterizable; that is, since an on-line simulator is not exploring different versions of the device itself, the simulator does not need to contain a functional model of the device.

1.2 Contributions

In this paper, we address how to implement an I/O scheduler that is aware of the underlying disk technology in a simple, portable, and robust manner. To achieve this goal, we introduce the Disk Mimic, which meets the requirements of an on-line simulator for disk scheduling. The Disk Mimic is based upon a simple table-based approach, in which input parameters to the simulated device are used to index into a table; the corresponding entry in the table gives the predicted output for the device. A table-based approach is appropriate for on-line simulation because it can portably capture the behavior of a variety of devices, requires no manual configuration, and can be performed with little computational overhead. However, there is a significant challenge as well: to keep the size of the table tractable, one must identify the input parameters that significantly impact the desired outputs. The method for reducing this input space depends largely upon the domain in which the on-line simulator is deployed.

We show that for disk scheduling, a single input of the logical distance between two requests is sufficient for predicting the positioning time. However, when using the inter-request distance for prediction, two issues must be resolved. First, inter-request distance is a fairly coarse predictor of positioning time; as a result, there is high variability in the times for different requests with the same distance. The implication is that the Disk Mimic must observe many instances for a given distance and use an appropriate summary metric for the distribution; experimentally, we have found that summarizing a small number of samples with the mean works well. Second, given the large number of possible inter-request distances on a modern disk drive, the Disk Mimic can not record all distances in a table of a reasonable size. We show that simple linear interpolation can be used to represent ranges of missing distances, as long as some number of the interpolations within each range are checked against measured values.

We propose a new disk scheduling algorithm, shortest-mimic'ed-time-first (SMTF), which picks the request that is predicted by the Disk Mimic to have the shortest positioning time. We demonstrate that the SMTF scheduler can utilize the Disk Mimic in two different ways; specifically, the Disk Mimic can either be configured off-line or on-line, and both approaches can be performed automatically. When the Disk Mimic is configured off-line, it performs a series of probes to the disk with different inter-request distances and records the resulting times; in this scenario, the Disk Mimic has complete control over which inter-request distances are observed and which are interpolated. When the Disk Mimic is configured on-line, it records the requests sent by the running workload and their resulting times. Note that regardless of whether the Disk Mimic is configured off-line or on-line, the simulation itself is always performed on-line, within an active system.

In this work, we show that the Disk Mimic can be used to significantly improve the throughput of disks with high utilization. Specifically, for a variety of simulated and real disks, C-LOOK and SSTF perform between 10% and 50% slower than SMTF. Further, we demonstrate that the Disk Mimic can be successfully configured on-line; we show that while the Disk Mimic is learning about the storage device, SMTF performs no worse than a base scheduling algorithm (*e.g.*, C-LOOK or SSTF) and quickly performs close to the off-line configuration (*i.e.*, after approximately 750000 requests).

The rest of this paper is organized as follows. In Section 2 we describe the SMTF scheduler in more detail and in Section 3 we describe the Disk Mimic. We describe our basic methodology for evaluation in Section 4. Next, we investigate the issues of configuring the Disk Mimic off-line in Section 5. We then describe the additional complexities of configuring the Disk Mimic on-line and show its performance in Section 6. Finally, we describe related work and conclude.

2 I/O Scheduler

In this section, we briefly describe the approach of a new I/O scheduler that leverages the Disk Mimic. We refer to the algorithm implemented by this scheduler as shortest-mimic'ed-time-first, or SMTF. The basic function that SMTF performs is to order the queue of requests such that the request with the shortest position time, as determined by the Disk Mimic, is scheduled next. However, given this basic role, there are different optimizations that can be made. The assumptions that we use for this paper are as follows.

First, we assume that the goal of the I/O scheduler is to optimize the *throughput* of the storage system. We do not consider the fairness of the scheduler. We believe that the known techniques for achieving fairness (*e.g.*, weighting each request by its age [11, 19]), that have been added to approaches such as SSTF can be added to SMTF as well.

Second, we assume that the I/O scheduler is operating in an environment with heavy disk traffic. Given that the queue lengths at the disk may contain hundreds or even thousands of requests [11, 19], the computational complexity of the scheduling algorithm is an important issue [2]. Given these large queue lengths, it is not feasible to perform an optimal scheduling decision that considers all possible combinations of requests. Therefore, we consider a greedy approach, in which only the time for the next request is minimized [11].

To evaluate the performance of SMTF, we compare to the algorithms most often used in practice: first-come-first-served (FCFS), C-LOOK, and shortest-seek-time-first (SSTF). To compare our performance to the best possible case, we have also implemented a best-case-greedy scheduler for our simulated disks; this best-case scheduler knows exactly how long each request will take on the simulated disk and greedily picks the request with the shortest positioning time next.

3 The Disk Mimic

In this section, we describe the Disk Mimic, which captures the behavior of a disk drive in a portable, robust, and efficient manner. To predict the performance of a disk, the Disk Mimic uses a simple table, indexed by the relevant input parameters to the disk. Thus, the Disk Mimic does not attempt to simulate the mechanisms or components internal to the disk; instead, it simply reproduces the output as a function of the inputs it has observed.

3.1 Reducing Input Parameters

Given that the Disk Mimic uses a table-driven approach to predict the time for a request as a function of the observable inputs, the fundamental issue is reducing the number of inputs to the table to a tractable number. If the I/O device is treated as a true black box, in which one knows nothing about the internal behavior of the device, then the Disk Mimic must assume that the service time for each request is a function of all previous requests. Given that each request is defined by many parameters (*i.e.*, whether it is a read or a write, its block number, its size, the time of the request, and even its data value), this leads to a prohibitively large number of input parameters as indices to the table.

Therefore, the only tractable approach is to make assumptions about the behavior of the I/O device for the problem domain of interest. Given that our goal is for the I/O scheduler to be portable across the realistic range of disk drives, and not to necessarily work on any hypothetical storage device, we can use high-level assumptions of how disks behave to eliminate a significant number of input parameters; however, the Disk Mimic will make as few assumptions as possible.

Our current implementation of the Disk Mimic predicts the time for a request from only a single input parameter: the logical distance from the ending block of the previous request; we refer to this input as the *inter-request distance*. This conclusion that inter-request distance is the key parameter agrees with that of previous researchers [26].

We now briefly argue why inter-request distance is a suitable parameter in our domain. We begin by summarizing the characteristics of modern disk drives; much of this discussion is taken from the classic paper by Ruemmler and Wilkes [16]; the interested reader is referred to their paper for more details.

3.1.1 Background

A disk drive contains one or many *platters*, where each platter *surface* has an associated disk head for reading and writing. Each surface has data stored in a series of concentric circles, or *tracks*. A single stack of tracks at a common distance from the spindle is called a *cylinder*. Modern disks also contain static RAM to perform caching; the caching algorithm is one of the most difficult aspects of the disk to capture and model [22, 30].

The disk appears to its client as a linear array of logical blocks; these logical blocks are then mapped to physical sectors on the platters. This indirection has the advantage that the disk can reorganize blocks to avoid bad sectors and to improve performance, but it has the disadvantage that the client does not know where a particular logical block is located. If a client wants to derive this mapping, there are multiple sources of complexity. First, different tracks have different numbers of sectors; specifically, due to zoning, tracks near the outside of a platter have more sectors (and subsequently deliver higher bandwidth) than tracks near the spindle. Second, consecutive sectors across track and cylinder boundaries are skewed, to adjust for head and track switch times; the skewing factor differs across zones as well. Third, flawed sectors are remapped through sparing; sparing may be done by remapping a bad sector (or track) to a fixed alternate location or by slipping the sector (or track) and all subsequent ones to the next sector (or track).

Accessing a block of data requires moving the disk head over the desired block. The time for this has two dominant components: moving the disk head over the desired track, *seek time*, and waiting for the desired block to rotate under the disk head, *rotation latency*. The seek time for reads is likely to be less than that for writes, since reads can be performed more aggressively (*i.e.*, a read can be performed when the block is not yet quite available without losing data, while a write cannot). The time for the platter to rotate is roughly constant, but it may vary around 0.5 to 1% of the nominal rate; as a result, it is difficult to predict the location of the disk head after the disk has been idle for many revolutions.

3.1.2 Inter-Request Distance

Given this basic behavior, the inter-request distance between logical block addresses captures some of the underlying characteristics of the disk, while missing others. We note that ordering requests based on the time for a given distance is significantly different than using the distance itself; due to the complexity of disk geometry, some requests that are separated by a larger logical distance can be positioned more rapidly.

In the opinion of Ruemmler and Wilkes, the following aspects of the disk should be modeled for the best accuracy: seek time (calculated with two separate functions depending upon the seek distance, and different for reads and writes), head and track switches, rotation latency, data layout (including reserved sparing areas, zoning, and track and cylinder skew), and data caching (both read-ahead and write-behind). We briefly discuss the extent to which each of these components is captured with our approach.

Our approach accounts for the combined costs of seek time, head and track switches, and rotation layout, but in a probabilistic manner. That is, for a given inter-request distance, there is some probability that a request crosses track or even cylinder

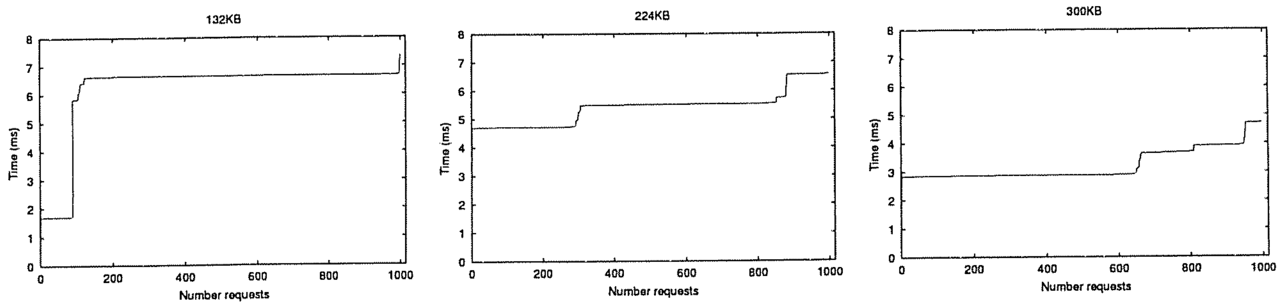


Figure 1: **Distribution of Off-Line Probe Times for Three Inter-Request Distances.** *Each graph shows a different seek distance: 132KB, 224KB, and 300KB. Along the x-axis, we show each of the 1000 probes performed (sorted by time) and along the y-axis we show the time taken by that probe. These times are for the IBM 9LZX disk.*

boundaries. Requests of a given distance that cross the same number of boundaries have the same total costs: the same number of track seeks, the same number of head and/or track switches, and the same amount of rotation.

We note that the table-based method for tracking seek time can be *more* accurate than that advocated by Ruemmler and Wilkes; instead of expressing seek time as a function of the seek distance, the Disk Mimic records the precise seek time for each distance. Although Ruemmler and Wilkes encourage treating the seek costs of reads and writes differently, this is not an issue for the Disk Mimic, since the difference is constant regardless of the I/O schedule created.

The cost incurred by the rotation of the disk has two components: the rotational distance between the previous and the current request and the elapsed time between the two requests (and thus, the amount of rotation that has already occurred). Although using inter-request distance probabilistically captures the rotational distance, the Disk Mimic does not record the amount of time that has elapsed since the last request. This omission is not an issue for disk scheduling in the presence of a full queue of requests; in this case, the inter-arrival time between requests at the disk is both negligible and identical and, thus, can be ignored. Ignoring time does cause inaccuracies when scheduling the first request after an idle period; however, if the disk is often idle, then I/O scheduling is not an important problem.

Data layout is incorporated fairly well by the Disk Mimic as well. The number of sectors per track and number of cylinders impact our measured values in that these sizes determine the probability that a request of a given inter-request distance crosses a boundary; thus these sizes impact the probability of each observed time in the distribution. Although zoning behavior and bad sectors are not tracked by our model, previous research has shown that this level of detail does not help with scheduling [29].

The aspect which we model the least directly is that of general caching. However, the Disk Mimic will capture the effects of simple prefetching, which is the most important aspect of caching for scheduling [29]. For example, if a read of one sector causes the entire track to be cached, then the Disk Mimic will observe the faster performance of accesses with distances less than that of a track. In this respect, configuring the Disk Mimic on-line by observing the actual workload could be more accurate than configuring off-line, since the locality of the workload is captured.

3.2 Results

To illustrate some of the complexity of using inter-request distance as the single predictor of request time, we show the distribution of times observed. For these experiments, we configure the Disk Mimic off-line as follows.

The Disk Mimic configures itself by probing the I/O device using fixed-size read operations (*i.e.*, 1KB). For each of the possible inter-request distances covering the disk (both negative and positive), the Disk Mimic samples a number of points of the same distance: it reads a block the specified distance from the previous block. To avoid any caching or pre-fetching performed by the disk, the Disk Mimic reads from a random location before each new probe of the required distance. The observed times are recorded in a table, indexed by the inter-request dis-

	Configuration	rotate	seek			head	cyl	track	cyl	sec/trk	heads
			1 cyl	400	3000	switch	switch	skew	skew		
1	**Base**	6	0.8	6.0	8	0.79	1.78	36	84	272	10
2	**Fast seek**	6	0.16	1.32	1.6	0.79	1.00	36	46	272	10
3	**Slow seek**	6	2.0	33.0	40.0	0.79	2.80	36	127	272	10
4	**Fast rotate**	2	0.8	6.0	8	0.79	1.78	108	243	272	10
5	**Slow rotate**	12	0.8	6.0	8	0.79	1.78	18	41	272	10
6	**Fast seek+rot**	2	0.160	1.32	1.6	0.79	1.00	108	136	272	10
7	**More capacity**	6	0.8	6.0	8	0.79	1.78	36	84	544	20
8	**Less capacity**	6	0.8	6.0	8	0.79	1.78	36	84	136	5

Table 1: **Disk Characteristics.** *Configurations of eight emulated disks. All times are in milliseconds. In most experiments, the base disk is used.*

tance.

In Figure 1 we show a small subset of the data collected on the IBM 9LZX disk. The figure shows the distribution of 1000 samples for three inter-request distances of 132KB, 224KB, and 300KB. In each case, the y-axis shows the request time of a sample and the points along the x-axis represent each sample, sorted by increasing request time.

We make two important observations from the sampled times. First, for a given inter-request distance, the observed request time is not constant; for example, at a distance of 132K, about 10% of requests require 1.8 ms, about 90% require 6.8 ms, and a few require almost $8ms$. Given this multi-modal behavior, the time for a single request cannot be reliably predicted from only the inter-request distance; thus, one cannot usually predict whether a request of one distance will be faster or slower than a request of a different distance. Nevertheless, it is often possible to make reasonable predictions based upon the probabilities: for example, from this data, one can conclude that a request of distance 132K is likely to take longer than one of 224K.

Second, from examining distributions for different inter-request distances, one can observe that the number of transitions and the percentage of samples with each time value varies across inter-request distances. The number of transitions in each graph corresponds roughly to the number of track (or cylinder) boundaries that can be crossed for this inter-request distance.

This data shows that a number of important issues remain regarding the configuration of the Disk Mimic. First, since there may be significant variation in request times for a single inter-request distance, what summary metric should be used to summarize the distribution? Second, how many samples are required to adequately capture the behavior of this distribution? Third, must each inter-request distance be sampled, or is it possible to interpolate intermediate distances? We investigate these issues in Section 5.

4 Methodology

To evaluate the performance of SMTF scheduling, we consider a range of disk drive technology. We have implemented a disk emulator that accurately models a seek curve, fixed rotation latency, track and cylinder skewing, and a simple segmented cache; this emulator is similar in spirit to that implemented elsewhere [6], but can use the memory of either a networked machine or the local machine as the storage media of the emulated disk. For most of our experiments, we configure the disk emulator to have performance characteristics similar to the IBM 9LZX disk. The characteristics of the eight disks that we emulate are summarized in Table 1. We also run our experiments on an IBM 9LZX disk.

To evaluate scheduling performance, we show results from the HP traces [15]; in most cases, we focus on the trace for the busiest disk from the week of 5/30/92 to 6/5/92. For our performance metric, we report the time the workload spent at the disk. To consider the impact of heavier workloads and longer queue lengths, we compress the inter-arrival time between requests. When scaling time, we attempt to preserve the dependencies across requests in the workload by observing the blocks being requested; we assume that if a request is repeated to a block that has not yet been serviced, that this request is dependent on the previous request first complet-

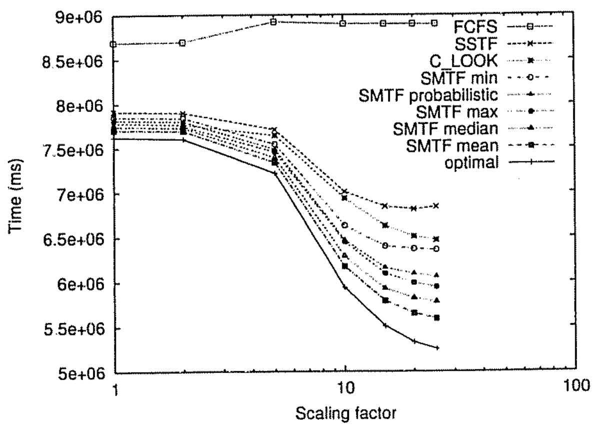


Figure 2: **Sensitivity to Summary Metrics.** *This graph compares the performance of a variety of scheduling algorithms on the base simulated disk and the week-long HP trace. For the SMTF schedulers, no interpolation is performed and 100 samples are obtained for each data point. The x-axis shows the compression factor applied to the workload. The y-axis reports the time spent at the disk.*

ing. Thus, we hold repeated requests, and all subsequent requests, until the previous identical request completes.

5 Off-Line Configuration

In this section, we explore the SMTF scheduler when the Disk Mimic has been configured off-line; again, although the the Disk Mimic is configured off-line, the simulation and predictions required by the scheduler are still performed on-line within the system. As described previously, configuring the Disk Mimic off-line involves probing the underlying disk with requests that have a range of inter-request distances. We note that even when the model is configured off-line, the process of configuring SMTF remains entirely automatic and portable across a range of disk drives. The main drawback to configuring the Disk Mimic off-line is a longer installation time when a new device is added to the system: the disk must be probed before it can be used for workload traffic.

5.1 Summary Data

To enable the SMTF scheduler to easily compare the expected time of all of the requests in the queue, the Disk Mimic must supply a summary value for each distribution as a function of the inter-request distance. Given the multi-modal characteristics of these distributions, the choice of a summary metric is non-obvious. Therefore, we evaluate five different summary metrics: `mean`, `median`, `maximum`, `minimum`, and `probabilistic`, which randomly picks a value from the sampled distribution according to its probability.

The results for each of these summary metrics on the base simulated disk are shown in Figure 2. For the workload, we consider the week-long HP trace, scaled by the compression factor noted on the x-axis. The graph shows that FCFS, SSTF, and C-LOOK all perform worse than each of the SMTF schedulers; as expected, the SMTF schedulers perform worse than the greedy-optimal scheduler, but the best approach is always within 7% for this workload. These results show that using inter-request distance to predict positioning time merits further attention.

Comparing performance across the different SMTF approaches, we see that each summary metric performs quite differently. The ordering of performance from best to worse is: `mean`, `median`, `maximum`, `probabilistic`, and `minimum`. It is interesting to note that the scheduling performance of each summary metric is not correlated with its accuracy. The accuracy of disk models is often evaluated according to its *demerit figure*, which is defined as the root mean square of the horizontal distance between the time distributions for the model and the real disk. This point is briefly illustrated in Figure 3, which shows the distribution of actual times versus the predicted times for three different metrics: `probabilistic`, `mean`, and `maximum`.[1]

As expected, the `probabilistic` model has the best demerit figure; with many requests, the distribution it predicts is expected to match that of the real device. However, the `probabilistic` model performs relatively poorly within SMTF because the time it predicts for any one request may differ significantly from the actual time for that request. Conversely, although the `maximum` value results in a poor demerit figure, it performs adequately for scheduling; in fact, SMTF with `maximum` performs significantly better than with `minimum`, even though both have similar demerit figures. Finally, using the `mean` as a summary of the distribution

[1] The relatively large differences between the actual and predicted distributions for small request times is due to a small discrepancy in our methodology: although we probed 1 KB requests, we schedule 4 KB requests. We will correct this error in the final version of the paper.

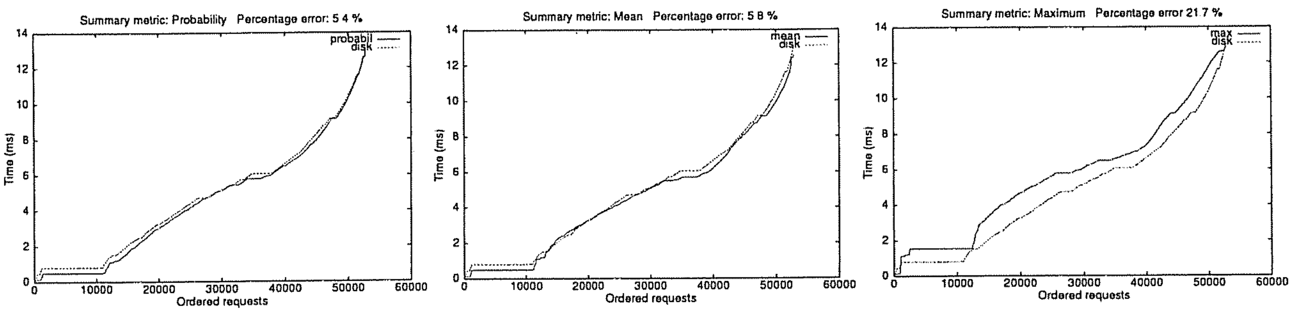


Figure 3: **Demerit Figures for SMTF with Probability, Mean, and Maximum Summary Metrics.** *Each graph shows the demerit figure for a different summary metric. These distributions correspond to the one day from the experiments shown in Figure 2 with a compression factor of 20.*

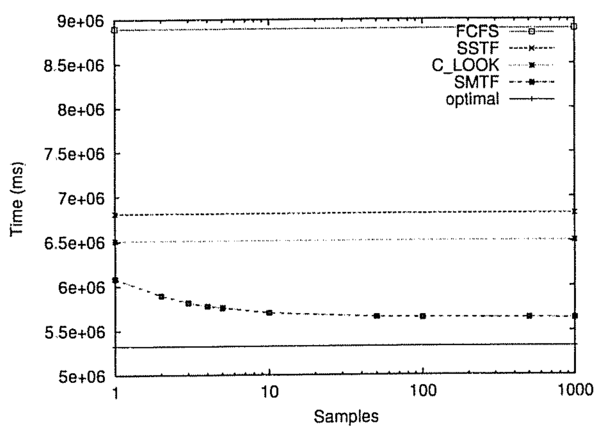


Figure 4: **Sensitivity to Number of Samples.** *The graph shows that the performance of SMTF improves with more samples. The results are on the simulated disk and the week-long HP trace with a compression factor of 20. The x-axis indicates the number of samples used for SMTF. The y-axis shows the time spent at the disk.*

achieves the best performance, even though it does not result in the best demerit figure; we have found that `mean` performs best for all other days from the HP traces we have examined as well. Thus, for the remainder of our experiments, we use the mean of the observed samples as the summary data for each inter-request distance.

5.2 Number of Samples

Given the large variation in times for a single inter-request distance, the Disk Mimic must perform a large number of probe samples to find the true mean of the distribution. However, to reduce the time required to configure the Disk Mimic off-line, we would like to perform as few samples as possible. Thus, we now evaluate the impact of the number of samples on SMTF performance.

Figure 4 compares the performance of SMTF as a function of the number of samples to the performance of FCFS, C-LOOK, SSTF, and optimal. As expected, the performance of SMTF increases with more samples; on this workload and disk, the performance of SMTF continues to improve up to approximately 10 samples. However, most interestingly, even with a single sample for each inter-request distance, the Disk Mimic performs better than FCFS, C-LOOK, and SSTF.

5.3 Interpolation

Although the number of samples performed for each inter-request distance impacts the running time of the off-line probe process, an even greater issue is whether each distance must be explicitly probed or if some can be interpolated from other distances. Due to the large number of potential inter-request distances on a modern storage device (*i.e.*, 2 * number of sectors for both negative and positive distances), not only does performing all of the probes take a significant amount of time, but storing each of the mean values is prohibitive as well. For example, given a disk of size 10 GB, the amount of memory required for the table can exceed 200 MB. Therefore, we explore how some distances can be interpolated without making detailed assumptions about the underlying disk.

To illustrate the potential for performing simple interpolations, we show the mean value as a function of the inter-request distance in Figure 5. The graph on the left shows the mean values for all inter-request distances on our simulated disk. The curve of the two bands emanating from the middle point corresponds to the seek curve of the disk (*i.e.*, for short seeks, the time is proportional to the square root of the distance, whereas for long, the time is linear with

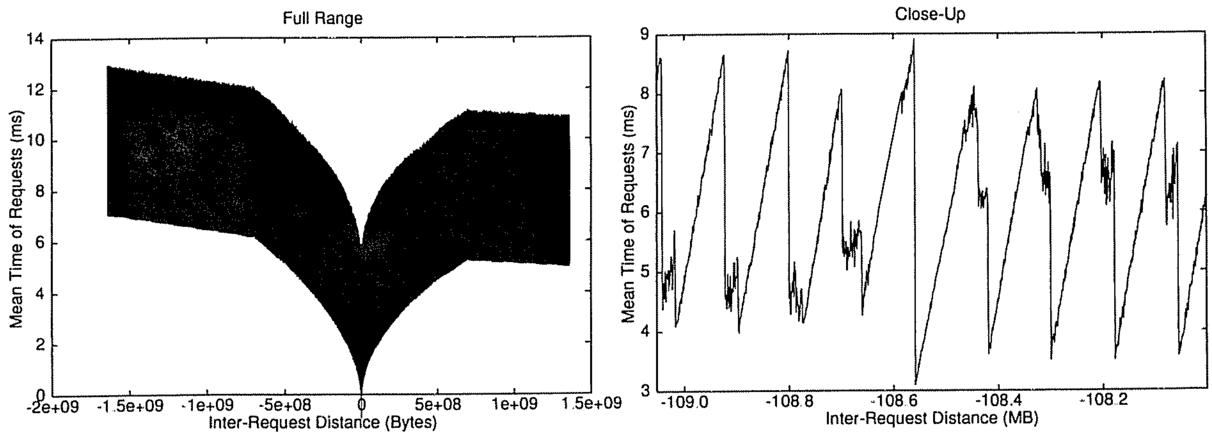


Figure 5: **Mean values for samples as a function of inter-request distance.** *The graph on the left shows the mean time for the entire set of inter-request distances on our simulated disk. The graph on the right shows a close-up for inter-request distances; other distances have qualitatively similar saw-tooth behavior.*

distance); the width of the bands is relatively constant and corresponds to the rotation latency of the disk. The graph on the right shows a close-up of the inter-request distances. This graph shows that the times follow a distinct saw-tooth pattern; as a result, a simple linear model can likely be used to interpolate some distances, but care must be taken to ensure that this model is applied to only relatively short distances.

Given that the length of the linear regions varies across different disks (as a function of the track and cylinder size), our goal is not to determine the particular distances that can be interpolated successfully. Instead, our challenge is determine when an interpolated value is "close enough" to the actual mean such that scheduling performance is impacted only negligibly.

Our basic off-line interpolation algorithm is as follows. After the Disk Mimic performs S samples of two inter-request distances *left* and *right*, it chooses a random distance *middle* between *left* and *right*; it then linearly interpolates the mean value for *middle* from the means for *left* and *right*. If the interpolated value for *middle* is within *error* percent of the probed value for *middle*, then the interpolation is considered successful and all the distances between *left* and *right* are interpolated. If the interpolation is not successful, the Disk Mimic recursively checks the two smaller ranges (*i.e.*, the distances between *left* and *middle* and between *middle* and *right*) until either the intermediate points are successfully interpolated or until all points are probed.

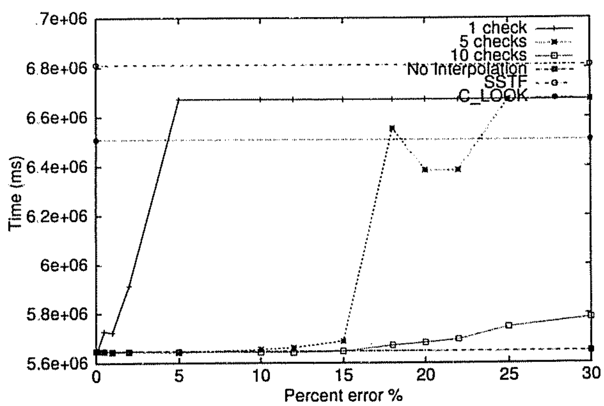


Figure 6: **Sensitivity to Interpolation.** *The graph shows performance with interpolation as a function of the percent of allowable error. Different lines correspond to different numbers of check points, N. The x-axis is the percent of allowable error and the y-axis is the time spent at the disk. These results use the base simulated disk and the week-long HP trace with a compression factor of 20.*

For additional confidence that linear interpolation is valid in a region, we consider a slight variation in which N points between *left* and *right* are interpolated and checked. Only if all N points are predicted with the desired level of accuracy is the interpolation considered successful. The intuition of performing more check points is that a higher error rate can be used and interpolation can still be successful.

Figure 6 shows the performance of SMTF when distances are interpolated; the graph shows the effect of increasing the number of intermediate points N that are checked, as well as increasing the acceptable error, *error*, of the interpolation. We make two

Check Points N	Acceptable Error
1	1 %
2	2 %
3	5 %
4	10 %
5	15 %
10	20 %

Table 2: **Allowable Error for Interpolation.** *The table summarizes the percentage within which an interpolated value must be relative to the probed value in order to infer that the interpolation is successful. As more check points are performed between two inter-request distances, the allowable error increases. The numbers were gathered by running a number of different workloads on the simulated disks and observing the point at which performance with interpolation degrades relative to that with no interpolation.*

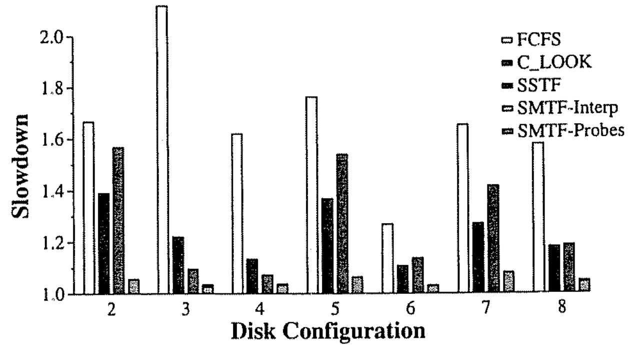


Figure 7: **Sensitivity to Disk Characteristics.** *This figure explores the sensitivity of scheduling performance to the disk characteristics shown in Table 1. Performance is shown relative to greedy-optimal. The bar for SMTF with interpolation is shown stacked on top of the bar for SMTF without interpolation (i.e., all probes).*

observations from this graph.

First, SMTF performance decreases as the allowable error of the check points increases. Although this result is to be expected, we note that performance decreases dramatically with the error not because the error of the checked distances is increased, but because the interpolated distances are inaccurate by much more. For example, with a single check point (*i.e.*, $N = 1$) and an error level of 5%, we have found that only 20% of the interpolated values are actually accurate to that level and the average error of all interpolated values increases to 25% (not shown). In summary, when disk time increases significantly, there was not a linear relationship for the distances between *left* and *right* and interpolation should not have been performed.

Second, SMTF performance for a fixed error increases with the number of intermediate check points N. The effect of performing more checks is to confirm that linear interpolation across these distances is valid. For example, with $N = 10$ check points and *error* = 95%, almost all interpolated points are accurate to that level and the average error is less than 1% (also not shown).

Table 2 summarizes our findings for a wider number of check points. The table shows the allowable error percentage as a function of the number of check points, N, to achieve scheduling performance that is identical to that with all probes. Thus, the final probe process can operate as follows. If the interpolation of one distance between *left* and *right* has an error less than 1%,, it is deemed successful. Otherwise, if two distances between *left* and *right* have errors less than 2%, the interpolation is successful as well. Thus, progressively more check points can be made with higher error rates to be successful. With this approach, 90% of the distances on the disk are interpolated instead of probed, and yet scheduling performance is virtually unchanged; thus, interpolation leads to a 10-fold memory savings.

5.4 Disk Characteristics

To demonstrate the robustness and portability of the Disk Mimic and SMTF scheduling, we now consider the full range of simulated disks described in Table 1. The performance of FCFS, C-LOOK, SSTF, and SMTF relative to greedy-optimal for each of the seven new disks is summarized in Figure 7. We show the performance of SMTF both with interpolation and without, but their performance is nearly identical. As expected, FCFS performs the worst across the entire range of disks; sometimes performing more than a factor of two slower than greedy-optimal. C-LOOK and SSTF perform relatively well when seek time dominates performance (*e.g.*, disks 3 and 4); SSTF performs better than C-LOOK in these cases as well. Finally, SMTF performs very well, even when rotational latency is a significant component of request positioning (*e.g.*, disks 2 and 4). In summary, across this range of disks, SMTF always performs better than both C-LOOK and SSTF scheduling and within 8% of the optimal algorithm.

To show that SMTF can handle the performance variation of real disks, we compare the performance of our implementation of SMTF to that of FCFS, SSTF, and C-LOOK when run on the IBM 9LZX disk. Note that due to the difficulties of performing an optimal simulation of a real disk, we do not compare to greedy-optimal in this case. For our initial results, we consider a synthetic workload in which a 4K request is randomly generated for a disk block between 400 MB and 450 MB; we assume a closed system with 50 outstanding requests. The results of this preliminary investigation are extremely promising: SMTF performs better than all of the other schedulers; specifically, C-LOOK and SSTF perform about 40% slower than SMTF and FCFS performs about 80% slower. We will investigate a more thorough set of workloads in the final version of the paper.

6 On-Line Configuration

In this section, we explore the SMTF scheduler when all configuration is performed on-line. With this approach, there is no overhead at installation time to probe the disk drive; instead, the Disk Mimic observes the behavior of the disk as the workload runs. As in the off-line version, the Disk Mimic records the observed disk times as a function of its inter-request distance, but in this case has no control over the inter-request distances it observes.

6.1 General Approach

For the on-line version, we assume that many of the lessons learned from off-line configuration hold. First, we continue to use the mean to represent the distribution of times for a given inter-request distance. Second, we continue to rely upon interpolation; note that when the Disk Mimic is configured on-line, interpolation is useful not only for saving space, but also for providing new information about distances that have not been observed.

The primary challenge that SMTF must address in this situation is how to schedule requests when some of the inter-request distances have unknown times (*i.e.*, this inter-request distance has not yet been observed by the Disk Mimic and the Disk Mimic is unable to confirm that it can be interpolated successfully). We consider two algorithms for comparison. Both algorithms assume that there is a base scheduler (either C-LOOK or SSTF) which is used when the Disk Mimic does not have sufficient information.

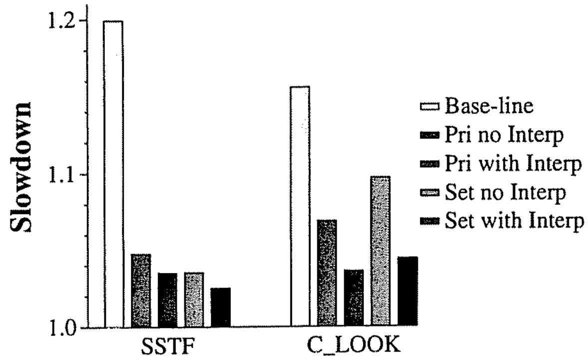
The first algorithm, `Online-Priority`, tries to schedule only those requests for which the Disk Mimic has information. Specifically, `Online-Priority` gives strict priority to those requests in the queue that have an inter-request distance with a known time; among those requests with known times, the request with the minimum mean time is picked. With `Online-Priority`, the base scheduler (*e.g.*, C-LOOK or SSTF) is only used when no inter-request distances for the current queue are known. There are two problems with this approach. First, given its preference for scheduling already known inter-request distances, `Online-Priority` may perform worse than its base scheduler. Second, schedules with a diversity of distances may never be produced and thus the Disk Mimic may never observe some of the most efficient distances.

The second algorithm, `Online-Set`, improves on both of these limitations by using the decision of the base scheduler as its starting point, and scheduling a different request only when the Disk Mimic has knowledge that performance can be improved. Specifically, `Online-Set` first considers the request that the base scheduler would pick. If the time for the corresponding distance is not known by the Disk Mimic, then this request is scheduled. However, if the time is known, then all of the requests with known inter-request distances are considered and the one with the fastest mean is chosen. Thus, `Online-Set` should only improve on the performance of the base scheduler and it is likely to schedule a variety of inter-request distances when it is still learning.

6.2 Experimental Results

To evaluate the performance of the on-line algorithms, we return to the base simulated disk. The leftmost graph of Figure 8 compares the performance of `Online-Priority` and `Online-Set`, when either C-LOOK or SSTF is used as the baseline algorithm and both with and without interpolation. Performance is expressed in terms of slowdown relative to the off-line version of SMTF. We make three observations from this graph.

First, and somewhat more surprising, although C-LOOK performs better than SSTF for this workload and disk, SMTF performs noticeably better



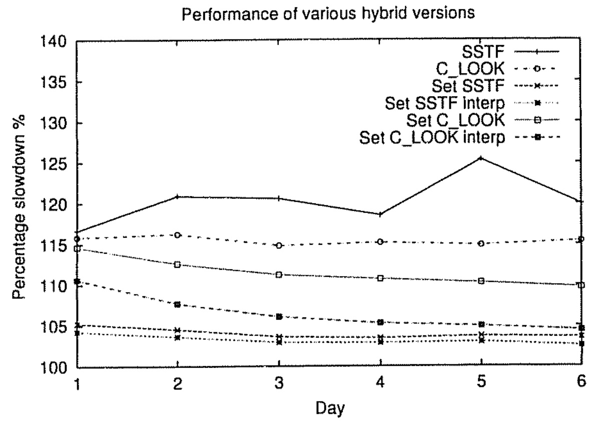


Figure 8: **Performance of On-Line SMTF.** *The first graph compares the performance of different variations of on-line SMTF; the performance of the last day of the week-long HP trace is shown relative to off-line SMTF. The second graph shows that the performance of* `Online-Set` *improves over time as more inter-request distances are observed.*

with SSTF than with C-LOOK as a base; with C-LOOK, the Disk Mimic is not able to observe inter-request distances that are negative (*i.e.*, backward) and thus does not discover distances that are close together. Second, `Online-Set` performs better than `Online-Priority` with SSTF as the base scheduler. Third, although interpolation does significantly improve the performance of `Online-Priority` and of `Online-Set` with C-LOOK, it leads to only a small improvement with `Online-Set` and SSTF. Thus, as with off-line configuration, the primary benefit of interpolation is to reduce the memory requirements of the Disk Mimic, as opposed to improving performance.

The right-most graph of Figure 8 illustrates how the performance of `Online-Set` improves over time as more inter-request distances are observed. We see that the performance of the `Online-Set` algorithms (with and without interpolation) is better than the base-line schedulers of SSTF and C-LOOK even after one day of the original trace (*i.e.*, approximately 150000 requests). The performance of `Online-Set` with SSTF converges to within 3% of the off-line version after four days, or only about 750000 requests.

At this point, we feel that there are two opportunities for further improving the performance of on-line SMTF relative to off-line SMTF. First, in our current on-line implementations, if a slow time for a particular distance is observed initially, the scheduler will avoid that distance even if the mean is much faster. To address this, we plan on requiring that a distance has a minimum number of samples before being classified as known. Second, our current algorithm does not leverage idle time. We plan on performing probes of known inter-request distances during idle times so that the Disk Mimic can learn more of the characteristics of the disk.

7 Related Work

7.1 Disk Modeling

The classic paper describing models of disk drives is that by Ruemmler and Wilkes [16]. The main focus of this work is to enable an informed trade-off between simulation effort and the resulting accuracy of the model. Ruemmler and Wilkes evaluate the aspects of a disk that should be modeled for a high level of accuracy, using the *demerit figure*. Other researchers have noted that additional non-trivial assumptions must be made to model disks to the desired accuracy level [12]; modeling cache behavior is a particularly challenging aspect [22].

Given that the detailed knowledge for modeling disks is not available from documentation, researchers have developed innovative methods to acquire the information. For example, Worthington *et al.*describe techniques for SCSI drives that extract time parameters such as the seek curve, rotation speed, and command overheads as well as information about the data layout on disk and the caching and prefetching characteristics[30]; many of these techniques are automated in later work [17].

Modeling storage devices using tables of past performance has also been explored in previous work; in most previous work [1, 5], high-level system parameters (*e.g.*, load, number of disks, and operation type) are used as indices into the table. Anderson also uses the results on-line, to assist in the reconfiguration of disk arrays. The approach most similar to ours is that of Thornock *et al.* [25]. In this work, the authors use stochastic methods to build a model of the underlying drive. However, the application of this model is to standard, off-line simulation; specifically, the authors study block reorganization, similar to earlier work by Ruemmler and Wilkes [14].

At a higher level, Seltzer and Small suggest *in situ* simulation as a method for building more adaptive operating systems [20]. In this work, the authors suggest that operating systems can utilize in-kernel monitoring and adaptation to make more informed policy decisions. By tracing application activity, the VINO system can determine whether the current policy is behaving as expected or if another policy should be switched into place. However, actual simulations of system behavior are performed off-line, as a "last resort" when poor performance is detected.

7.2 Disk Scheduling

Disk scheduling has long been a topic of study in computer science [28]. Rotationally-aware schedulers came into existence in the early 1990's, through the work of Seltzer *et al.* [19] and Jacobson and Wilkes [11]. However, perhaps due the difficulty of implementation, those early works focused solely upon simulation to explore the basic ideas. Only recently have implementations of rotationally-aware schedulers been described within the literature, and those are crafted with extreme care [9, 31].

More recently, Worthington *et al.* [29] examine the benefits of even more detailed knowledge of disk drives within OS-level disk schedulers. They find that algorithms that mesh well with the modern prefetching caches perform best, but that detailed logical-to-physical mapping information is not currently useful.

Anticipatory scheduling is a recent scheduling development that is complementary to our on-line simulation-based approach [10]. An anticipatory scheduler makes the assumption that there is likely to be locality in a stream of requests from a given process; by waiting for the next request (instead of servicing a request from a different process), performance can be improved. The authors also note the difficulty of building a rotationally-aware scheduler, and instead use a empirically-generated curve-fitted estimate of disk access-time costs; the Disk Mimic would yield a performance benefit over this simplified approach.

8 Conclusions

In this paper, we have explored some of the issues of using simulation within the system to make runtime scheduling decisions; in particular, we have focused on how a disk simulator can automatically model e a range of disks without human intervention. We have shown that the Disk Mimic can model the time of a request by simply observing the logical distance from the previous request and predicting that it will behave similarly to other requests with the same distance in the past. The Disk Mimic can configure itself for a given disk by either probing the disk off-line or, at a slight performance cost, by observing requests sent to the disk on-line. We have demonstrated that a disk scheduler, shortest-mimic'ed-time-first (SMTF), based upon the Disk Mimic, can significantly improve disk performance relative to FCFS, SSTF, and C-LOOK for a range of disk characteristics.

In the future, we plan to show that SMTF scheduling is appropriate for a range of storage devices other than disk drives. For example, RAID systems [13], network-attached storage devices [4], MEMS-based devices [18], tapes [7], and non-volatile memory [27] may all be used as building blocks in a storage system. Each of these devices has its own complex performance characteristics and it would be ideal if the I/O scheduler could automatically adapt to any of these devices.

References

[1] E. Anderson. Simple table-based modeling of storage devices. Technical Report HPL-SSP-2001-04, HP Laboratories, July 2001.

[2] Andrews, Bender, and Zhang. New algorithms for the disk scheduling problem. In *FOCS: IEEE Symposium on Foundations of Computer Science (FOCS)*, 1996.

[3] G. R. Ganger, B. L. Worthington, and Y. N. Patt. The DiskSim Simulation Environment - Version 2.0 Reference Manual. http://citeseer.nj.nec.com/article/ganger99disksim.html.

[4] G. A. Gibson, D. F. Nagle, K. Amiri, F. W. Chang, E. M. Feinberg, H. Gobioff, C. Lee, B. Ozceri, E. Riedel, D. Rochberg, and J. Zelenka. File Server Scaling with Network-Attached Secure Disks. In *Proceedings of the 1997 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 272–284, Seattle, WA, June 1997.

[5] C. Gotlieb and G. MacEwen. Performance of movable-head disk storage devices. *Journal of the Association for Computing Machinery*, 20(4):604–623, 1973.

[6] J. L. Griffin, J. Schindler, S. W. Schlosser, J. S. Bucy, and G. R. Ganger. Timing-accurate Storage Emulation. In *Proceedings of the First USENIX Conference on File and Storage Technologies (FAST '02)*, Monterey, CA, January 2002.

[7] B. Hillyer and A. Silberschatz. On the modeling and performance characteristics of a serpentine tape drive. In *Proceedings of 1996 SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 170–179, 1996.

[8] M. Hofri. Disk scheduling: FCFS vs.SSTF revisited. *Communications of the ACM*, 23(11):645–653, 1980.

[9] L. Huang and T. Chiueh. Implementation of a rotation latency sensitive disk scheduler. Technical Report ECSL-TR81, SUNY, Stony Brook, March 2000.

[10] S. Iyer and P. Druschel. Anticipatory scheduling: A disk scheduling framework to overcome deceptive idleness in synchronous I/O. In *18th ACM Symposium on Operating Systems Principles*, October 2001.

[11] D. M. Jacobson and J. Wilkes. Disk scheduling algorithms based on rotational position. Technical Report HPL-CSP-91-7, HP Laboratories, 1991.

[12] D. Kotz, S. B. Toh, and S. Radhakrishnan. A detailed simulation model of the HP 97560 disk drive. Technical Report TR94-220, Dartmouth College, 1994.

[13] D. A. Patterson, G. Gibson, and R. H. Katz. A case for redundant arrays of inexpensive disks (RAID). *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 17(3):109–116, September 1988.

[14] C. Ruemmler and J. Wilkes. Disk Shuffling. Technical Report HPL-91-156, Hewlett Packard Laboratories, October 1991.

[15] C. Ruemmler and J. Wilkes. Unix disk access patterns. In *Proceedings of the USENIX Winter 1993 Technical Conference*, pages 405–420, 1993.

[16] C. Ruemmler and J. Wilkes. An introduction to disk drive modeling. *IEEE Computer*, 27(3):17–28, 1994.

[17] J. Schindler and G. Ganger. Automated disk drive characterization. Technical Report CMU-CS-99-176, Carnegie Mellon University, November 1999.

[18] S. W. Schlosser, J. L. Griffin, D. Nagle, and G. R. Ganger. Designing computer systems with MEMS-based storage. In *Architectural Support for Programming Languages and Operating Systems*, pages 1–12, 2000.

[19] M. Seltzer, P. Chen, and J. Ousterhout. Disk scheduling revisited. In *Proceedings of the USENIX Winter 1990 Technical Conference*, pages 313–324, Berkeley, CA, 1990.

[20] M. I. Seltzer and C. Small. Self-Monitoring and Self-Adapting Systems. In *Proceedings of the 1997 Workshop on Hot Topics on Operating Systems*, Chatham, MA, May 1997.

[21] P. Shenoy and H. Vin. Cello: A Disk Scheduling Framework for Next-generation Operating Systems. In *Proceedings of the 1998 SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, June 1998.

[22] E. A. M. Shriver, A. Merchant, and J. Wilkes. An analytic behavior model for disk drives with readahead caches and request reordering. In *Proceedings of 1998 SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 182–191, 1998.

[23] N. Talagala, R. H. Arpaci-Dusseau, and D. Patterson. Microbenchmark-based Extraction of Local and Global Disk Characteristics. Technical Report CSD-99-1063, University of California, Berkeley, 1999.

[24] T. J. Teorey and T. B. Pinkerton. A comparative analysis of disk scheduling policies. *Communications of the ACM*, 15(3):177–184, 1972.

[25] N. C. Thornock, X.-H. Tu, and J. K. Flanagan. A Stochastic Disk I/O Simulation Technique. In *Proceedings of the 1997 Winter Simulation Conference*, pages 1079–1086, 1997.

[26] N. C. Thornock, X.-H. Tu, and J. K. Flanagan. A stochastic disk i/o simulation technique. In *Proceedings of the 29th conference on Winter Simulation*, pages 1079–1086, December 1997.

[27] A.-I. A. Wang, P. Reiher, G. J. Popek, and G. H. Kuenning. Conquest: Better Performance Through a Disk/Persistent-RAM Hybrid File System. In *The Proceedings of the USENIX Annual Technical Conference (USENIX '02)*, Monterey, CA, June 2002.

[28] N. C. Wilhelm. An anomaly in disk scheduling: a comparison of FCFS and SSTF seek scheduling using an empirical model for disk accesses. *Communications of the ACM*, 19(1):13–17, 1976.

[29] B. L. Worthington, G. R. Ganger, and Y. N. Patt. Scheduling algorithms for modern disk drives. In *Proceedings of the 1994 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 241–251, Nashville, TN, USA, 16–20 1994.

[30] B. L. Worthington, G. R. Ganger, Y. N. Patt, and J. Wilkes. On-line extraction of SCSI disk drive parameters. Technical Report CSE-TR-323-96, Carnegie Mellon University, 19 1996.

[31] X. Yu, B. Gum, Y. Chen, R. Y. Wang, K. Li, A. Krishnamurthy, and T. E. Anderson. Trading capacity for performance in a disk array. In *Proceedings of the 2000 Symposium on Operating Systems Design and Implementation*, pages 243–258, San Diego, 2000. USENIX Association.