ABSTRACT

Predicting access times is a crucial part of predicting hard disk drive performance. Existing approaches use white-box modeling and require intimate knowledge of the internal layout of the drive, which can take months to extract. Automatically learning this behavior is a much more desirable approach, requiring less expert knowledge, fewer assumptions, and less time. Others have created behavioral models of hard disk drive performance, but none have shown low per-request errors. A barrier to machine learning of access times has been the existence of periodic behavior with high, unknown frequencies. We show how hard disk drive access times can be predicted to within 0.83 ms using a neural network after these frequencies are found using Fourier analysis.

Categories and Subject Descriptors
I.6.5 [Model Development]: Modeling methodologies; D.4.8 [Performance]: Modeling and prediction; D.4.2 [Storage Management]:

General Terms
Performance

Keywords
black-box, storage, modeling, hard disk, machine learning, neural network, Fourier transform

1. INTRODUCTION

Hard disk drive performance models are often used as part of a much larger system simulation [22], for parallel file system simulation [23], for storage configuration [1], and for data placement [15]. The most accurate models are white-box models such as DiskSim [3], which is used by many people [4, 33, 23]. These models require extensive parameterization.

Manufacturers do not release details such as sector layout required by white-box models, which makes parameterizing the models a long and difficult process. In fact, one of us (Oldfield) spent several months configuring DiskSim to model an existing device. Tools such as DIG can extract some of this information [12]. Unfortunately, they require assumptions about the internal structure of the hard disk drive. This structure is likely to change in the future due to the introduction of shingled hard disk drives or other optimizations, as has happened in the past with Zoned Bit Recording, serpentine layouts, etc. Since manufacturers do not release this information, researchers must reverse-engineer a device before modifying DIG and DiskSim to support the new layout.

A more desirable approach is to use machine learning to generate models that can reproduce the behavior with as few assumptions as possible. Some progress has been made in behavioral modeling of hard disk drive performance [17, 5, 30, 33], but none can accurately model individual requests.

Access time has stubbornly resisted efforts to model it, which is what we focus on in this paper. We use workloads that read random single sectors, which minimizes caching and readahead effects. In this case, request latency consists of two components:

1. Queue time — the time the request spends in the device’s queue, waiting to be processed. Requests may queue to some maximum depth in the device, then be serviced out of order.

2. Access time — the time it takes to start reading sector B, given that sector A was just read. This includes seek time, rotational latency, and settle time. Medium and large seeks are relatively easy to model, because the time can be closely approximated by a simple, smooth function of the logical block numbers (LBNs). Settle time can be modeled as a constant and is easily subsumed into the seek model. Small seeks and rotational latency are difficult to model because there are very high-frequency functions in LBN-space.

After decomposing the latency into these two parts, we discard the queue time and focus on the access time, which we approximate as the time the request was completed minus the time the previously completed request was completed.
We tested decision trees and neural nets for access time prediction. One of the complications in the problem is the existence of unknown, high frequency components caused by the rotational aspect of the drive. Unfortunately, a limitation of traditional neural nets and decision trees is their inability to recognize periodic patterns in the data. This can be seen with the checkerboard problem [29] as well as the two-spirals problem [7, 27]. We overcome this limitation by finding the frequencies of these periodic patterns with Fourier analysis and then feeding them into the neural net explicitly by augmenting the feature vector.

2. RELATED WORK

2.1 Predicting request latencies

DiskSim [3] is a well-regarded disk model based on discrete event simulation. It has been validated to produce request-level accuracy. However, it is computationally expensive and difficult to configure for modern disks.

Many analytic models exist, including work by Lebrecht, Dingle, and Knottenbelt [21]. Analytic models are relatively easy to understand and extremely fast due to their compact formulae. Unfortunately, they require very detailed expert knowledge to create. Often, they are limited to certain classes of workloads and are not useful alone in generalized contexts. Our approach shares the last two limitations (for now) but does not require experts.

Kelly et al. describe a black-box probabilistic model [17] similar to table-based models such as the one by Garcia et al. [9]. Requests are categorized based on features including size, LRU stack distance, number of pending reads, number of pending writes, and some RAID-specific information. Table-based models are limited by the table size. The work by Kelly et al. ameliorates the issue by essentially not requiring the entire table to be filled in, but the problem is not fully solved.

Mesnier et al. create a model for relative performance of storage devices [24]. Unfortunately, their approach still requires an accurate base model.

A popular approach is to use regression trees to predict response time. Dai et al. predict performance with a combination of regression trees and support vector regression [5]. However, their models are workload-specific, and their prediction errors are based on one-second averages rather than per-request latencies. Wang et al. also calculate errors based on windows, using one-minute windows in their case [30].

None of the machine-learning-based approaches have shown low per-request errors.

2.2 Learning periodic functions

Neural nets can be trained on periodic functions in many ways. Some setups predict the value of the function directly from the input. These invariably use a fixed interval with a very small number of oscillations [14, 13, 26, 8, 32]. Extrapolating beyond the training range leads to poor performance [19]. This approach is infeasible for our problem because the number of oscillations (roughly, the number of tracks) is on the order of a million.

If the function is known to have period \( p \), another approach is to map the input \( x \) into the range \((0,p)\) using \( x \mod p \). Common examples include time-of-day or day-of-year inputs for functions that are daily or yearly periodic [20, 6]. An alternative is to map it to \( \sin(2\pi x/p) \) and \( \cos(2\pi x/p) \) [10]. (This pair of functions has nice properties; they are both continuous and bounded, and weighted sums are equal to other sinusoids with varying phase.) Either way, this approach requires the period to be known precisely ahead of time, and that the input be exactly periodic.

Neurons in a neural net calculate their output by taking a weighted sum of the inputs and applying an activation function or transfer function, typically \( \tanh(x) \) or \( \frac{1}{1+e^{-x}} \). Rather than using one of these as the activation function, one may use \( \sin(x) \). Unfortunately, \( \sin(x) \) does not approach a limit for large \( x \), while \( \tanh(x) \) and \( \frac{1}{1+e^{-x}} \) do. Lack of a such a limit can lead to instability [31]. Periodic activation functions also introduce many local minima [28].

A powerful tool for time series data is recurrent or delay networks [25, 16]. These feed the network back into itself, so that the network predicts \( y \) values from other \( y \) values, rather than from \( x \) values. This is usually applied to problems with one dimension of recursion, but ours has two, one for the previously accessed sector and one for the current sector to access. With dense data, only one level of recursion is necessary, because the other \( y \) values are already known. If the data is sparse, missing values must be recursively computed. Our sampling is (by necessity) so sparse that the recursion would be extremely deep, making computation time impractical.

3. FOURIER ANALYSIS

We assume that the access time function may have components that are locally periodic. This is a fairly lax assumption. First of all, it is well-known that these functions do have locally periodic components for existing hard drives, due to track lengths and track skews that are constant within a serpentine. Periodicity is likely to occur in other devices as well, because of the benefits of regular repetition in designs. Secondly, as described in section 3.2 the cost is minor if these functions do not have locally periodic components.

3.1 Finding strong frequencies

We refer to the previously accessed sector as the start sector, which is the current position of the head, and the first sector of the current request as the end sector, which will be the new position of the head. Let \( f(a,b) \) be the access time function, where \( a \) and \( b \) are the LBNs of the start and end sectors, and \( f \) returns the access time in milliseconds. To find periodic behavior in \( f \), the Fourier transform is an obvious place to start. It can only find periods that are directly correlated to the value of \( f \), but it is much faster than training a neural net for every period to see which ones are useful. Furthermore, if many periods are useful, their individual impact on the neural net’s accuracy may be swamped by noise. Using the Fourier transform allows us to pinpoint useful periods relatively quickly and with high precision.

Capturing the access time for every sector pair takes a very long time. Given our test device’s mean access time of 15.5 ms and 976,773,168² pairs of sectors, the time to capture all of them would be roughly 469 million years. Obviously, this is infeasible, so we are limited to a very sparse sampling.

The sparsity means that we cannot use the Fast Fourier Transform, so we fall back to the brute force method of calculation. Let the vectors \( x = (a,b) \) and \( \xi = (u,v) \), where \( u \) and \( v \) are coordinates in frequency space. Further define
$N$ as the number of data points (in our case, the number of requests in the trace), $M$ as the number of frequencies to search, and $\hat{f}$ as the Fourier transform of $f$. The discrete Fourier transform is defined:

$$\hat{f}(\xi) = \frac{1}{N} \sum_{k} f(x_k) e^{-2\pi i x_k \xi}$$

which takes $O(N)$ time to calculate for a single frequency vector $\xi$, so calculating $\hat{f}$ for $M$ frequencies takes $O(MN)$ time. Scanning all frequencies in the 2D space leads to infeasible computation time for large datasets or datasets over larger regions of the drive.

Note that the access time depends mostly on the difference between the sectors. This is intuitively true, as the time to move from sector 0 to sector 9 should be roughly the same as the time to move from sector 1 to sector 10. This causes stripes in the access time function along $a = b$ (see fig. 4a), and the Fourier transform of a function with stripes has strong components in the direction orthogonal to the stripes [11], which means that strong components of the access time function should lie on $u = -v$. We see this empirically in fig. 1. By searching only this diagonal instead of the entire space, the computation time becomes feasible.

Figure 1: Full 2D Fourier spectrum for the first $K = 237,631$ sectors out to $200 \times 200 = 8.42 \times 10^{-4} \times 8.42 \times 10^{-4}$, which corresponds to periods of at least $1/(8.42 \times 10^{-4}) = 1188,155$ sectors. Plot is clipped to magnitude 1 to show detail, but central spike goes up to 8.6, and other diagonal spikes go up to 3.9. This figure shows that strong frequencies do lie on the diagonal $v = -u$.

Another way to approach the Fourier analysis, which we did not test, is to use a search-based method. Assuming the number of strong frequencies is small, this should be much faster, roughly $O(kN \log M)$ for $k$ strong frequencies. Unfortunately, these methods require performing convolutions, and it is not clear if performing a convolution on sparse data is possible or meaningful.

Off-diagonal frequencies occur, but they are mostly related in a straightforward fashion to frequencies on the diagonal. An example would be a dataset that includes track sizes of 10 and 11. Strong coefficients are likely at periods (10, 10), (10, 11), (11, 10), and (11, 11). Since the diagonal provides (10, 10) and (11, 11) (or just 10 and 11 in 1D), the locations of the others could be inferred, although that is not actually necessary.

To determine a threshold for strong frequencies, we sample 1000 frequencies at random and calculate $|\hat{f}|$. The threshold is then set to the mean plus six standard deviations. Assuming the magnitudes are normally distributed, this means that a frequency has roughly a 1 in 500 million chance of being spuriously flagged.

We scan across $v = -u$ with a step size of 0.1/blockCount looking for local maxima above the threshold. The peaks were often not centered on integer multiples of 1/blockCount, which is why the step size is not 1/blockCount. After finding a maximum, we perform a local search to fine-tune its position.

Due to structures such as serpentine, $f$ may have higher-order periods. In other words, $f$ may have period $p_1$ for a subrange, then $p_2$, then $p_1$, then $p_2$, etc., with the switch between $p_1$ and $p_2$ being periodic. Currently, we have no method for detecting these explicitly, although we have seen them show up as interesting periods in their own right. These actually cause further problems by reducing the utility of the lower-level periods. The multiple regions using period $p$ may be out of phase, causing $|\hat{f}|$ to drop. Essentially, this is a form of mixed interference which results in a weaker signal than if the signals interfered purely constructively.

### 3.2 Input augmentation

Once all interesting periods have been found, the input vectors for the neural net are augmented. Given an input $(a, b)$ and periods $p_1, \ldots, p_k$, the augmented input vector is $(a, \cos(2\pi a/p_1), \sin(2\pi a/p_1), \ldots, \cos(2\pi a/p_k), \sin(2\pi a/p_k), b, \cos(2\pi b/p_1), \sin(2\pi b/p_1), \ldots, \cos(2\pi b/p_k), \sin(2\pi b/p_k))$. We use sinusoids of $a$ and $b$ separately rather than sinusoids of $b-a$ because the period changes for each input separately. For example, if $a$ is in a region where $p_1$ dominates, and $b$ is in a region where $p_2$ dominates, then $\sin(2\pi a/p_1)$ and $\sin(2\pi b/p_2)$ are useful, but $\sin(2\pi (b-a)/p_3)$ is not likely to be useful for any period $p_3$.

If no interesting periods are found, then the input vectors are unchanged. This means that for devices with no periodicity, the cost of the periodicity assumption is just the time to search for periods. The neural net is unchanged, and therefore the speed and accuracy of the model is unchanged.

### 4. SHARED WEIGHTS

A useful and device-independent assumption is that all sectors map from LBNs to geometrical space in the same way. This is encoded into the neural net by applying weight sharing across two subnets (see fig. 2). This means that the weights in each neuron in subnet 1 are identical to the weights in the corresponding neuron in subnet 2. This is essentially a very simple convolutional neural net with only two instances of the convolving subnet and no overlap of inputs. A similar approach was used by Kindermann et al. for solving functional equations with neural nets [18]. Another interpretation is that the subnets perform feature generation, knowing that the two sectors are instances in the same input space.

More formally, the net assumes $f$ can be expressed as

$$f(a, b) = h(g(a), g(b))$$

where $g$ (the subnet) maps from LBNs to geometry, and $h$ (the main net) maps from geometry to access time. Reuse of $g$, which is equivalent to weight sharing in the neural
network, formalizes the assumption that all sectors map to geometry in the same way. Technically, \( f \) is not restricted as long as the intermediate space is at least as large as the input space, since one could always use \( g(x) = x \) and \( h = f \), but this predisposes the neural net to learn more useful decompositions.

---

Figure 2: Network architecture when subnets are used. Note that the weights in subnet 1 are identical to the weights in subnet 2.

Each subnet has 32 hidden and 10 output units, and the main net has 10 hidden units. All neurons use a sigmoidal activation function except the final output, which is linear. This is common practice for neural nets performing regression [8]. Note that when 6 periods are used, the number of parameters is nearly equal with or without subnets, so the configurations should have the same expressive power.

We know of no analogous operation for decision trees, so they were unmodified in this respect.

5. EXPERIMENTAL SETUP

The device we modeled is a Western Digital Caviar Black WD5002AALX. It has a capacity of 500GB (976,773,168 sectors), rotational speed of 7200 RPM, cache size of 32 MB, and NCQ queue length of 32. The drive was connected with SATA. The host runs 64-bit Ubuntu Linux, kernel version 2.6.35-28-server. It has an Intel Core i7 quad-core CPU (plus hyper-threading) running at 2.80 GHz and 12 GB of RAM. Block-level traces were captured using blktrace. Request latency was defined to be the D2C or device-to-completion time. This is the time between the OS IO scheduler sending the request to the device driver and receiving a response from the device driver. Excluding device driver times is difficult and requires a hardware setup. In modern computers, the time spent in the device driver should be negligible compared to the time spent in the hard drive.

The dataset is of \( N = 32,000 \) random reads of the first 94 tracks, or 237,631 sectors, which is the first 0.024% of the drive. This corresponds to the first part of the first serpentine, so all tracks have the same size (2528 sectors).

Decision trees were tested with Weka. In all cases, the minimum leaf weight (minNum) was set to 0. Bagging, a popular ensemble method used to improve the accuracy of decision trees, was also tested. Bagging generates many decision trees trained on random subsets of the data, then averages their predictions [2]. When using bagging, the number of decision trees was set to 100.

We trained and tested each configuration five times and reported the mean and standard deviation. For tests with random periods, each run used different random periods between 0 and 5000 sectors.

---

6. RESULTS

For reference, we ran DIG on our test hard disk drive. From the DIG data, the track length at the beginning of the drive is 2528 sectors, and the skew is 1.1908 ms, which corresponds to 361.37 sectors (given the rotation time of 8.33 ms). We expected to see a dominant period of 2528 – 361.37 = 2166.63 sectors, which is a frequency of 1/2166.63 = 4.62 \cdot 10^{-4}.

From the Fourier analysis, the actual dominant period is 1/(4.52 \cdot 10^{-4}) = 2212.99 sectors, which is close to our prediction of 2166.63.

Access time errors are listed in table 1. The first entry is the error for a model that always predicts the mean value of the dataset.

When using decision trees, adding the sines and cosines as additional features reduced the error noticeably. Bagging the decision trees reduced error further, but only when the periodic information was included. The lowest error achieved with decision trees was 1.1906 ms.

When using neural nets, adding the sines and cosines reduced the error significantly. Weight sharing reduced the error further by a small but noticeable amount. The lowest error achieved with neural nets was 0.830 ms.

Most of the time, neural net tests with random periods fared worse than tests with no period information. However, one test with random periods performed much better, although not as well as tests with the correct period information. This appears to have been caused by one of the random periods being nearly equal to twice the most important period.

When comparing the actual and predicted access time functions (fig. 4), we see that the shape is matched very well. Even the notches in the stripes are in the predicted locations.

7. CONCLUSIONS
Neural nets achieved lower error than decision trees in all tests. We suspect this is caused by two things: 1) the structure imposed by weight sharing cannot be implemented with decision trees, and 2) many important features are individually uncorrelated with the output, which decision trees do not learn from well.

Adding periodic features requires only a mild assumption and improves multiple machine learning algorithms significantly. This approach is likely to be a crucial part of future behavioral modeling of storage device performance.

8. FUTURE WORK

The analysis needs to be extended to an entire hard disk drive. Other types of devices also need to be tested, such as SSDs and RAID arrays.

One important large assumption made by this approach is that access time is a meaningful and useful concept. For other block storage devices, this may not be the case, and removing this assumption would be useful future work. Generalizing, we may predict the latency (not access time) of a request given \( k \) previous requests. The neural net is easily extended to predict \( f(x_1, x_2, \ldots, x_k) = h(g(x_1), g(x_2), \ldots, g(x_k)) \) in time linear to \( k \). The difficulty lies in finding useful periods to feed the neural net. One possibility is to generalize the diagonal search. Previously, we searched \( u = -v \). Letting \( \xi = (u, v) \), this can be expressed as a line through the origin normal to \( c \cdot \xi = 0 \) with \( c_1 = 1, c_2 = -1 \). An intuitive expansion is all lines through the origin normal to some hyperplane \( c \cdot \xi = 0 \) where \( c_i \in \{-1, 0, 1\} \). This yields \( 3^k \) lines, but this is still far better than searching an entire \( k \)-dimensional space.

9. ACKNOWLEDGEMENTS

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.

10. ADDITIONAL AUTHORS

Additional authors: Patrick Widener (Sandia National Laboratories, email: pwidene@sandia.gov).
11. REFERENCES


