A Performance Model of Zoned Disk Drives with I/O Request Reordering

Abigail S. Lebrecht Nicholas J. Dingle William J. Knottenbelt

Department of Computing, Imperial College London,
180 Queen's Gate, London SW7 2BZ, United Kingdom.
{asl102,njd200,wjk}@doc.ic.ac.uk

***Abstract*—Disk drives are a common performance bottleneck in modern storage systems. To alleviate this, disk manufacturers employ a variety of I/O request scheduling strategies which aim to reduce disk head positioning time by dynamically reordering queueing requests. An analytical model of this phenomenon is best represented by an $M/G/1$ queue with queue length dependent service times. However, there is no general exact result for the response time distribution of this variety of queue with generalised service time distributions. In this paper, we present a novel approximation for the response time distribution of such a queue. We then apply this method to the specific case of a zoned disk drive which implements I/O request reordering. A key contribution is the derivation of realistic service time distributions with minimised positioning time. We derive analytical results for calculating not only the mean but also higher moments and the full distribution of I/O request response time. We validate our model against measurements from a real disk to demonstrate the accuracy of our approximation.**

I. Introduction

Over the past two decades disk drive performance improvements have significantly lagged behind all other system component performance enhancements [1]. Moreover the dramatic increase in capacity without a corresponding increase in bandwidth has made disk drives the greatest performance bottleneck in many storage systems. To address this issue, I/O request scheduling algorithms have been developed that attempt to minimise disk head positioning time [2], [3], [4], [5]. Disk head positioning time consists of seek time and rotational latency. Seek time is the time it takes the disk head to settle over the correct track containing the target sector. Rotational latency is the time the disk takes to rotate the target sector under the disk head. The best way to minimise the total response time of all queueing I/O requests is to dynamically reorder them so that the next request chosen to be serviced has the lowest disk head positioning time of all queueing requests. With this strategy employed, as queue lengths increase response times do not suffer excessively since service times are reduced.

In this paper we present an analytical queueing model of a zoned[1] disk drive with this type of scheduling. We base our model on an existing zoned disk model [6], [7] in which each disk drive is represented as a first-come first-served (FCFS) $M/G/1$ queue with a fixed service time distribution. The present work models the operation of a disk drive with Shortest Access Time First (SATF) scheduling by using an $M/G/1$ queue with queue-length dependent service time distributions. There does not currently exist a generally applicable exact result for the response time distribution of this variety of queue. We present a novel approximation for the response time distribution of such a queue. It is a non-trivial challenge to derive realistic service time distributions for each queue length such that expected positioning time is minimised. We demonstrate the accuracy of our model by comparing model predictions with real device measurements.

The remainder of this paper is organised as follows. In Section II we discuss prior work in this area. We survey existing scheduling strategies, modelling techniques and existing approaches for modelling queues with state-dependent service times. We also briefly recap the existing zoned disk model. In Section III we present a new approximation for calculating the response time distribution of $M/G/1$ queues with state-dependent service times. We then apply this method to the zoned disk model, deriving a queue length dependent service time distribution for the disk drive. Section IV validates our model against real device measurements. Finally, Section V concludes and considers directions for future work.

II. Background

A. Disk Head Positioning Optimisation

Bursty workloads can result in long queues of pending I/O requests [8]. In such circumstances, it is the role of the disk scheduler to reorder requests to minimise disk head positioning time. This reduces the time needed to service each request which inevitably reduces overall request response times [9]. There exist many possible scheduling algorithms to choose the order in which requests are serviced.

The Shortest Seek Time First (SSTF) algorithm minimises track-to-track seek time only. SSTF can be implemented using the SCAN algorithm [3] in which requests are serviced in order of the disk cylinder number in a particular direction. The main drawback of SSTF is that it does not consider rotational latency; the latter makes up an increasing proportion of disk

[1]On modern hard drives there are more blocks on cylinders on the outside of the platter than those closer to the centre. Cylinders with the same number of blocks are grouped together in zones. Disks rotate with a constant angular velocity and so data throughput is higher for outer zones than for inner ones.

head positioning time as recent advances in disk technology have shortened seek times significantly, while rotational speeds have increased only slightly [4]. To address this, Jacobson and Wilkes [4] and Seltzer et al. [5] introduce Shortest Access Time First (SATF), where access time is disk head positioning time. This strategy introduces the possibility that certain requests can suffer from starvation. The Aged Shortest Access Time First (ASATF) algorithm avoids this by basing ordering on a metric that takes into account the amount of time that a request has been queueing.

Worthington et al. [10] carry out a simulation study of FCFS, SSTF and SATF and resolve that SATF provides the best overall performance, and that FCFS can yield particularly poor performance. Burkhard and Palmer [2] present an SATF-like scheduling algorithm for optimising positioning time that takes into account the fact that an aggressive head movement may fail to settle in time to read from the target sector. In this case, the disk must complete a full rotation before data transfer can begin. The probability of this occurring is known as the miss probability, and is drive-dependent. Seagate disks implement Rotational Positioning Ordering (RPO) using Native Command Queueing (NCQ) which aims to optimally re-order commands to maximise performance [11].

B. Models of Disk Head Positioning Optimisation

Modelling response times for disks with minimised disk head positioning time is analytically difficult, and hence there do not exist many analytical models and none for zoned disk drives. Chen et al. [12] present a model for a scheduling algorithm that only minimises seek time. Shriver et al. [13] define the distance (in terms of number of bytes) between two random requests with minimised positioning time as

$$\frac{\textit{no\_of\_Cylinders} \times \textit{Bytes\_per\_Cylinder}}{E[\textit{Queue\_length}] + 2}$$

However, this is not applicable in the context of zoned disks since *Bytes_per_Cylinder* is not constant. The most comprehensive existing analytical performance model including queue re-ordering is that of Gotlieb and MacEwen [14]. However, this only models SSTF, not SATF. They use the theory of state-dependent queues in their model, whereby the service time distribution can depend on queue length at the start of a service. This work is primarily based on that of Harris [15].

There are a number of studies of $M/M/1$ queues with state-dependent service times, including those by Harris [15] and Morrison [16]. A number of other studies consider the simpler case of two service time states [17], [18], [19]. Brill and Posner [17] allow for different service rates depending on whether or not there are customers queueing behind a request at the start of service. Gray and Wang [19] study the case in which the service rate changes when the queue length exceeds a given number (N) and then changes back when the queue length is less than K ($K \leq N$).

We note that no general result exists for response time in $M/G/1$ queues with state-dependent service times.

C. Zoned Disk Model

The service time density of an access to a random location on a single disk drive is the convolution of the seek time, rotational latency and data transfer time probability density functions. An important subtlety that needs to be taken into account is that modern disks are *zoned*, with more sectors on the outer tracks than inner tracks. Therefore, a random request is more likely to be directed to a sector on an outer track. Similarly, zoning means that it is faster to transfer data on a track close to the circumference than the centre of the disk. The seek time and data transfer models must take these factors into account.

In our model we use the seek time and rotational latency probability distributions defined in [7] and the data transfer time distribution from [6]. We denote the random variables of seek time, rotational latency and k-block transfer time as S, R and T_k respectively. We represent a disk as an $M/G/1$ queue. A full description of the derivation of these probability distributions is included in Appendix A.

III. State-Dependent Service Times for an M/G/1 queue

A. Theory

In an $M/G/1$ queue with state-dependent service times, we assume that from time $t = 0$, customers $C_0, C_1, \ldots, C_n, \ldots$ arrive at the queue. Let X_n denote the queue length immediately after customer C_n has completed service, and let Z_n denote the number of customers that arrive in the queue during the service of customer C_{n+1}. Then,

$$X_{n+1} = \begin{cases} X_n - 1 + Z_n & X_n > 0 \\ Z_n & X_n = 0 \end{cases}$$

Given state-dependent service times, the number of arrivals during a service period, Z_n, is dependent on the service time, which itself is dependent on the queue length at the start of customer C_{n+1}'s service, X_n. Given i requests in the queue at the start of service, we denote the service time by the random variable B_i. Since arrivals are Markovian with arrival rate λ, the probability of j arrivals in a pre-defined service period x is:

$$P(Z_n = j \mid B_{X_n} = x) = \frac{(\lambda x)^j}{j!} e^{-\lambda x} \quad (j \geq 0)$$

Therefore, by the law of total probability, the probability of j arrivals during a service period given all possible service times for a queue length of i at the start of a service is:

$$p_{j,i} = P(Z_n = j \mid X_n = i) = \int_0^\infty \frac{(\lambda x)^j}{j!} e^{-\lambda x} dF_{B_i}(x) \quad (1)$$

The probability generating function for Z_n given a queue length of i at the start of a service is:

$$G_i(z) = \int_0^\infty e^{\lambda x z} e^{-\lambda x} dF_{B_i}(x) = B_i^*[\lambda(1 - z)] \quad (2)$$

where B_i^* is the Laplace-Stieltjes Transform (LST) of B_i.

The embedded Markov chain of queue population has transition matrix $Q = (q_{ij} \mid i, j \geq 0)$ where:

$$\begin{aligned} q_{0j} &= P(X_{n+1} = j \mid X_n = 0) = p_{j,1} \\ q_{ij} &= P(X_{n+1} = j \mid X_n = i) \\ &= \begin{cases} p_{j-i+1,i} & j \geq i-1 \geq 0 \\ 0 & 0 \leq j \leq i-2 \end{cases} \end{aligned}$$

$$Q = \begin{bmatrix} p_{0,1} & p_{1,1} & p_{2,1} & p_{3,1} & \cdots \\ p_{0,1} & p_{1,1} & p_{2,1} & p_{3,1} & \cdots \\ 0 & p_{0,2} & p_{1,2} & p_{2,2} & \cdots \\ 0 & 0 & p_{0,3} & p_{1,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The steady-state equations for the Markov chain, $\pi = \pi Q$ are:

$$\pi_j = \pi_0 p_{j,1} + \sum_{i=1}^{j+1} \pi_i p_{j-i+1,i} \tag{3}$$

where π_i is the steady-state probability of there being i requests in the queue (including the customer currently in service).

Then the queue length generating function $\Pi(z) = \sum_{i=0}^{\infty} \pi_i z^i$, if it exists, is [15]:

$$\begin{aligned} \Pi(z) &= \pi_0 \sum_{j=0}^{\infty} p_{j,1} z^j + \sum_{j=0}^{\infty} \sum_{i=1}^{j+1} \pi_i p_{j-i+1,i} z^j \\ &= \pi_0 \sum_{j=0}^{\infty} p_{j,1} z^j + \sum_{j=0}^{\infty} \pi_1 p_{j,1} z^j + \\ &\quad \sum_{j=1}^{\infty} \pi_2 z p_{j-1,2} z^{j-1} + \sum_{j=2}^{\infty} \pi_3 z^2 p_{j-2,3} z^{j-2} + \ldots \\ &= \pi_0 G_1(z) + \frac{1}{z} \sum_{i=1}^{\infty} \pi_i z^i G_i(z) \end{aligned} \tag{4}$$

This is dependent on the chain being stationary, the condition for which is that $\Pi(1) = 1$ [20]. Since the $G_i(z)$ are all probability generating functions, $\forall i \; G_i(1) = 1$ and

$$\Pi(1) = \pi_0 + \sum_{i=1}^{\infty} \pi_i$$

By definition of the steady-state probabilities, $\sum_{i=0}^{\infty} \pi_i = 1$, hence $\Pi(1) = 1$.

Using an approach similar to the derivation of $G_i(z)$ in Equation (2), it can be observed that $\Pi(z)$ is related to the response time, W as follows [20]:

$$\begin{aligned} \Pi(z) &= \int_0^{\infty} e^{\lambda x z - \lambda x} dF_W(x) \\ &= W^*[\lambda(1-z)] \end{aligned} \tag{5}$$

Hence, by substituting into Equation (4),

$$W^*(\theta) = \pi_0 B_1^*(\theta) + \frac{\lambda}{\lambda - \theta} \sum_{i=1}^{\infty} \pi_i \left(\frac{\lambda - \theta}{\lambda} \right)^i B_i^*(\theta) \tag{6}$$

In practice we would need to know the service time distribution for all possible queue lengths to be able to apply this equation. An elegant simplification that eradicates this problem assumes that if the queue length is greater than or equal to a specified length n then all corresponding service times are represented by the random variable B_n. This is an increasingly accurate approximation when there is a relatively low probability of high queue lengths or if the service time distributions are similar for higher queue lengths. Then,

$$\begin{aligned} \Pi(z) &= \pi_0 G_1(z) + \frac{1}{z} \sum_{i=1}^{n-1} \pi_i z^i G_i(z) + \\ &\quad \frac{1}{z} \sum_{i=n}^{\infty} \pi_i z^i G_n(z) \\ &= \pi_0 G_1(z) + \frac{1}{z} \sum_{i=1}^{n-1} \pi_i z^i G_i(z) + \\ &\quad \frac{1}{z} G_n(z) (\Pi(z) - \sum_{i=0}^{n-1} \pi_i z^i) \end{aligned}$$

$$= \frac{z\pi_0 G_1(z) + \sum_{i=1}^{n-1} \pi_i z^i G_i(z) - G_n(z) \sum_{i=0}^{n-1} \pi_i z^i}{z - G_n(z)}$$

We need to ensure that $\Pi(1) = 1$ to fulfil the stationary condition. Using L'Hôpital's rule to find the limit as $z \to 1$, it becomes apparent that in order for $\Pi(1) \to 1$ as $z \to 1$, the following equation must hold:

$$\pi_0 = \frac{1 - \lambda E[B_n] - \sum_{i=1}^{n-1} \pi_i (\lambda E[B_i] - \lambda E[B_n])}{1 + \lambda E[B_1] - \lambda E[B_n]} \tag{7}$$

Solving the set of linear equations arising from Equations (3) and (7), the queue length probabilities $\pi_0, \pi_1, \ldots, \pi_n$ can be calculated.

The response time LST can be calculated using Equation 5:

$$\begin{aligned} &W^*(\theta) = \\ &\frac{1}{\lambda(1 - B_n^*(\theta)) - \theta} \Bigg(\pi_0 ((\lambda - \theta) B_1^*(\theta) - \lambda B_n^*(\theta)) + \\ &(\lambda - \theta) \sum_{i=1}^{n-1} (\pi_i \left(\frac{\lambda - \theta}{\lambda} \right)^{i-1} (B_i^*(\theta) - B_n^*(\theta))) \Bigg) \end{aligned}$$

By differentiating this equation m times and evaluating at $\theta = 0$, a recurrence relation for moments of response time can be derived:

$$\begin{aligned} E[W^m] &= \frac{1}{(m+1)(1 - \lambda E[B_n])} \Bigg(\pi_0 \left(\lambda E[B_1^{m+1}] \right. \\ &\quad \left. + (m+1) E[B_1^m] - \lambda E[B_n^{m+1}] \right) \\ &\quad + \sum_{i=1}^{n-1} \pi_i \lambda \sum_{j=0}^{\min[i, m+1]} \binom{m+1}{j} \binom{i}{j} \frac{j!}{\lambda^j} \\ &\quad \left(E[B_i^{m+1-j}] - E[B_n^{m+1-j}] \right) + \\ &\quad \lambda \sum_{j=2}^{m+1} \binom{m+1}{j} E[B_n^j] E[W^{m+1-j}] \Bigg) \end{aligned} \tag{8}$$

B. Application to Zoned Disk Model

In the case of RPO, we define service time as the minimum disk head positioning time of all queueing I/O requests plus any additional rotations needed if the head fails to settle in time to read from target sectors. The probability that the disk head misses the correct rotational position at the end of a seek (termed a *latency miss*) is denoted as p_{miss}. If there are i requests in the queue immediately prior to the start of a service, the service time of a request is thus:

$$B_i = \min_{l=1,\ldots,i}(S_l + R_l) + p_{miss}R_{max} + T_k$$

where R_{max} is the time to complete a complete disk revolution and S, R and T_k are seek time, rotational latency and k-block data transfer time respectively. In order to calculate the probability distribution of B_i we employ order statistics [21]. We find the first order statistic (i.e. minimum) of i convolutions of seek time and rotational latency ($S + R$). If a set of independent and identically distributed (iid) random variables, $X_1, X_2, \ldots, X_i$ are ordered in terms of size, the cumulative distribution function (cdf) of the smallest, $X_{(1)}$, will be:

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \leq x) = 1 - P(X_{(1)} > x) \\ &= 1 - \forall j P(X_{(j)} > x) \quad j = 1, 2, \ldots, i \\ &= 1 - \forall j (1 - P(X_{(j)} \leq x)) \\ &= 1 - (1 - F_X(x))^i \end{aligned}$$

In our case X is $S + R$ which has a convolved cdf of:

$$\begin{aligned} F_{R+S}(x) &= \frac{1}{R_{max}} \int_0^{R_{max}} F_S(x - z)dz \\ &= \frac{1}{R_{max}} \int_{x-R_{max}}^{x} F_S(u)du \end{aligned}$$

The probability density function (pdf) of a random variable M that models the occurrence of a latency miss, based on a single Bernoulli trial, is:

$$f_M(x) = \begin{cases} 1 - p_{miss} & x = 0 \\ p_{miss} & x = R_{max} \\ 0 & \text{otherwise} \end{cases}$$

It should be noted that the latency miss is only present when RPO is switched on. Since for the case $n = 1$ there is no queue re-ordering, there will be no latency misses. If the convolved minimum positioning time and transfer time have density function $f_{Y_i}(x)$ then convolving $f_{Y_i}(x)$ with $f_M(x)$ yields

$$f_{B_i}(x) = \begin{cases} f_{Y_i}(x) & i = 1 \\ (1 - p_{miss})f_{Y_i}(x) & \\ +p_{miss}f_{Y_i}(x - R_{max}) & i > 1 \end{cases} \quad (9)$$

Here x is bounded between the minimum transfer time, and the sum of maximum seek time, maximum latency (which is the time to complete two full disk revolutions) and maximum transfer time, irrespective of how much request reordering occurs.

Using Equation (8) the mean, variance and further moments of response time can be calculated. In order to do this it must be noted that the mth moment of service time is

$$E[B_i^m] = \begin{cases} E[Y_i^m] & i = 1 \\ (1 - p_{miss})E[Y_i^m] + & \\ p_{miss}\sum_{j=0}^{m}\binom{m}{j}E[Y_i^j]R_{max}^{j-i} & i > 1 \end{cases}$$

where

$$E[Y_i^m] = \sum_{j=0}^{m}\binom{m}{j}E[((R + S)_i)^j]E[T_k^{m-j}]$$

The service time pdf, $f_{B_i}(x)$, cannot be obtained analytically, and is expensive to evaluate numerically. Hence, it is very difficult to calculate the response time pdf, $f_W(x)$, exactly, either analytically or numerically. However, $f_W(x)$ can be readily approximated from its first four moments (calculated from Equation (8) using the Generalised Lambda Distribution (GLD) [22].

IV. Validation

Our experimental platform consists of a Seagate ST3500630NS disk drive with 60 801 cylinders. A sector is 512 bytes and we have approximated, based on measurements from the disk drive, that the time to write a single physical sector on the innermost and outermost tracks are 0.012064ms (t_{max}) and 0.005976ms (t_{min}) respectively. We define a block as 128KB, and therefore there are 256 sectors per block. The time for a full disk revolution is 8.33ms. A track to track seek takes 0.8ms and a full-stroke seek requires 17ms for a read; the same measurements are 1ms and 18ms respectively for a write [23]. Based on information from the disk manufacturer, we set the miss probability at 0.05. To obtain response time measurements from this system, we implemented a benchmarking program that issues read and write requests using a master process and multiple child processes. These child processes are responsible for issuing and timing I/O requests, leaving the master free to spawn further child processes without the need for it to wait for previously-issued operations to complete.

In order to validate the analytical model effectively, it was necessary to minimise the effects of buffering and caching as these are not currently represented in the model. We therefore disabled the system's write-back cache, set the read-ahead buffer to 0 and opened the device with the `O_DIRECT` flag set. We also disabled the operating system's I/O scheduler. For each of the experiments presented below, 100 000 I/O requests were issued. To ensure a high disk utilisation with long queue lengths (i.e. a suitable environment for RPO), the mean arrival rate of I/O requests was set to either 0.03 or 0.04 requests/ms.

A. Service Time

In order to validate our service time model of Equation (9), we measured service times for various fixed queue lengths. Figure 1 plots measured and modelled mean service times against constant queue lengths. We observe moderate agreement between model and measurement with similar trends. We note that these results are based on using a value of $p_{miss} = 0.05$ according to manufacturer advice. However, substantially better agreement is observed for a value of $p_{miss} = 0.17$. One

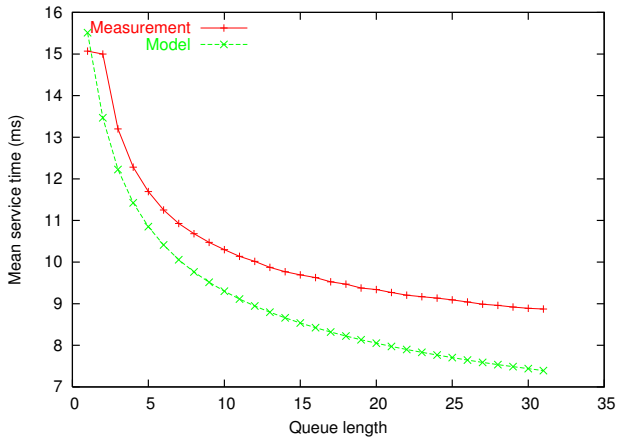


Fig. 1. Comparison of measured and modelled mean service times for various fixed queue lengths

avenue of future work is to devise experiments to determine the exact value of p_{miss} for our specific disk drive.

B. I/O Request Response Time

Figures 2 and 3 demonstrate the change in mean response time when different values are chosen for the queue length at which it is assumed that the service time distribution no longer changes for increasing queue lengths. A straight line is plotted to indicate the measured response time. For higher assumed maximum queue lengths, we observe excellent agreement between model and measurement for mean response times independent of arrival rate and request size. It can be observed, particularly for smaller sized requests and smaller arrival rates (e.g. Figures 2(a), 2(b), 3(a)), that the assumed maximum queue length does not have to be very high before convergence of the mean response times is observed. The impact of RPO on disk performance is magnified for larger request sizes and arrival rates. In many of these cases it can be observed that if RPO is not modelled (i.e. when the assumed maximum queue length is 1), the modelled mean response time is very high or the model is saturated (e.g. Figures 2(e), 3(c), 3(d)), whereas this does not occur in RPO-enabled measurements.

Although the mean response times show excellent agreement between model and measurement, our modelled variances compare less favourably with measurements. Table I presents variances for the same cases as Figures 2 and 3 using an assumed maximum queue length chosen at the length that the respective mean response time converges. For increasing arrival rates, the model presents significantly smaller variances than the measurements. Inevitably, this will affect skew and kurtosis (input parameters for the GLD with the mean and variance) to an even greater degree.

To test the accuracy of the GLD approximation that we use to approximate our response time densities, we first compare the approximation with a known pdf. In [6], an analytical model is introduced for response time distributions of single disks without RPO. In Figure 4, we compare this model with the GLD approximation of it, for single block transfers and arrival rate 0.01 requests/ms. We observe excellent agreement between approximate and exact models.

	$\lambda = 0.03$		$\lambda = 0.04$	
Size	Measured	Modelled	Measured	Modelled
1	129.3658	71.4639	234.3871	105.61
2	208.1058	110.839	383.8498	184.18
3	320.1849	175.285	822.2696	330.98
4	628.6987	280.2	2081.566	614.12
6	1568.488	737.56	10598.82	2494.2
7	3055.687	1229.4	25867.46	5745.9
8	6824.624	2106.4	sat	sat
9	11976.34	3809.4	sat	sat

TABLE I
MEASURED AND MODELLED VARIANCES FOR READ REQUEST RESPONSE TIMES ON A SINGLE DISK WITH DIFFERENT SIZED REQUESTS AND ARRIVAL RATE λ REQUESTS/MS

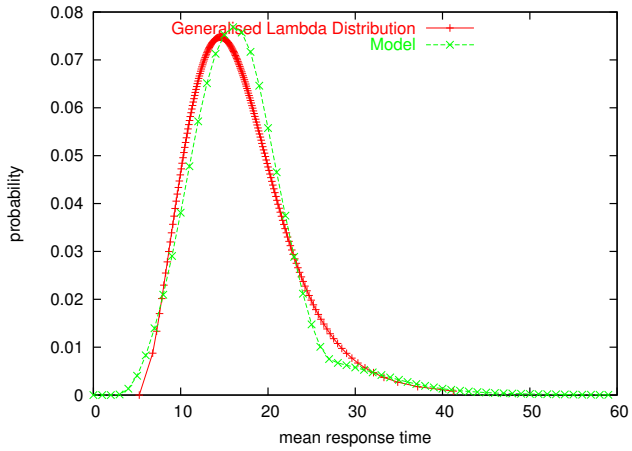


Fig. 4. Comparison of actual model and Generalised Lambda Distribution approximation for a 1-block read request to a single disk, arrival rate 0.01 requests/ms

In Figures 5 and 6 we present GLD approximations of the I/O response density of various request sizes and arrival rates of 0.03 and 0.04 requests/ms. Again we use a maximum queue length chosen at the length that the respective mean response time converges. We generally observe good agreement between model and measurement. However, the increase in difference between measured and modelled variances for larger request sizes causes increasing disagreement between model and measurement, despite still having excellent agreement for mean response time.

V. CONCLUSION

This paper has presented two contributions. Firstly, we introduced an approximation to the response time distribution of an $M/G/1$ queue with state-dependent service times. Secondly, we illustrated the effectiveness of this approximation by using it to model I/O request response times in zoned disk drives that intelligently re-order incoming requests. In order to do this we derived service time distributions according to queue length for use in the state-dependent service time model.

In the future we aim to extend the workloads that the model can support to include mixtures of read and write requests, requests of varying size and bursty arrivals. Additionally, caching is not yet supported in our model. We will also extend the disk model to represent RAID systems made up of RPO-enabled disk drives.

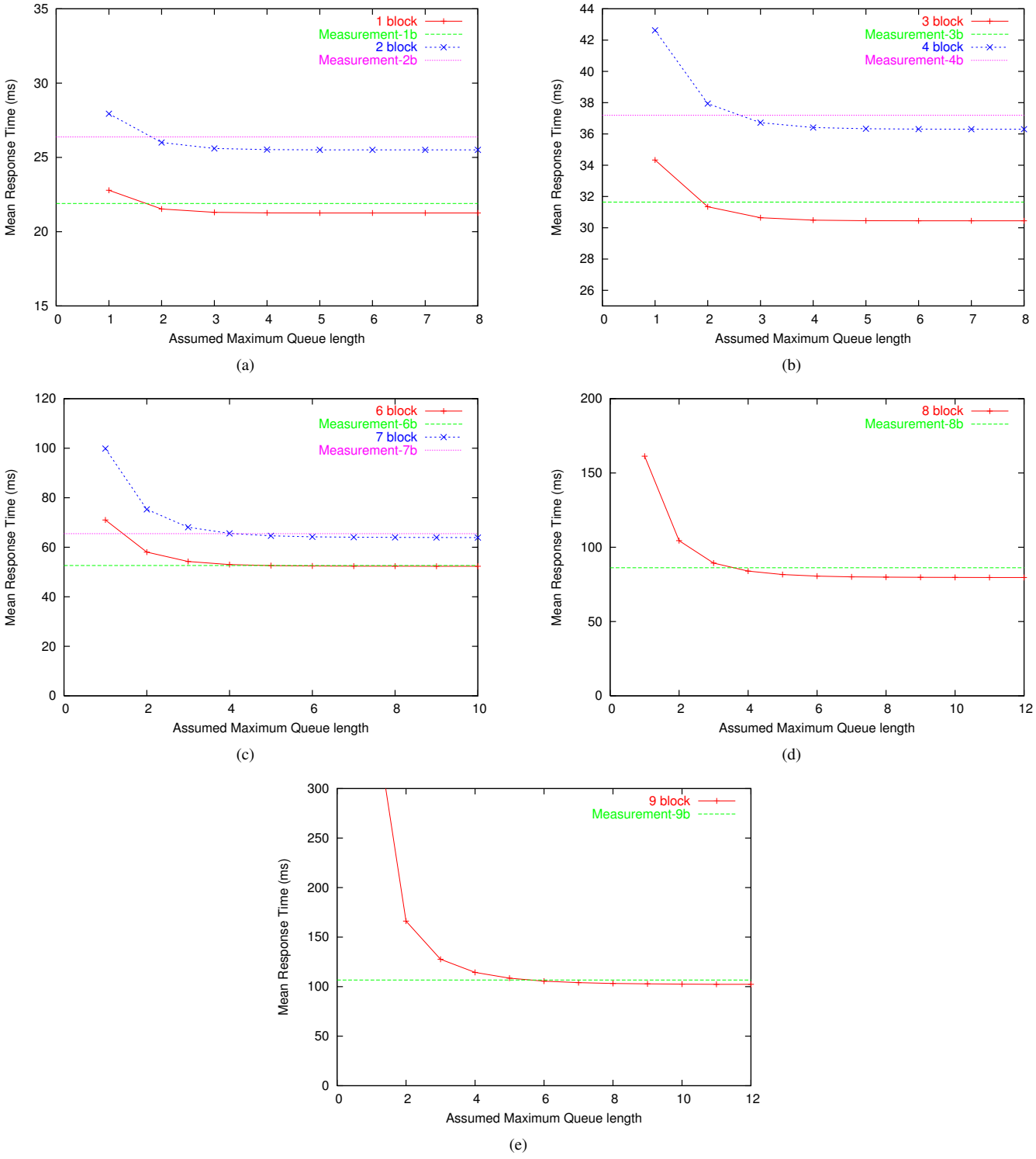


Fig. 2. Mean response time against assumed maximum queue length and measurements for different sized read requests (0.03 requests/ms)

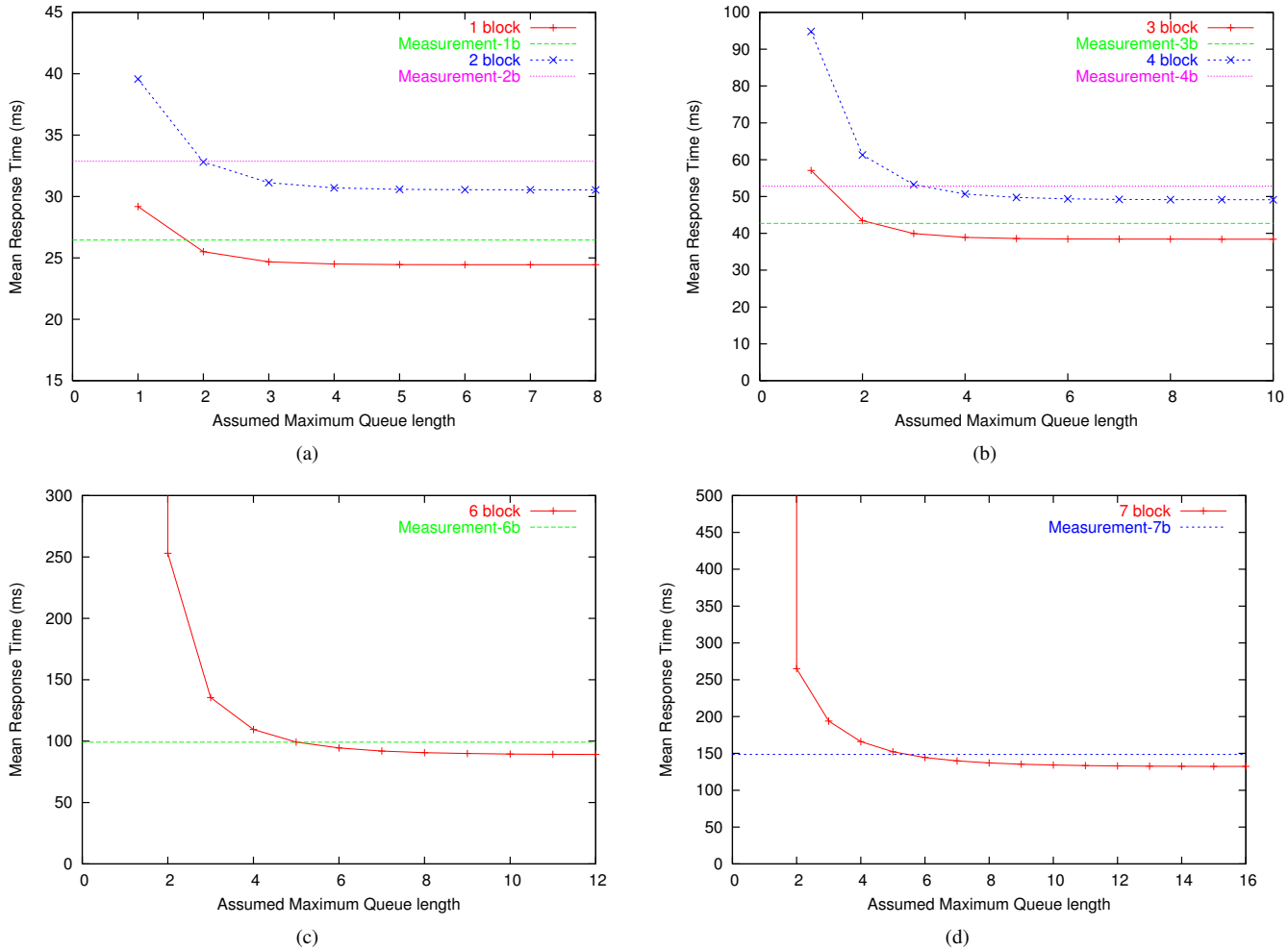


Fig. 3. Mean response time against assumed maximum queue length and measurements for different sized read requests (0.04 requests/ms)

ACKNOWLEDGEMENTS

We are grateful to Peter Harrison for helpful discussions. This work is supported by EPSRC research grant "Intelligent Performance Optimisation of Virtualised Data Storage Systems" (iPODS) (EP/F010192/1).

REFERENCES

[1] Seagate, "Economies of Capacity and Speed: Choosing the most cost-effective disc drive size and RPM to meet IT requirements," Seagate, Whitepaper, May 2004.

[2] W. A. Burkhard and J. D. Palmer, "Rotational position optimization (RPO) disk scheduling," University of California at San Diego, Tech. Rep., 2001.

[3] P. J. Denning, "Effects of scheduling on file memory operations," in *Proc. AFIPS Spring Joint Computer Conference*, vol. 31, 1967, pp. 9–21.

[4] D. M. Jacobson and J. Wilkes, "Disk scheduling algorithms based on rotational position," HP Laboratories, Tech. Rep. HPL-CSP-91-7rev1, 1991.

[5] M. Seltzer, P. Chen, and J. Ousterhout, "Disk Scheduling Revisited," in *Proc. USENIX Winter Technical Conference*. USENIX Association, 1990, pp. 313–324.

[6] A. S. Lebrecht, N. J. Dingle, and W. J. Knottenbelt, "A response time distribution model for zoned RAID," in *15th Int. Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA)*, June 2008.

[7] S. Zertal and P. G. Harrison, "Multi-RAID queueing model with zoned disks," in *High Performance Computing and Simulation Conference (HPCS'07)*, June 2007.

[8] C. Ruemmler and J. Wilkes, "Unix disk access patterns," in *Proc. USENIX Winter Conference*, San Diego, CA, 1993, pp. 405–420.

[9] W. W. Hsu and A. J. Smith, "The performance impact of I/O optimizations and disk improvements," *IBM Journal of Research and Development*, vol. 48, no. 2, pp. 255–289, 2004.

[10] B. L. Worthington, G. R. Ganger, and Y. N. Patt, "Scheduling algorithms for modern disk drives," in *Proc. ACM SIGMETRICS*, 1994, pp. 241–251.

[11] A. Huffman and J. Clark, "Serial ATA Native Command Queueing," Intel Corporation and Seagate Technology, Whitepaper, July 2003.

[12] S. Chen, J. A. Stankovic, J. F. Kurose, and D. Towsley, "Performance evaluation of two new disk scheduling algorithms for real-time systems," *Real-Time Systems*, vol. 3, no. 3, pp. 307–336, 1991.

[13] E. Shriver, A. Merchant, and J. Wilkes, "An analytic behavior model for disk drives with readahead caches and request reordering," in *Proc. ACM SIGMETRICS*. ACM, 1998, pp. 182–191.

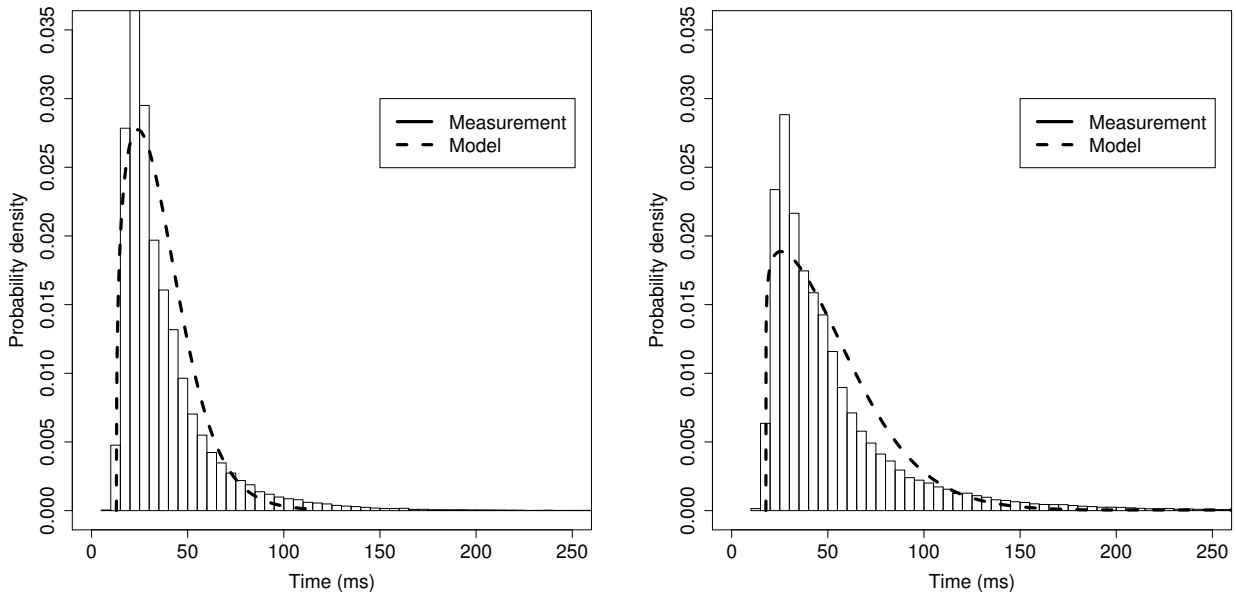
[14] C. C. Gotlieb and G. H. MacEwen, "Performance of movable-head disk storage devices," *Journal of the ACM*, vol. 20, no. 4, pp. 604–623, 1973.

[15] C. M. Harris, "Queues with state-dependent stochastic service rates," *Operations Research*, vol. 15, no. 1, pp. 117–130, 1967.

[16] J. A. Morrison, "Sojourn and waiting times in a single-server system with state-dependent mean service rate," *Queueing Systems*, vol. 4, no. 3, pp. 213–235, 1989.

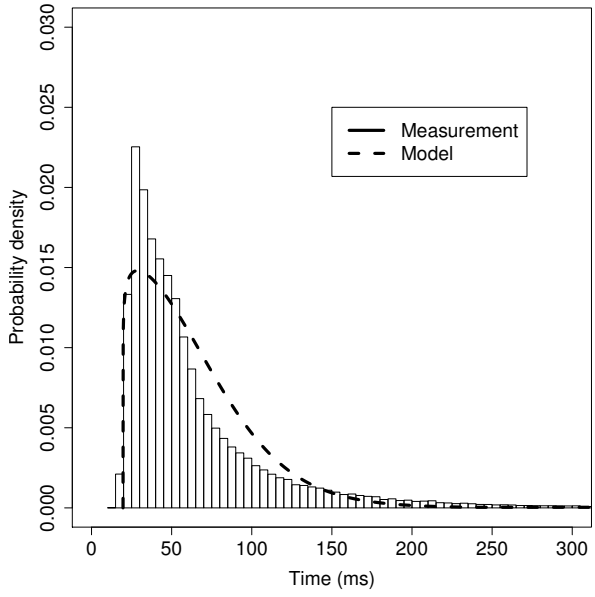
[17] P. H. Brill and M. J. M. Posner, "Level crossings in point processes applied to queues: Single server case," *Operations Research*, vol. 25, no. 4, pp. 662–674, 1977.

[18] D. I. Choi, C. Knessl, and C. Tier, "A queueing system with queue length dependent service times, with applications to cell discarding in ATM

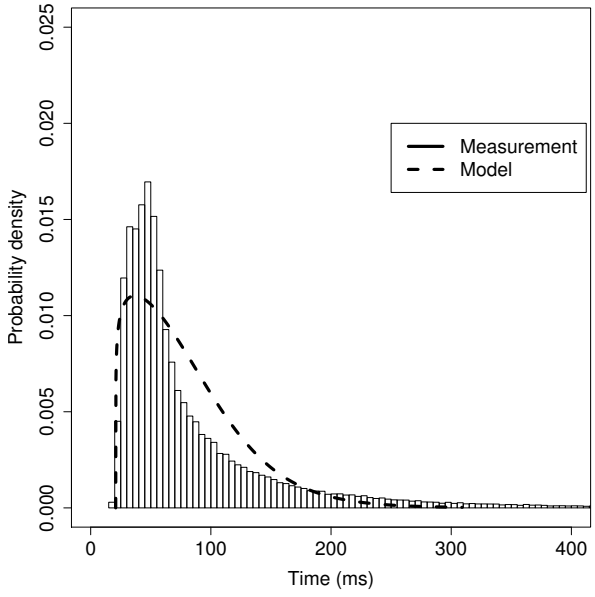


(a) 4 block

(b) 6 block



(c) 7 block



(d) 8 block

Fig. 5. Comparison of measurements and approximations of the modelled pdfs for response times of different sized read requests to a single disk with arrival rate 0.03 requests/ms

networks," *J. Applied Mathematics and Stochastic Analysis*, vol. 12, no. 1, pp. 35–62, 1999.

[19] W. J. Gray, P. Wang, and M. Scott, "An M/G/1-type queuing model with service times depending on queue length," *Applied Mathematical Modelling*, vol. 16, no. 12, pp. 652 – 658, 1992.

[20] P. G. Harrison and N. M. Patel, *Performance Modelling of Communication Networks and Computer Architectures*. Addison-Wesley, 1993.

[21] H. A. David, *Order Statistics*. John Wiley and Sons, Inc, 1981.

[22] A. Lakhany and H. Mausser, "Estimating the parameters of the Generalized Lambda Distribution," *Algo Research Quarterly*, vol. 3, no. 3, pp. 47–58, December 2000.

[23] Seagate, "Barracuda ES Data Sheet," 2007, http://www.seagate.com/docs/pdf/datasheet/ disc/ds_barracuda_es.pdf.

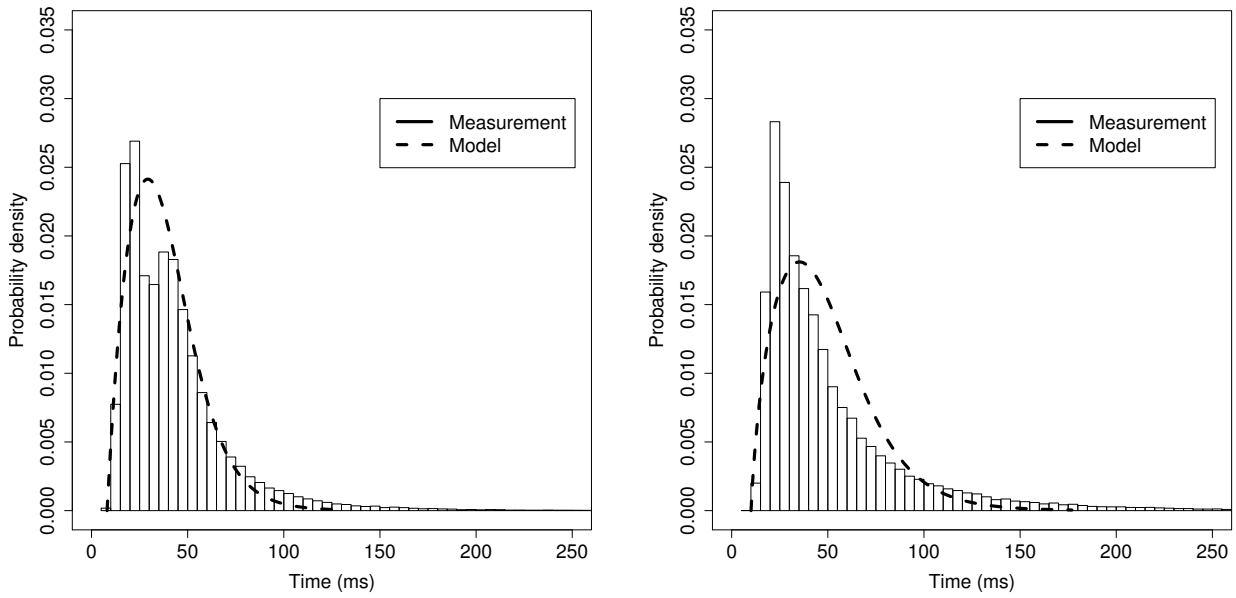
[24] S. Chen and D. Towsley, "The design and evaluation of RAID 5 and parity striping disk array architectures," *IEEE Transactions on Parallel and Distributed Systems*, vol. 17, no. 1-2, pp. 58–74, 1993.

[25] ——, "A performance evaluation of RAID architectures," *IEEE Transactions on Computers*, vol. 45, no. 10, pp. 1116–1130, 1996.

APPENDIX

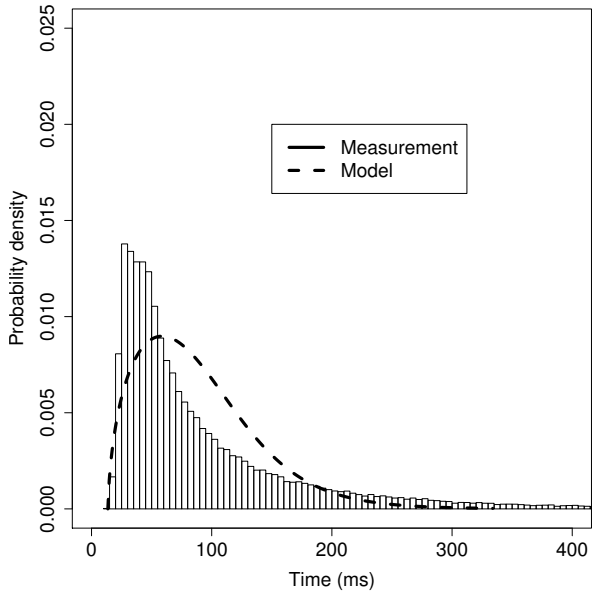
We present a summary of the derivation of the existing zoned disk model presented in [6].

In making performance predictions for a disk array or storage system, it is fundamental to model disk service time accurately. To this end, we model a disk drive as an $M/G/1$ queue where the service time density is the convolution of seek time, rotational latency and data transfer time densities.

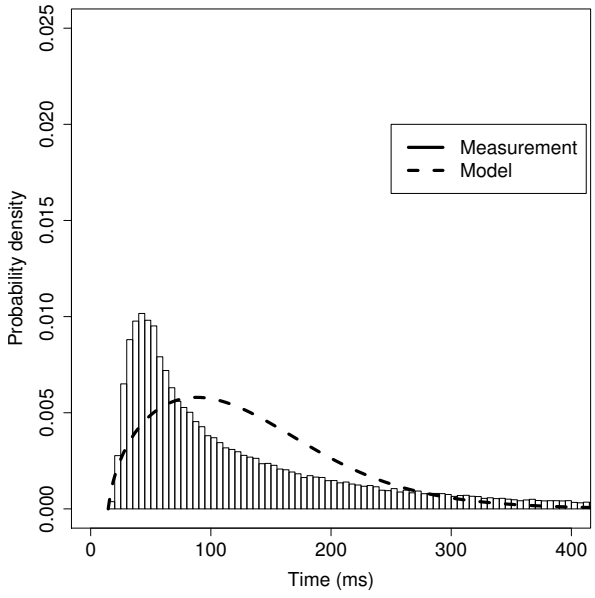


(a) 3 block

(b) 4 block



(c) 6 block



(d) 7 block

Fig. 6. Comparison of measurements and approximations of the modelled pdfs for response times of different sized read requests to a single disk with arrival rate 0.04 requests/ms

Defining random variables for seek time, S, rotational latency, R, and block transfer time, T, we describe their distributions below.

A. Seek Time

A seek, S, is the time taken for the disk head to move from the cylinder where it is currently located, C_2, to the cylinder containing a target sector, C_1. We define a random variable, $D = |C_1 - C_2|$, as the seek distance. Seek time can then be defined in terms of seek distance. Specifically [24],

$$S(D) = \begin{cases} 0 & \text{if } D = 0 \\ a + b\sqrt{D} & \text{otherwise} \end{cases}$$

where a and b are constants defined in terms of the disk geometry, and are given by:

$$a = \frac{minseek\ \sqrt{Cyls - 1} - maxseek}{\sqrt{Cyls - 1} - 1}$$

$$b = \frac{maxseek - minseek}{\sqrt{Cyls - 1} - 1}$$

Here $Cyls$ is the total number of cylinders on the disk, $minseek$ is the track-to-track seek time and $maxseek$ is the full-stroke seek time.

The disk model must reflect the layout of a zoned disk accurately. As cylinders get closer to the disk edge, their circumference increases and the number of sectors per cylinder increases. Therefore, a random request has an increased probability of being directed to a sector on an outer cylinder. Let C be a random variable representing the cylinder number of a randomly selected disk sector. Then the probability distribution of C can be approximated by assuming that the number of sectors per track increases linearly [7]. That is,

$$f_C(x) = \frac{\alpha + \beta x}{\gamma} \qquad x = 0, 1, \ldots, Cyls - 1$$

with constants α, β and γ defined as:

$$\begin{aligned} \alpha &= \frac{SEC[0]}{spb} \\ \beta &= \frac{SEC[Cyls-1] - SEC[0]}{(Cyls-1)\ spb} \\ \gamma &= \alpha(Cyls-1) + \frac{\beta}{2}(Cyls-1)^2 \end{aligned}$$

where $SEC[0]$ and $SEC[Cyls-1]$ are the number of sectors on the innermost and outermost tracks respectively and spb is the number of physical sectors per logical block. α represents the number of logical blocks on the innermost track and β charts the rate of increase in blocks per cylinder.

Often the disk specifications do not provide the number of sectors on the innermost and outermost tracks. However, it is possible to take measurements from the disk drive to ascertain the mean transfer time to a single sector on the innermost (t_{min}) and outermost (t_{max}) tracks. Then α and β are calculated from the transfer time definition in Equation 11.

The pdf of seek distance is calculated by assuming the two random variables, C_1 and C_2 are two distinct cylinder numbers, and calculating the seek distance between all possible cylinder numbers. This is split into two terms, one for the case when $C_1 \leq C_2$ and one for the case where $C_1 > C_2$:

$$\begin{aligned} f_D(x) &= \int_0^{Cyls-1-x} f_C(y) f_C(x+y) dy + \\ &\quad \int_x^{Cyls-1} f_C(y) f_C(y-x) dy \end{aligned}$$

This can be shown to equate to

$$f_D(x) = A + Gx + Ex^3 \quad 0 \leq x \leq Cyls - 1$$

where,

$$\begin{aligned} A &= \frac{V(C-1)}{3\gamma^2} \\ G &= -\frac{V + \beta^2(Cyls-1)^2}{3\gamma^2} \\ E &= \frac{\beta^2}{3\gamma^2} \\ V &= 6\alpha^2 + 6\alpha\beta(Cyls-1) + 2\beta^2(Cyls-1)^2 \end{aligned}$$

The cumulative distribution function (cdf) of seek time, $F_S(t)$, can be defined in terms of the cdf of $f_D(x)$, $F_D(x)$, as [24]:

$$F_S(t) = \begin{cases} F_D(0) & 0 \leq t < a + b \\ F_D\left(\left(\frac{t-a}{b}\right)^2\right) & \text{otherwise} \end{cases}$$

B. Rotational Latency

Rotational latency, R, is the time to rotate to the angle of a target sector. R has a uniform distribution with a range between 0 and the time for a full disk revolution, R_{max} [25].

$$f_R(x) = 1/R_{max} \qquad 0 \leq x \leq R_{max} \tag{10}$$

C. Data Transfer Time

The time to transfer k logical blocks on cylinder x of a zoned disk can be approximated as [7]:

$$t(x) = \frac{k\ spb\ R_{max}}{\alpha + \beta x} \tag{11}$$

Denoting T_k as the random variable of the time to transfer k blocks of data, its cdf is:

$$\begin{aligned} F_{T_k}(t) &= \int P(T_k \leq t \mid C = x) f_C(x) dx \\ &= \int P(x \geq \frac{kspbR_{max}}{\beta t} - \frac{\alpha}{\beta}) f_C(x) \\ &= \int_{\max(\phi_k(t),0)}^{Cyls-1} f_C(x) dx \end{aligned} \tag{12}$$

where

$$\phi_k(t) = \frac{k\ spb\ R_{max}}{\beta t} - \frac{\alpha}{\beta}$$

calculates the minimum cylinder number it is possible to transfer k logical blocks of data in less than t ms. The solution of the integral in Equation (12) is a function of t with a domain bounded between the minimum and maximum possible k-block transfer times.

Equation (12) expands to:

$$F_T(t) = \begin{cases} 0 & t < kspbt_{min} \\ \frac{\alpha}{\gamma}(Cyls-1) + \frac{\alpha^2}{2\beta\gamma} + & \\ \frac{\beta(Cyls-1)^2}{2\gamma} - \frac{k^2 R_{max}^2 spb^2}{2t^2\beta\gamma} & t < kspbt_{max} \\ 1 & \text{otherwise} \end{cases} \tag{13}$$