Non-linear Seek distance for optimal accuracy of zoned disks seek time in Multi-RAID storage systems

Soraya Zertal[1] and Peter Harrison[2]
[1] PRiSM, Université de Versailles, 45 Av. des Etats-Unis, 78000 Versailles, France
Zertal@prism.uvsq.fr
[2] Imperial College London, South Kensington Campus, London SW7 2AZ, UK
pgh@doc.ic.ac.uk

ABSTRACT

Models of multi-RAID storage systems, implemented on modern zoned disks, are simplified by using an approximation for the seek distance that assumes that the number of sectors per track varies linearly with the cylinder number. Results obtained in this way have matched well against simulation, but the relationship of the number of sectors per track against cylinder number has turned out to be rearkably linear for the specific RAID systems modelled so far. This will not always be the case, necessarily, and in this paper, we go a step further, calculating exactly the seek distance moments, then those of the seek time on zoned disks for specific manufacturer's specifications. This is to ensure the highest accuracy possible for our model, especially for the non-sequential request streams. The results show good accuracy, even in highly non-linear systems, and indicates a threshold for the model parameters at which the linear approximation becomes unacceptable.

KEYWORDS

Multi-RAID, Zoned disks, Seek distance distribution, M/G/1 queues, I/O modeling and Simulation.

1 Introduction

A continual, heavy and increasing pressure persists on storage systems, necessitating accurate models of their operation, capable of analysing and predicting the performance and quality of service (QoS) these systems can deliver. To fulfill these requirements, models should provide detailed abstractions of a wide range of possible real system architectures. Many storage system models do exist and we can split them into three broad categories: the first focuses on calculating the service time for a given RAID configuration and a specific type of workload [10]; the second – widely investigated – concerns the analysis of the performance of a given RAID configuration in a specific working mode [2, 3, 16, 17, 18, 11, 1, 13]; and the third category studies the effect of caching and controller optimizations on a disk array's performance [15, 14]. Regardless of the main goal of a study among these categories, the objective is always one and only one RAID configuration per disk array. We proposed previously, in [28], a Multi-RAID model which is – as far as we know – the only one that models a RAID storage system with multiple, coexisting RAID organisations, using modern zoned disks. In that model, the probability distribution of the seek distance is considered to be a continuous random variable, assuming that the number of sectors per track varies linearly with the cylinder number. In this paper, we calculate exactly the seek distance moments to provide the highest accuracy for our model, which would become applicable to arbitrary zoned disks, possibly with a highly non-linear relationship between number of sectors and cylinder sequence number.

In the rest of the paper, section 2 considers the technology issues in the design of RAIDs of zoned disks as well as the details of our analytical model. Section 3 describes the calculation of the moments of the seek distance and their impact on our zoned Multi-RAID model. Results are presented and discussed in section 4, and the paper concludes in section 5 with a summary and some future related research objectives.

2 Technological and modeling context

A RAID storage system consists of a disk system manager and a collection (array) of independent disks. The disk system manager is a software component of the RAID controller. It is responsible for the logical to physical mapping of requests according to the prevailing RAID organisation scheme [4, 5].

We proposed a dynamic Multi-RAID architec-

Proceedings of the 2008 High Performance
Computing & Simulation Conference ©ECMS
Waleed W. Smari (Ed.)
ISBN: 978-0-9553018-7-2 / ISBN: 978-0-9553018-6-5 (CD)

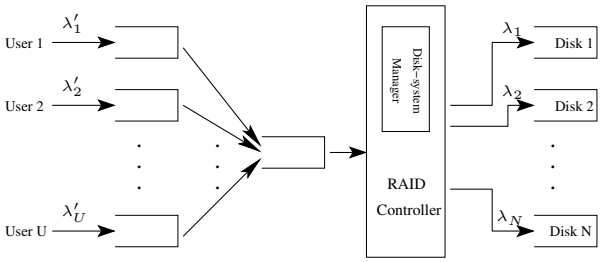

Figure 1: Requests flow in a RAID storage system

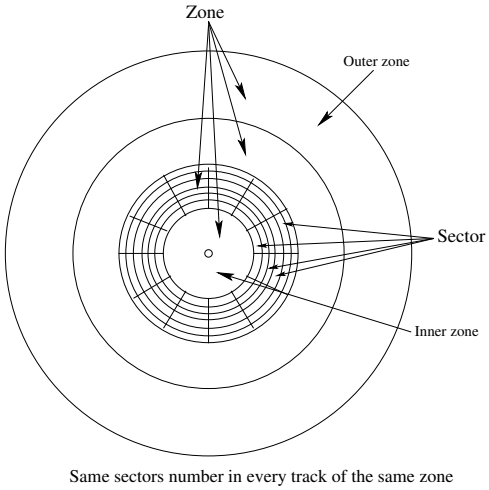

Figure 2: Zoned disk technology

ture, on which various RAID organisation schemes coexist, with a dynamic selection of a redundancy pattern during the data lifetime, in order to provide enhanced space use and access time, [26] . Related requests' independent executions on such asynchronous disks lead to Fork-Join-type modeling problems. We modelled this architecture, using a collection of M/G/1 queues with various extensions to account for the parallel disk (physical) accesses corresponding to a logical request [27, 8]. The response time of each physical request, to an individual disk, is composed of four components: the queueing time (Q), the seek time (S), the rotational latency (R) and the transfer time, which itself is divided into two components, t and T_{bus}, corresponding to the transfer time between the disk's cylinder and its buffer, and between this buffer and the controller via the bus, respectively. We first considered uniform disks[1] to validate our model [9] and extended our initial model to modern zoned disks using more accurate access time functions [28].

On zoned disks (see figure 2), the number of sectors per cylinder is variable. Consecutive cylinders are collected into groups, called *zones*, such that within each zone, the track capacity (number of sectors) and the transfer rate are fixed. However, these two parameters decrease from the outer to the inner zones. These disks have become very popular due to their greater storage capacitiy and transfer rate. Their average rotational latency is constant but the variable seek and transfer times necessitate more complex calculations in terms of the assumed statistical workload and disk operation principles in a storage model [20, 21].

The model we presented in [28] deals with an assumed linear relationship between the number of sectors per track and the cylinder number. This gives a good approximation to the seek distance distribution on such modern zoned disks in the context of a Multi-RAID system. As far as we know, there is no analytical model for a non-linear seek distance distribution on zoned disks. We approximated it using a linear function in [28], rather than by using interpolation, as in [24], or using parameters based solely on preliminary simulations, as in [19]. An approximation for seek time in terms of seek distance is proposed in [7]: proportional to the square root of seek distance, when below some threshold value, otherwise linear in the seek distance. A Chernoff bound on the transfer time of a 'sweep' of N requests is found, assuming equidistant seek positions, giving a constant total seek time for the sweep. Further, this model assumes that all zones have the same number of tracks and that the track capacity increases linearly, which is much more restrictive than our approximation in [28]. Here, we calculate the moments of seek distance exactly, for uniform address accesses, and hence of seek time, allowing highly non-linear regimes to be investigated. A glossary of the notation used is provided by Table 1.

[1] Sectors are uniformly distributed across the tracks and all the tracks have the same number of sectors.

3 The non-linear seek distance

Seek time is one of the main components of a disk's access time. The morphology of a zoned disk improves its performance but makes its modelling much more complicated because of having to deal with a non-constant (here, not even linear) seek distance between the read/write head's current position and its target one in any disk access. We assume that the incoming logical requests' addresses are independent random variables, uniformly distributed over the disk-address space. This does not mean that the accesses are uniformly distributed over the disk's physical space, of course – the access distribution over the physical space follows its density, i.e. higher on the outer zone and decreasing towards the

Param.	Description
N	Number of disks in the storage system.
C	Number of cylinders on a disk.
SEC	Number of sectors on the disk.
N_z	Number of zones in the disk.
SEC_c	Number of sectors on cylinder c.
spb	Number of sectors per block.
B	Logical request size (transfer block).
K_i	The number of blocks generated by a logical request at disk i
D_i	Seek distance on a disk i.
S_i	Seek time on a disk i. buffer and its cylinder.
λ	Logical request arrival rate to the storage system.
p_i	Probability that disk i is used.
λ_i	Physical request arrival rate to disk i.
λ_{Rj}	Physical request arrival rate to the RAIDj area.
λ_{iRj}	Physical request arrival rate to a RAIDj area on disk i.
P_{raid_j}	$RAIDj$ area's proportion in the whole storage system space.
p_w	Probability that a request is a write.
p_r	Probability that a request is a read.
p_s	Probability of a sequential access.
z_i	The probability to access a sector in zone i.
C_i	The first cylinder's number of zone i.
d_i	The number of cylinders in zone i.

Table 1: Notation for the RAID model's parameters

inner zone. Since the number of cylinders C is large, the seek distance D can be well approximated by a continuous random variable. On Uniform disks, the seek distance density function is [9]:

$$f_D(x) = p_s\delta(x) + (1-p_s)\frac{2(C-x)}{(C-1)^2}$$

for $0 \le x \le C-1$, where p_s is the probability that consecutive accesses are sequential, i.e. on the same track, requiring no seek for the second access.

The term $\frac{2(C-x)}{(C-1)^2}$ is the probability density function of the difference between two uniform random variables on $[0, C-1]$, and $\delta(x)$ is the Dirac delta-function (unit impulse). Turning now to zoned disks, assuming that the number of sectors (and hence blocks) per track varies linearly with the cylinder number, the density function of D can be shown to be [28]:

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

Thus, the nth moment of the seek distance D can be calculated as:

$$M_n = (C-1)^{n+1}[\frac{A}{n+1} + \frac{G(C-1)}{n+2} + \frac{E(C-1)^3}{n+4}]$$

where

$$\begin{aligned} A &= \frac{V(C-1)}{3\gamma^2} \\ G &= -\frac{V+\beta^2(C-1)^2}{3\gamma^2} \\ E &= \frac{\beta^2}{3\gamma^2} \\ V &= 6\alpha^2 + 6\alpha\beta(C-1) + 2\beta^2(C-1)^2 \\ \gamma &= \alpha(C-1) + \beta(C-1)^2/2 \\ \alpha &= SEC_{C-1}/spb, \\ \beta &= (SEC_0 - SEC_{C-1})/(spb(C-1)) \end{aligned}$$

We implemented this linear approximation and the obtained results showed good agreements when applied to a real disk device : Fujitsu-MAN3397 disk [6], as we can see in figure 3. This figure plots the exact number of sectors per track (as specified in the disk's technical data) against cylinder sequence number (with the higher numbered, inner cylinders on the left) over the whole disk. This graph is compared with the model's linearised version in the same figure, showing close agreement. However, we cannot rely on such an approximate linear relationship holding good on all zoned disks (see section 4). The reason is that on such disks, the streaming bandwith varies by over 50% from one part of the disk to another [22], showing clear non-linearity.

Very few applications writers exploit this non linearity, however, specifying explicitly the data placement so that the behaviour of an application can be predicted [23, 25]. Most programmers completely ignore the data placement that their applications will use. Models of such applications that employ (approximately) linearized seek time distributions do not, therefore, represent real system behaviour faithfully, and so predict performance poorly. This motivated our exact seek distance calculation, using real zoned disks' published hardware characteristics. This results in seek time moments being estimated by the formulae given in the proposition below:

Proposition 1 *Consider a disk with N_z zones, numbered $0, 1, \ldots, N_z - 1$ and C cylinders, counting from the outside of the disk. Let the first cylinder in zone i be numbered C_i, so that the number of cylinders in zone i is $d_i = C_{i+1} - C_i$. Then the n^{th} moment M_n of seek distance, assuming uniform access to sectors over the whole disk, is*

$$M_n = \frac{2}{(n+1)(n+2)} \sum_{0 \le j < i}^{N_z-1} \frac{z_i z_j}{(d_i-1)(d_j-1)} \times$$

$$[(C_i - C_{j+1} + 1)^{n+2} + (C_{i+1} - C_j - 1)^{n+2} - (C_i - C_j)^{n+2} - (C_{i+1} - C_{j+1})^{n+2}] + \frac{2}{(n+1)(n+2)} \sum_{i=0}^{N_z-1} z_i^2 (d_i - 1)^n$$

where $z_i = \frac{d_i \times SEC_i}{SEC}$ is the probability of a single request accessing a sector in zone i.

Proof Let the random variable D denote the seek distance for an access on the disk, assuming that all accesses are uniformly distributed over the sectors. Then, the probability that a given access is to a sector in zone i is z_i, as defined. Hence we have:

$$\begin{aligned}
&M_n = E[D^n] \\
&= E[E[D^n \mid \text{seek between cylinders } i \text{ and } j,\ i \geq j]] \\
&= \sum_{j<i} \frac{z_i z_j}{(d_i - 1)(d_j - 1)} \times \\
&\quad \int_{x=0}^{d_i-1} \int_{y=0}^{d_j-1} (C_i - C_j + x - y)^n \mathrm{d}x\mathrm{d}y \\
&\quad + \sum_{i=0}^{N_z-1} \frac{z_i^2}{(d_i - 1)^2} \int_{x=0}^{d_i-1} \int_{y=0}^{d_i-1} |x - y|^n \mathrm{d}x\mathrm{d}y \\
&= \frac{2}{(n+1)(n+2)} \sum_{0 \leq j < i}^{N_z-1} \frac{z_i z_j}{(d_i - 1)(d_j - 1)} \times \\
&\quad [(C_i - C_{j+1} + 1)^{n+2} + (C_{i+1} - C_j - 1)^{n+2} \\
&\quad - (C_i - C_j)^{n+2} - (C_{i+1} - C_{j+1})^{n+2}] \\
&\quad + 2 \sum_{i=0}^{N_z-1} \frac{z_i^2}{(d_i - 1)^2} \int_{x=0}^{d_i-1} \int_{y=0}^{x} (x - y)^n \mathrm{d}x\mathrm{d}y
\end{aligned}$$

The result follows on evaluating the integral. $\diamondsuit$

The seek time is calculated according to the widely accepted formula of Lee [12] :

$$S_i(D) = \begin{cases} 0 & \text{if } D_i = 0 \\ a\sqrt{D_i} + b(D_i - 1) + c & \text{otherwise} \end{cases}$$

where a, b, c are hardware-related constants :

$$\begin{aligned}
a &= (-10 \times MinSeek + 15 \times AvgSeek \\
&\quad -5 \times MaxSeek)/(3 \times \sqrt{C}) \\
b &= (7 \times MinSeek - 15 \times AvgSeek \\
&\quad +8 \times MaxSeek)/(3 \times C) \\
c &= MinSeek
\end{aligned}$$

and the three first moments required for the response time moment calculation on the Multi-RAID system are, as in [28]:

$$\begin{aligned}
\overline{S} &= (c - b) + aM_{1/2} + bM_1 \\
\overline{\overline{S}} &= (c - b)^2 + 2a(c - b)M_{1/2} \\
&\quad +[a^2 + 2b(c - b)]M_1 + 2abM_{3/2} + b^2 M_2 \\
\overline{\overline{\overline{S}}} &= (c - b)^3 + a(c - b)^2 M_{1/2} \\
&\quad +3(c - b)[a^2 + b(c - b)]M_1 \\
&\quad +[a^3 + 6ab(c - b)]M_{3/2} \\
&\quad +3[a^2 b + b^2(c - b)]M_2 + 3ab^2 M_{5/2} + b^3 M_3
\end{aligned}$$

4 Results and discussion

We calculated and compared the first three moments of the seek time using the linear seek distance approximation [28], the non-linear seek distance calculation of proposition 1 and real system simulation using our Multi-RAID system simulator. The system modelled was composed of 16 disks with 200,000 simulated small logical requests arriving at different rates from 10req/s to 1000req/s. For our experiments, two kinds of zoned disks were used: a fujitsu-MAN3367 and a fictitious one with a highly non-linear relationship between sectors per track and cylinder number – see the characteristics in table 2.

It is not unusual to use a fictituous component to validate an extreme case. For example, [22] used a hypothetical fast disk, called Uberdisk, the parameters of which were scaled from contemporary disks, to approximate the performance of a MEMstore. Here, we use a fictitious disk as well, parameterized in terms of capacity (total number of Gigabytes and number of sectors), performance (rotation time and minimum/average/maximum seek times) and morphology (numbers of data heads and cylinders). These were scaled from a real disk (Fujitsu-MAN3367), apart from a reduced number of zones, to yield a significantly non-linear distribution of seek distance. This is to show the accuracy of the approximation as well as the value of the exact calculation for disks with a non-linear morphology. Figures 3 and 4 show that the number of zones changes from 18, on the Fujitsu disk, to 4, on the fictitious disk, but that the distribution of storage density changes drastically from linear to clearly non-linear, as intended.

The three first seek moments using the Fujitsu-MAN3367 disk in table 3 highlight the accuracy of the linear approximation for this kind of device. In fact, on such devices with a certain linearity of the recorded density distribution, the distribution of sectors across the zones shows consequently a certain linearity which makes the linear approximation and the non linear 'exact' calculation close to each other, matching very well with the simulation results. However, we notice the small advantage of the latter, nonlinear calculation.

Param.	value	value
	Fuj-MAN3367	Fictif
capacity	36,74 GB	36,74 GB
Sectors	$18,37 \times 10^7$	$18,37 \times 10^7$
Rotation	10000 rpm	10000 rpm
Cylinders	29950	29950
Min Seek	0,4 - 0,6 ms	0,4 - 0,6 ms
Avg Seek	4,5 - 5 ms	4,5 - 5 ms
Max Seek	11 - 12 ms	11 - 12 ms
Data Heads	4	4
Zones nbr	18	4

Table 2: Fujitsu-MAN3367 Vs Fictitious disks characteristics

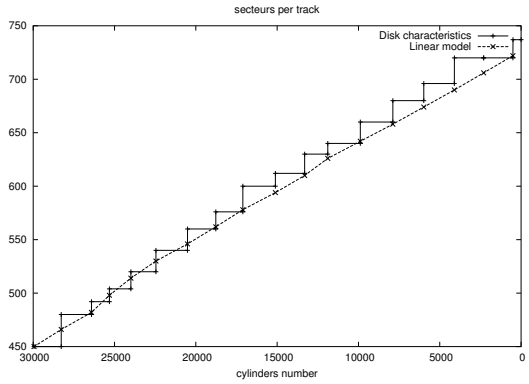

Figure 3: Sectors per track (FujitsuMAN3367)

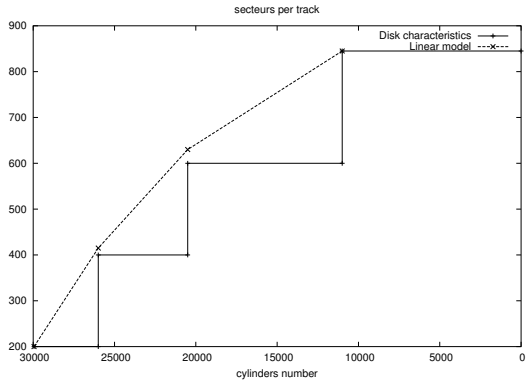

Figure 4: Sectors per track (Fictitious disk)

This linear distribution of sectors across the zones is not always respected, however. Even when it is, the seek time does not depend solely on the seek distance but also on the start position (or end position) of a seek. This is more important when these two positions belong to different zones, which is generally the case for non-sequential accesses. For such an access profile, using an approximation for the seek time calculation introduces a delay, accumulating along the request stream and so generating a larger discrepancy against the real system. Even if we can consider the seek time approximation accurate enough for sequential access, we certainly cannot consider it thus for non-sequential access, due to the cumulative delays introduced in a sequence of accesses. The more these non-sequential accesses appear, the more significant is the re-positioning delay and the less accurate is the seek time approximation.

	Model approx. linear Di	Model calc. Non linear Di	Simul.
$\overline{S}$	4.7	4.72	4.73
$\overline{\overline{S}}$	28.54	28.56	28.61
$\overline{\overline{\overline{S}}}$	200.35	199.98	199.21

Table 3: Seek time moments comparison (Fujitsu-MAN3367

Hence, the linear approximation cannot always give accurate results regardless of the device's morphology. Consequently, we used the fictitious disk, with the same space storage capacity and performance characteristics, but with a non linear distribution of sectors on zones to calculate the superiority of the non-linear seek distance calculation against its linear counterpart, with the Multi-RAID system simulation as a reference. The results obtained in table 4 confirm this superiority.

	Model approx. (linear Di)	Model calc. (Non linear Di)	Simul.
$\overline{S}$	4.47	4.41	4.36
$\overline{\overline{S}}$	26.00	24.90	24.56
$\overline{\overline{\overline{S}}}$	176.26	163.85	161.35

Table 4: Seek time moments comparison (Fictitious disk)

5 Conclusion

In this paper, we have compared our previously published model for the moments of seek distance in RAID models, which approximated the number of sectors on a track as a linear function of the track's sequence number, against an exact, explicit calculation. We showed that the proposed approximation is

still accurate for devices showing a certain degree of linearity in the density of storage over zones. However, the results obtained show the clear superiority of the non-linear moments calculation, regardless of the disk morphology. Using this result, we can guarantee almost perfect accuracy in the seek time calculation for uniform disk block accesses, which is a major component of the I/O response time, especially for non-sequential request streams.
We thereby constructed an accurate disk array model and in the near future, we expect to extend it to account for the operation of the bus that connects disks to the controller, which is subject to congestion and a consequent bus-queue, introducing further delay at high request arrival rates.

REFERENCES

[1] E. Bachmat and J. Schindler. Analysis od methods for scheduling low priority disk drive tasks. In *ACM Sigmetrics*, 2002.

[2] S. Chen and D. Towsley. The design and evaluation of raid5 and parity striping disk array architectures. *Parallel and distributed computing*, 17, 1993.

[3] S. Chen and D. Towsley. A performance evaluation of raid architectures. *IEEE Transactions on Computers*, 45(10), October 1996.

[4] G. Gibson D. A. Patterson and R. H. Katz. A case for redundant arrays of inexpensive disks (RAID). In *Proceedings of SIGMOD Conference*, June 1988.

[5] G. Gibson D. A. Patterson, P. M. Chen and R. H. Katz. Introduction to redundant arrays of inexpensive disks (RAID). In *IEEE COMPCON*, 1989.

[6] Fujitsu. *Disk Drives. Products/Maintenance Manual*. Fujitsu, 2001.

[7] Peter Muth Guido Nerjes and Gerhard Weikum. Stochastic service guarantees for continuous data on multi-zone disks. In *Symposium on Principles On Database Systems*, 1997.

[8] P.G. Harrison and S. Zertal. Queueing models with maxima of service times. In *Proceedings of TOOLS Conference*, 2003.

[9] P.G. Harrison and S. Zertal. Queueing models of RAID systems with maxima of waiting times. *Performance Evaluation*, 2007.

[10] M. Y. Kim and A.N. Tantawi. Asynchronous disk interleaving: approximating access delays. *IEEE Transactions on Computers*, 40(7), 1991.

[11] A. Kuratti and W. H. Sanders. Performance analysis of the raid5 disk array. In *Proc. IEEE Int'l Computer Performance and dependability Symp.*, 1995.

[12] E.K. Lee. *Perfomance modelling and analysis of disk arrays*. ph.D. thesis, University of California, Berkeley, USA, 1993.

[13] E.K. Lee and R.H. Katz. An analytic performance model of disk arrays. In *Proc. ACM SIGMETRICS*, May 1993.

[14] G. A. Alvarez M. Uysal and A. Merchant. A Modular, Analytical Throughput Model for Modern Disk Arrays. In *Proceedings of the International Symposium onModelling, Analysis and Simulationof Computer and Telecommunications Systems (MASCOTS)*, August 2001.

[15] J. Menon. Performance of raid 5 disk arrays with read and write caching. *Distributed ND Parallel Databases*, 2(3), July 1994.

[16] A. Merchant and P.S. Yu. An analytical model or reconstruction time in mirrored disks. *Performance evaluation*, 20, May 1994.

[17] A. Merchant and P.S. Yu. Analytic modeling and comparisons of striping strategies for replicated disk arrays. *IEEE Transactions on Computers*, 44(3), March 1995.

[18] A. Merchant and P.S. Yu. Analytic modeling of clustered raid with mapping based on nearly random permutation. *IEEE Transactions on Computers*, 45(3), March 1996.

[19] R. Nelson and A. N. Tantawi. Approximate analysis of fork-join synchronisation in parallel queues. In *IEEE Trans. Computers*, volume 37, June 1998.

[20] S. Christodoulakis P. Triantafillou and C. A. Georgiadis. A Comprehensive Analytical Performance Model for Disk Devices Under Random Workloads. *IEEE Transactions on Knowledge and data Engineering*, 14(1), 2002.

[21] Sangsoo Park and Heonshik Shin. *Rigourous Modeling of Disk Performance for Real-Time Applications*, volume 2986. Springer Berlin, 2004.

[22] Steven W. Scholosser and Gregory R. Ganger. Mems-based storage devices and standard disk interfaces: A square eg in a round hole. Technical Report CMU-PDL-03-102, Carnegie Mellon University, December 2003.

[23] seon ho kim Shahram Ghandeharizadeh and cyrus Shahabi. Continuous display of video objects using multi-zone disks. Technical Report 94-592 USC, University of South California, April 2003.

[24] S. Varma and A. M. Makowski. Interpolation approximations for symmetric fork/join queues. In *Performance Evaluation Journal*, volume 20, 1994.

[25] Jun Wang and Yiming Hu. Profs - performance-oriented data organization for log-structured file system on multi-zone disks. In *9th Internantional Symposium on Modeling, Analysis and Simulation on Computer and Telecommunication Systems*, 2001.

[26] S. Zertal. *Dynamic redundancy mechanisms for storage customisation on multi disks storage systems*. ph.D. thesis, University of Versailles, France, January 2000.

[27] S. Zertal and P.G. Harrison. Multi-level raid storage system modelling. In *Proceedings of 2003 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (Spects)*, 2003.

[28] S. Zertal and P.G. Harrison. Multi-raid queueing model with zoned disks. In *High Performance Computing and Simulation*, 2007. Best paper Award.