

# Zoned-RAID

SEON HO KIM

University of Denver

HONG ZHU

University of Southern California

and

ROGER ZIMMERMANN

National University of Singapore

---

The RAID (Redundant Array of Inexpensive Disks) system has been widely used in practical storage applications for better performance, cost effectiveness, and reliability. This study proposes a novel variant of RAID named Zoned-RAID (Z-RAID). Z-RAID improves the performance of traditional RAID by utilizing the zoning property of modern disks which provides multiple zones with different data transfer rates within a disk. Z-RAID levels 1, 5, and 6 are introduced to enhance the effective data transfer rate of RAID levels 1, 5, and 6, respectively, by constraining the placement of data blocks in multizone disks. We apply the Z-RAID to a practical and popular application, streaming media server, that requires a high-data transfer rate as well as a high reliability. The analytical and experimental results demonstrate the superiority of Z-RAID to conventional RAID. Z-RAID provides a higher effective data transfer rate in normal mode with no disadvantage. In the presence of a disk failure, Z-RAID still performs as well as RAID.

Categories and Subject Descriptors: E.2 [**Data Storage Representations**]—*Composite structures*; C.4 [**Performance of Systems**]—*Performance attributes; Design studies*; H.2.4 [**Database management**]: *Systems—Multimedia database*

General Terms: Storage Systems Architecture, Design, and Validation

Additional Key Words and Phrases: RAID, multizone disks, disk array

## ACM Reference Format:

Kim, S. H., Zhu, H., and Zimmermann, R. 2007. Zoned-RAID. *ACM Trans. Storage* 3, 1, Article 1 (March 2007), 17 pages. DOI = 10.1145/1227835.1227836 <http://doi.acm.org/10.1145/1227835.1227836>

---

## 1. INTRODUCTION

The need for storage has increased rapidly over the years in almost every field that involves computing. Many artifacts that used to be stored in analog form

---

Author's address: S. H. Kim, University of Denver, Denver, CO; email: seonkim@cs.du.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or [permission@acm.org](mailto:permission@acm.org). © 2007 ACM 1553-3077/2007/03-ART1 \$5.00 DOI 10.1145/1227835.1227836 <http://doi.acm.org/10.1145/1227835.1227836>

(e.g., on paper) are now either converted to digital formats or are directly produced as such. Adding to this growth trend are storage-intensive media such as pictures and video. At the current time, magnetic disk storage is still preferred for many applications because of its performance and cost efficiency. However, while technological improvements for individual disk drives have increased the storage capacity per device and also improved their reliability, the capabilities of a single disk are insufficient for large data archives which are commonly built by combining many disks into disk arrays.

In their ground-breaking paper on RAID (Redundant Array of Independent Disks) Patterson et al. [1988] demonstrated that disk arrays can improve both performance (through parallelism) and fault tolerance (through redundant coding) compared to single large disks. To date, most high-performance storage systems are based on disk arrays and the basic algorithms have not changed all that much. Some new RAID levels, such as level 0 + 1 and 6 have been introduced, but out of the original RAID levels 1 through 5, both 1 and 5 are still widely used for fault tolerant systems [Gray and Shenoy 2000].

In order to design a high-performance disk array, it is imperative to understand the technological trends for magnetic disk drives. The capacity and speed of disks has improved steadily over the last decade. According to Gray and Shenoy's article on the trends in data engineering [Gray and Shenoy 2000], the storage capacity of magnetic disks has increased at the rate of about 60% per year. At the same time, the data transfer rate of magnetic disks has increased at a lesser rate of about 40% per year. Other sources have reported similar trends: the storage capacity of disks has accelerated to approximately 50% annually. The disk transfer rate (i.e., the bandwidth) follows a similar but less aggressive trend with an annual improvement of approximately 40% [Gray and Graefe 1997; Consortium 1998; Ng 1998; Grochowski 1999]. The increase in storage capacity decreases the cost-per-megabyte at a rate of approximately 40% per year [Grochowski 1999]. Higher recording densities and faster rotations of the disk platters (faster revolutions per minute) increase the data transfer rate of a disk drive. Thus, the imbalance in the growth rates between the disk space and data transfer rate has widened. One consequence of this trend is that the data transfer rate (i.e., bandwidth) is a scarce resource and one wants to optimize for bandwidth rather than for space [Gray and Shenoy 2000].

In this article, we present a technique called *Zoned-RAID (Z-RAID)* that is specifically designed to boost the effective bandwidth of a disk array. Z-RAID achieves this by taking advantage of a technical feature that is now common in all modern disk drives, *zoned recording (or zoning)*. This is an approach utilized by disk manufacturers to increase the storage capacity of magnetic disks [Ng 1998]. The technique groups adjacent disk cylinders into zones [Ruemmler and Wilkes 1994; Ng 1998]. Tracks are longer towards the outer portions of a disk platter compared to the inner portions, hence more data can be recorded in the outer tracks when the maximum linear density, that is, bits-per-inch, is applied to all tracks. A zone is a contiguous collection of disk cylinders whose tracks have the same storage capacity, that is, the number of sectors per-track is constant in the same zone. Hence, outer tracks have more sectors per-track than inner zones. Different disk models have different

numbers of zones. Even though zoning was originally introduced to increase the storage capacity of disk drives, it results in another important effect: different zones provide different transfer rates as a result of the following two facts. First, the storage capacity of the tracks for each zone is different, and second, the disk platters rotate at a fixed number of revolutions per second. We can observe a significant difference in data transfer rates between the minimum and maximum (approximately 50% difference) [Dashti et al. 2003; Ng 1998]. By intelligently placing data blocks (e.g., redundant and parity information) in the zones of each disk drive, one can increase the effective transfer rate of a storage array.

Z-RAID provides an improvement for any workload that is highly storage intensive. However, for a more focused discussion, we are presenting the advantages of our approach using an exemplar of such applications that requires not only large amounts of space, but also high data rates, namely, streaming media (SM). In recent years there has been a proliferation of multimedia databases especially for handling streaming media types such as digital audio and video. These media have become a part of everyday life, including not only electronic consumer products, but also online streaming media services on the Internet. Due to 1) successful standards for compression and file formats, such as (MPEG-4<sup>1</sup>), 2) increased network capacity for local area networks (LAN) and the Internet, and 3) advanced streaming protocols (e.g., RTSP<sup>2</sup>), more and more multimedia database applications, combined with the Internet, are providing streaming media services such as remote viewing of video clips.

Disk arrays have been the storage platform of choice for SM servers due to their high data transfer rate, large storage capacity, random access capability, and low price<sup>3</sup>. Consequently, many studies have investigated a hiccup-free display of streaming media using magnetic disk drives [Anderson and Homsy 1991; Berson et al. 1994; Berson et al. 1995; Chen and Little 1993; Gemmell et al. 1995; Rangan and Vin 1993; Rangan et al. 1992; Reddy and Wyllie 1994; Tobagi et al. 1993; Yu et al. 1993; Ghandeharizadeh et al. 1997].

Streaming media such as audio and video have two main characteristics. First, SM data must be displayed at a prespecified rate. For example, a commercial satellite broadcasting network, DirecTV, transmits a MPEG-2 encoded video stream at the rate of 4Mb/s for a TV channel<sup>4</sup> [Fogg 1995]. Any deviation from this real-time requirement may result in undesirable artifacts, disruptions, and jitters, collectively termed *hiccups*. Second, SM objects are large in size. For example, the size of a two-hour MPEG-2 encoded digital movie requiring 4Mb/s for its display is 3.6GB. Due to these characteristics, the overall design of servers in support of SM has been different from that of conventional

<sup>1</sup>The Motion Picture Expert Group (MPEG) has standardized several video and audio compression formats.

<sup>2</sup>The Real Time Streaming Protocol is an Internet Engineering Task Force (IETF)-proposed standard for the control of streaming media on the Internet.

<sup>3</sup>Magnetic disk drives provide a low-price solution for both storage capacity ( $\frac{\$}{MB}$ ) and bandwidth ( $\frac{\$}{MB/s}$ ).

<sup>4</sup>DirecTV changes its bit allocation per-program-channel based on the complexity of the video images. For example, a sports channel that includes quick movements and complicated images is assigned 6Mb/s.

databases [Tobagi et al. 1993; Gemmell et al. 1995]. However, both categories require high-performance storage solutions.

In large scale multimedia database servers in support of streaming media, it is critical to both optimize disk bandwidth and provide disk-based fault tolerance. Many studies [Heltzer and Menon 1993; Birk 1995; Ghandeharizadeh et al. 1996; Wang et al. 1997; Ghandeharizadeh and Kim 1999] have discussed data placement on multizone disks to maximize the effective data transfer rate. Kang and Yeom [2003] provided MRS (MultiRate Smoothing) data placement on multizone disks for a smooth transmission of variable-bit-rate data over a network. However, none of these studies cover the reliability issue. RAID has been widely used for fault-tolerant streaming servers as well as conventional file servers. Various reliability strategies in video servers, including RAID, were surveyed and compared in Gafsi and Biersack [2000]. However, no study considered one of the most important characteristics of modern disk drives, the variable data transfer rates from multiple zones in a disk. Therefore, they place data blocks without any constraints on a disk. This results in less optimized disk performance because the data transfer rate significantly varies depending on the location of a data block in multizone disks.

This study proposes a novel data placement scheme, Zoned-RAID (Z-RAID), to optimize the data transfer rate of a RAID system using multizone disks by constraining the data placement. We present the approach in the context of an example streaming media server application since it requires both a high data transfer rate as well as fault tolerance. The idea of combining the data placement of RAID with multizone disks was proposed in our previous work ([Dashti et al. 2005]) for the first time. It provided the basic idea of constrained data placement that stores primary data blocks (for normal access) in faster zones and secondary blocks (for standby in case of a disk failure) in slower zones. This study augments our previous work, including new extensions of various RAID levels (1, 5, and 6) and includes far more comprehensive experiments with both analytical and simulation results to compare the performance of RAID and Z-RAID systems. Our results demonstrate the superiority of Z-RAID to conventional RAID. Z-RAID provides a higher effective data transfer rate in normal mode with no disadvantage. In the presence of a disk failure, Z-RAID still performs as well as RAID.

## 2. Z-RAID

Since RAID [Patterson et al. 1988] was proposed in 1988, it has been widely implemented in many systems requiring fault tolerance. Originally, RAID levels 1-5 were proposed but many variants such as level 0 and 6 have been studied and commercialized. However, level 1 (mirroring) and 5 (block-based parity encoding) received most attention in many applications due to their cost effectiveness and implementation efficiency [Gray and Shenoy 2000]. Thus, this study focuses on extending RAID level 1, 5, and 6 to our proposed Zoned-RAID (Z-RAID) approach.

Z-RAID is based on the physical characteristics of multizone disks and their performance models. A multizone disk can be modeled in the following way. A

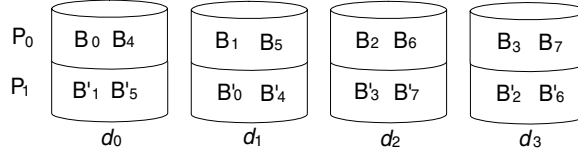


Fig. 1. Z-RAID level 1 with four disks.

disk with total space,  $S$ , has  $n$  zones, where zone 0 is the innermost (slowest) and zone  $n - 1$  is the outermost (fastest). The number of cylinders in each zone is  $Cyl(i)$ ,  $0 \leq i < n$ , and the total number of cylinders is  $Cyl$ . Cylinders are numbered from the innermost to the outermost. The size of a cylinder is  $S(i)$  bytes,  $0 \leq i < Cyl - 1$ . The data transfer rate of each cylinder is  $R_c(j)$ ,  $0 \leq j < Cyl$ , ( $R_c(0) \leq R_c(1) \leq \dots \leq R_c(Cyl - 1)$ ). Note that all cylinders in the same zone have the same data transfer rate. A rotational latency,  $l_{rot}$ , is one disk revolution time of a disk. A seek time between two locations in a disk, say  $x$  cylinders apart, can be calculated using a practical nonlinear approximation, *seek*( $x$ ) [Ruemmler and Wilkes 1994]. Then, an actual block retrieval time consists of a seek time, a rotational latency, and block reading time.

### 2.1 Z-RAID Level 1

RAID level 1 utilizes a replication of disks, called *mirroring*. When we have two disks,  $d_0$  and  $d_1$ , the primary copy of a block,  $B_i$ , is placed on  $d_0$  and the secondary copy, say  $B'_i$ , is placed on  $d_1$ . Blocks are arbitrarily distributed across cylinders inside a disk. This implies the system uses the average data transfer rate of a multizone disk and the average seek time (one half of the worst seek which is calculated from the outermost cylinder to the innermost cylinder). Then, the effective data transfer rate of a disk with no overhead (no seek time, no rotational latency) is:

$$R_R = \sum_{i=0}^{Cyl-1} \left( R_c(i) \times \frac{S(i)}{S} \right). \quad (1)$$

In a streaming media server whose access unit is a block ( $B$ ), each block access includes the worst seek time and average rotational latency to support real-time block retrieval even in the worst case [Gemmell et al. 1995; Yu et al. 1992; Ghandeharizadeh et al. 1997; Ozden et al. 1995]. Thus, the effective data transfer rate of RAID level 1 in a streaming media server is:

$$R_{RB} = \frac{B}{seek(Cyl) + l_{rot} + B/R_R}. \quad (2)$$

Z-RAID level 1 also uses mirroring in the same manner as RAID level 1. However, it utilizes only the faster zones of disks for primary copies of blocks. All secondary copies are placed on slower zones. With Z-RAID 1, each disk is divided into two logical partitions of equal size ( $P_0 = P_1 = S/2$ );  $P_0$  occupies the faster zones ( $S/2$  from the outermost cylinders), and  $P_1$  occupies the slower zones (remaining  $S/2$ ). All primary blocks,  $B_i$ , are assigned in  $P_0$ , while all secondary blocks,  $B'_i$ , are stored in  $P_1$  (see Figure 1). Let us say that  $P_0$  consists

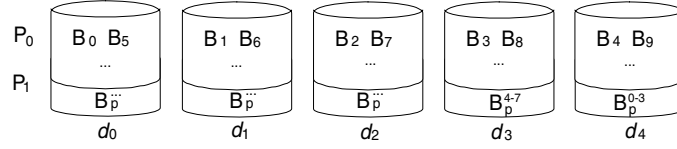


Fig. 2. Z-RAID level 5 with five disks.

of cylinders from  $m$  to  $Cyl - 1$ , where  $m$  is the cylinder number that divides the disk space in half (i.e.,  $\sum_{i=0}^{m-1} S(i) = S/2$ ). Note that the value of  $m$  and  $Cyl$  should be determined using real disk characteristics because different disk models have different zone characteristics. A more general allocation of blocks is as follows. When Z-RAID consists of  $k$  disks, if  $B_i$  resides on  $P_0$  of disk  $j$ ,  $B_i$  is stored in  $P_1$  of disk  $(j + 1) \bmod k$ .

Thus, in normal mode without disk failure, blocks are retrieved from  $P_0$ s of disks. Because  $P_0$ s are located in the faster zones of a disk, Z-RAID will increase the effective data transfer rate of the disk. Moreover, because the maximum cylindrical distance inside  $P_0$  is far shorter than  $Cyl$ , Z-RAID will decrease the required seek time between two adjacent block retrievals. Both will result in a significantly enhanced effective data transfer rate:

$$R_{ZR} = \sum_{i=m}^{Cyl-1} \left( R_c(i) \times \frac{S(i)}{S/2} \right) \quad (3)$$

$$R_{ZRB} = \frac{B}{seek(Cyl - m - 1) + lrot + B/R_{ZR}}. \quad (4)$$

## 2.2 Z-RAID Level 5

RAID level 5 uses a block-based parity encoding. It distributes parity blocks across disks in a parity group so that both normal blocks and parity blocks can be placed on a disk. Blocks are arbitrarily distributed in a disk. Thus, in normal mode, the effective data transfer rate of RAID level 5 is identical to RAID level 1, that is, Equations (1) and (2).

Z-RAID level 5 follows in the same manner as RAID level 5 to distribute parity blocks across disks. However, the location of parity blocks inside a disk is constrained to the slower zone areas. For example, when we form a parity group with 5 disks, 4 data blocks and a parity block will be distributed across 5 disks. Thus, 20% of each disk space consisting of corresponding innermost tracks will store all parity blocks, while 80% of the disk space with outer tracks store data blocks. For example, each disk has two logical partitions,  $P_0$  (outer 80% of disk space) and  $P_1$  (inner 20% space). Normal data blocks are stored in  $P_0$  and all parity blocks are in  $P_1$  (see Figure 2). The same advantages of Z-RAID level 1 in Section 2.1 are expected, namely, a higher effective data transfer rate and shorter average seek time in normal mode.

When  $d$  disks are in a parity group,  $1/d$  of each disk space will be used to store parity blocks. Then,  $P_0$  consists of cylinders from  $m$  (where  $\sum_{i=0}^{m-1} S(i) = S/d$ ) to  $Cyl - 1$ . Equations (3) and (4) for Z-RAID level 1 can be used for Z-RAID level 5 with a different value of  $m$  that is a function of  $d$ .

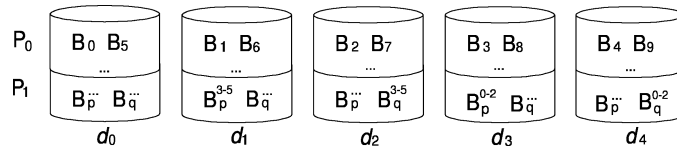


Fig. 3. Z-RAID level 6 with five disks.

### 2.3 Z-RAID Level 6

RAID level 6 is a variant of RAID level 5 which distributes parity blocks across disks in a parity group so that both normal blocks and parity blocks can be placed on a disk. However, RAID level 6 utilizes two types of parity blocks,  $P$  and  $Q$  [Chen et al. 1994]. The result is a significant increase in the reliability of the system so that it can tolerate two simultaneous disk failures in a parity group. Due to the extra parity block requirement, RAID level 6 is more expensive than RAID level 5. Because blocks are distributed within a disk, the effective data transfer rate of RAID level 6 is identical to RAID level 5 in normal mode.

Z-RAID level 6 follows the same method as RAID level 6 in distributing parity blocks across disks. However, as in Z-RAID level 5, the location of parity blocks inside a disk is constrained to the slower zone areas. For example, when we form a parity group with 5 disks, 3 data blocks and 2 parity blocks ( $P$  and  $Q$ ) will be distributed across 5 disks. Thus, 40% of each disk space consisting of corresponding innermost tracks will store all parity blocks, while 60% of the disk space with outer tracks store data blocks. Each disk has two logical partitions,  $P_0$  (outer 60% of disk space) and  $P_1$  (inner 40% space). Normal data blocks are stored in  $P_0$ , and all parity blocks are in  $P_1$  (see Figure 3).

Formally, when  $d$  disks are in a parity group,  $2/d$  of each disk space will be used to store parity blocks. Then,  $P_0$  consists of cylinders from  $m$  (where  $\sum_{i=0}^{m-1} S(i) = 2S/d$ ) to  $Cyl - 1$ . Equations (3) and (4) for Z-RAID level 1 can be used for Z-RAID level 6 with a different value of  $m$  that is a function of  $d$ .

### 2.4 Z-RAID for Streaming Media Applications

Because Z-RAID can provide a higher effective data transfer rate with the same fault-tolerant disk system compared to a conventional RAID, it can be used wherever a RAID can be used. However, some applications such as streaming applications that require a large page (block) size benefit the most from Z-RAID because a block retrieval time depends more on data transfer time than other near constant factors, that is, seek time and rotational latency. Note that  $B/R_{ZR}$  becomes a dominant factor (see the denominator in Equation (4)) as  $B$  grows larger.

Another important observation in real streaming applications is that objects may have different popularity or access frequency. For example, in a movie-on-demand system, more than half of the total user requests might reference only a handful of recently released hot movies. It is widely understood that the popularity distribution among objects in video-on-demand systems can be well represented by the Zipf distribution [Nussbaumer et al. 1995], which is a very skewed distribution.

Table I. Parameters for two Seagate Disks

Model	ST336752LC	ST3200822A
Series	Cheetah X15	Barracuda 7200.7 plus
Manufacturer	Seagate Technology	Seagate Technology
Capacity $S$	37 GB	200 GB
Transfer rate $R_c$	See Table 4.a	See Table 4.b
Spindle speed	15,000 rpm	7,200 rpm
Avg. rotational latency	2 msec	4.16 msec
Worst case seek time	7.2 msec	15 msec

Zone #	Size (GB)	Read Transfer Rate (MB/s)
0	12	57.5
1	3.5	55.4
2	3.0	54.7
3	4.0	52.7
4	3.0	50.6
5	2.5	48.1
6	3.0	45.6
7	2.5	43.6
8	2.5	41.9

Zone #	Size (GB)	Read Transfer Rate (MB/s)
0	48	65.2
1	17	63.8
2	14	61.5
3	21	58.2
4	9	56.0
5	12	54.1
6	14	52.4
7	9	50.6
8	6	49.5
9	13	46.8
10	9	44.1
11	6	42.2
12	8	39.7
13	8	37.6
14	6	35.3

(a) Cheetah X15

(b) Barracuda 7200.7

Fig. 4. Zoning information of two Seagate disks.

Z-RAID can well take advantage of this skewed popularity distribution because the distribution of data transfer rates across zones is also skewed. With  $n$  objects in the system, one can sort objects in descending order based on their popularity. Then one assigns blocks of objects from the outermost tracks in a disk which has the fastest data transfer rate towards the inner tracks, track-by-track. When the blocks of the first, most popular, object are all assigned, then the next object is assigned in the same way from the next track. This process is repeated until all objects are assigned.

### 3. COMPARISONS

In our experiments, we used two Seagate disk drives, Cheetah X15 and Barracuda 7200.7 plus. Cheetah X15 provides one of the fastest revolution speeds, 15000 revolutions per-minute (RPM) with a very short average seek time, 3.6ms. This has been a typical high-performance disk in the market for years (introduced in 2000). Barracuda 7200.7 is a typical cost-effective high-capacity disk with 7,200 RPM and 8.5ms of average seek time (introduced in 2004). Table I and Figure 4 show zone characteristics of Cheetah X15 and Barracuda 7200.7.

### 3.1 Analytical Comparison

First, we calculated and compared the effective data transfer rates of RAID and Z-RAID with the two disk drives in Table I using equations in Section 2.1 and 2.2. Our comparison assumed two typical approaches widely used in the design of streaming media servers. First, in guaranteed approaches that support 100% hiccup-free displays, one should assume the worst seek time and the worst rotational latency per each data block retrieval. Many round-robin data placement and retrieval schemes [Yu et al. 1993; Berson et al. 1994; Ghandeharizadeh and Kim 1995; Tobagi et al. 1993; Ghandeharizadeh et al. 1997; Ozden et al. 1995] followed the guaranteed approaches so they fall into the category of the worst-case analysis. To quantify the effective data transfer rates of these approaches, we performed a worst-case analysis assuming the worst seek (7.2ms for Cheetah X15 and 15ms for Barracuda 7200.7) and worst-rotational latency (4ms for Cheetah X15 and 8.3ms for Barracuda 7200.7) in the calculations. Second, in statistical approaches that allow a nonzero hiccup probability, one can take advantage of the average seek time and average rotational latency per each data block retrieval. Many random data placement and retrieval schemes [Muntz et al. 1997; Kim 2001] followed this statistical approach to enhance the performance of the system at an expense of a minor degradation of display quality, that is, occasional hiccups. For these approaches, we performed an average case analysis assuming the average seek (3.6ms for Cheetah X15 and 8.5ms for Barracuda 7200.7) and average rotational latency (2ms for Cheetah X15 and 4.16ms for Barracuda 7200.7) in the calculations.

It is a well-known fact that the performance of streaming media servers, especially the performance of their disk subsystems, significantly varies depending on the size of the data block that is the unit of access to the disks [Ghandeharizadeh et al. 1995; Ghandeharizadeh and Kim 1995; Ozden et al. 1995; Yu et al. 1993]. Thus, we calculated the effective data transfer rates while varying the size of the data blocks from 128KB to 8MB which are most reasonable ranges for streaming media servers.

Figures 5 and 6 show the effective data transfer rates of RAID and Z-RAID with Cheetah X15. RAID1 denotes the traditional RAID level 1, Z-RAID1 means the proposed Z-RAID level 1, and Z-RAID5 means the proposed Z-RAID level 5. Note that the effective rate of RAID5 is identical to that of RAID5 in normal mode because all data blocks are arbitrarily distributed across all zones without any constraints. In our calculation, the size of parity group of Z-RAID5 was 5 disks so that 20% of the disk space (from the slowest zone) in each disk is dedicated to store parity blocks. Figures 5 and 6 are results from the worst-case analysis and the average case analysis, respectively. Compared to RAID1, Z-RAID1 demonstrates enhancement in rates from 10.5% to 38.6% in the worst-case analysis and from 9.5% to 33.1% in the average case analysis. Compared to RAID5, the percentage enhancement of Z-RAID5 ranges from 4.8% to 12.7% in the worst-case analysis and from 4.5% to 11.4% in the average case analysis. Figures 7 and 8 show analytical results with Barracuda 7200.7. The results and trends are similar to those for Cheetah X15. Z-RAID1 outperforms RAID1 with 18.5% to 46.8% of rate enhancement in the worst-case analysis, and from

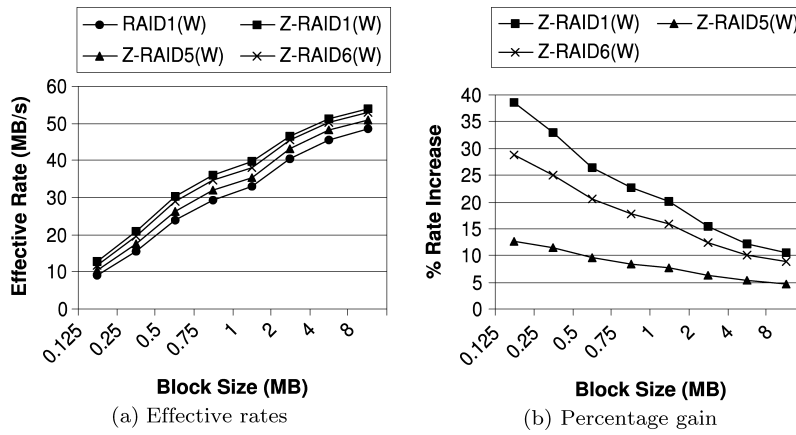


Fig. 5. Effective data rate of a Seagate X15 disk (worst analysis).

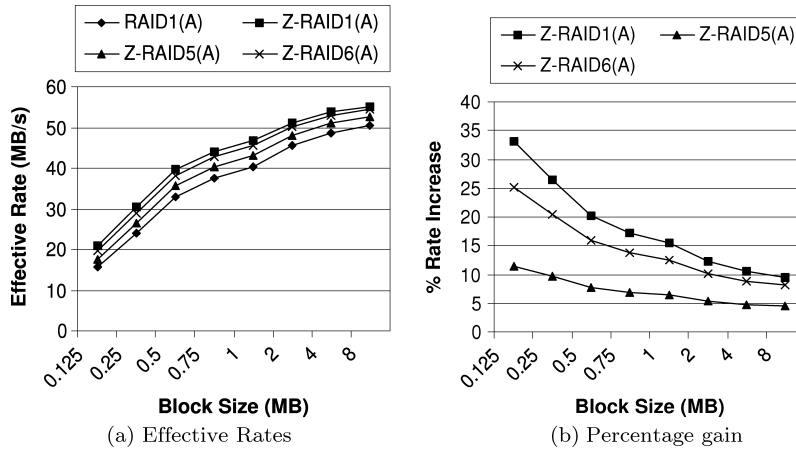


Fig. 6. Effective data rate of a Seagate X15 disk (average analysis).

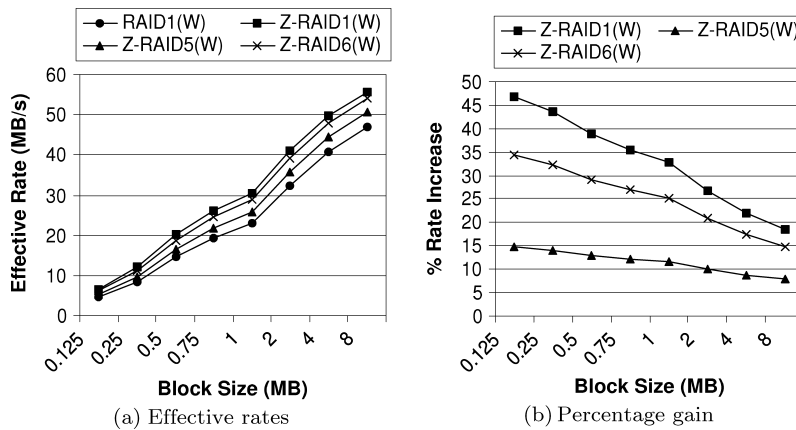


Fig. 7. Effective data rate of a Seagate 7200.7 disk (worst analysis).

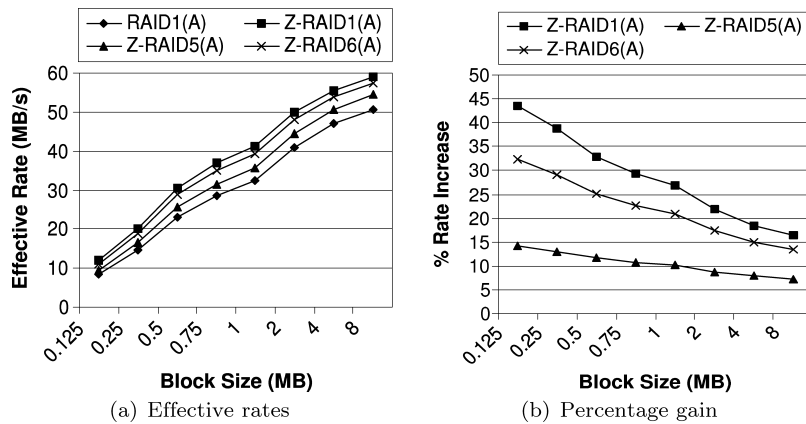


Fig. 8. Effective data rate of a Seagate 7200.7 disk (average analysis).

16.5% to 43.6% in the average case analysis. Compared to RAID5, the percentage enhancement of Z-RAID5 ranges from 7.9% to 14.7% in the worst-case analysis, and from 7.3% to 14.1% in the average case analysis. Z-RAID6 also shows similar trends compared to RAID6.

For all comparisons, Z-RAID outperforms RAID. The percentage enhancement of effective data transfer rate is greater when the size of the block is smaller. This is because reduced seek time is the dominant factor in determining the effective rate when the block size is small. The performance enhancement diminishes as the block size increases because the dominant factor shifts from seek time to actual block-reading time, see the divisors in Equations (2) and (4). The reduced seek time, due to the confined disk access within a region, is also the reason why Z-RAID1 gains a higher percentage increase than Z-RAID5 and Z-RAID6, that is, a shorter seek time. With Z-RAID 5 and 6, the performance enhancement decreases as the size of parity group, increases. When we had a smaller group, such as three disks, we achieved a higher effective rate than with a larger group.

### 3.2 Simulation Results

The analytical models of the previous section provide some compelling evidence that the Z-RAID method can provide an increased performance. However, they cannot encompass the full complexity of a storage system, and hence are based on some arguable simplifying assumptions. Thus, to further evaluate the performance of the Z-RAID technique, we implemented a simulator. It included a detailed disk model that was calibrated with parameters extracted from commercially available disk drives. To model user behavior, the simulator included a module to generate synthetic workloads based on various Poisson and Zipf distributions [Zipf 1949].

**3.2.1 Simulation Infrastructure.** The simulator was implemented using the C programming language on a Sun server running Solaris. Other than the standard libraries, no external support was needed. The simulator was

Table II. Experimental Parameters for the Z-RAID Level 1 Simulator

Z-RAID Level 1	18 Disks (Seagate Cheetah X15)
Block size $\mathcal{B}$	0.25, 0.5, 1, 2, 4, 8MB
Time period $T_p$	$(\frac{\mathcal{B}}{1.5\text{Mb/s}})$ sec
Throughput $\mathcal{N}_{Tot}$	<4800
No. of stored clips	47
Object type	MPEG-1 (1.5 Mb/s)
Object size (length)	675 MB (1 h)
Access distribution	Zipf

implemented in a modular fashion and consists of the following components.

- (1) The *disk emulation* module imitates the response and behavior of a magnetic disk drive. The level of detail of such a model depends largely on the desired accuracy of the results. Several studies have investigated disk modeling in great detail [Ruemmler and Wilkes 1994; Worthington et al. 1994; Ghandeharizadeh et al. 1995]. Our model includes mechanical positioning delays (seek time and rotational latency) as well as variable transfer rates due to the common zone-bit-recording technique.
- (2) The *file system* module provides the abstraction of files on top of the disk models and is responsible for the allocation of blocks and the maintenance of the free space. We selected the Everest file system for its flexibility and real-time performance [Ghandeharizadeh et al. 1996]. Either random or constrained block allocation were selectable with this file system.
- (3) The *loader* module generates a synthetic set of continuous media objects that are stored in the file system as part of the initialization phase of the simulator.
- (4) The *scheduler* module translates a user request into a sequence of real-time block retrievals. It implements the concept of a time period and enables the round-robin migration of consecutive block reads on behalf of each stream. Furthermore, it ensures that all real-time deadlines are met.
- (5) Finally, the *workload* generator models user behavior and produces a synthetic trace of access requests to be executed against the stored objects. Both the distribution of the request arrivals as well as the distribution of the object access frequency can be individually specified. For the purpose of our simulations, the request interarrival times were Poisson distributed, while the object access frequency was modeled according to Zipf's law [Zipf 1949].

**3.2.2 Results.** For the evaluation of RAID and Z-RAID level 1, the simulator was configured with a total of 18 disks of the Seagate Cheetah X15 (model ST336752LC), each with 37GB of space. Table II summarizes the rest of the simulation parameters.

We compared RAID level 1 with Z-RAID level 1. For regular mirroring, the data blocks were randomly distributed across all the zones of a disk. For Z-RAID mirroring, the primary copies of the data was constrained to the faster half of

the disk drives. We tested retrieval block sizes of 0.25, 0.5, 1, 2, 4, and 8MB, and we executed the simulation with a nominal workload of  $\lambda = 1900, 1800, 1700$  requests per-hour. The simulated database consisted of video clips whose display time was one hour long and which required a constant retrieval rate of 1.5Mb/s (e.g., MPEG-1). This resulted in a uniform storage requirement of 675MB per clip. We also performed simulation of RAID and Z-RAID level 5,6 with the parity group size 5,6.

The frequency of access to different media clips is usually quite skewed for a video-on-demand system, that is, a few newly released movies are very popular, while most of the rest are accessed infrequently. The distribution pattern can be modeled using Zipf's law [Zipf 1949], which defines the access frequency of movie  $i$  to be  $F(i) = \frac{c}{i^{1-d}}$ , where  $c$  is a normalization constant and  $d$  controls how quickly the access frequency drops off. In our simulations,  $d$  was set to equal 0.271. This value has been shown to approximate empirical data for rental movies [Dan et al. 1994]. For each experiment, the server had to service requests that arrived based on a Poisson distribution.

We focused on the disk utilization to compare the two techniques. A lower disk utilization, given a fixed workload, indicates a higher effective data transfer rate and a higher maximum throughput, for the overall system. Because the effective bandwidth of a disk drive increases with larger block sizes, we expected to see a drop in disk utilization with increased block sizes.

Figure 9 (a), (b), and (c) show the results of the simulations using 18 Cheetah X15 disks, which depicts the reduction of the overall disk utilization of Z-RAID level 1, 5, and 6 with a constant workload as compared with standard RAID level 1, 5, and 6. Z-RAID 1, 5, and 6 outperformed RAID 1, 5 and 6, respectively. For example, when the block size is 1MB in Figure 9 (b), the disk utilization of RAID1 was 49.7%, while that of Z-RAID1 was 45.7% to service the same number of requests. The average percentage reduction of disk utilization between Z-RAID1 and RAID1 ranges around 8.5% in Figures 9 (a)-(c). The utilization of Z-RAID5 was higher than that of Z-RAID1 but still lower than that of RAID5 (5% average reduction).

We performed more simulations with different configuration using Barracuda 7200.7 (model ST3200822A). We used 33 disks and the workload was the same,  $\lambda = 1,600, 1,500, 1,400$  requests per-hour. Figure 10 (a)-(c) show similar results as previous simulations with Cheetah X15. The average percentage reduction of disk utilization between Z-RAID1 and RAID1 ranges around 12.5%. The overall performance enhancement as a form of utilization reduction was a little lower than expected in the analytical comparison.

We also compared Z-RAID level 5 with ZRAID level 6 in an array of 4, 5, 6 disks in Figure 9 (d)-(f) and Figure 10 (d)-(f). The utilization of Z-RAID level 6 is around 2% lower than that of Z-RAID level 5.

Finally, we directly compared the performance of two disks using RAID 1 and Z-RAID1. The configuration changed to 18 disks and the workload was  $\lambda = 1,500$  requests per-hour. Comparing results from two disks, see Figure 11, X15 provided a lower utilization than 7200.7 when the block size is small. This is because the overhead portion (seek time + rotational latency) is larger than the actual block reading time when the block size is small. X15 has a far shorter seek

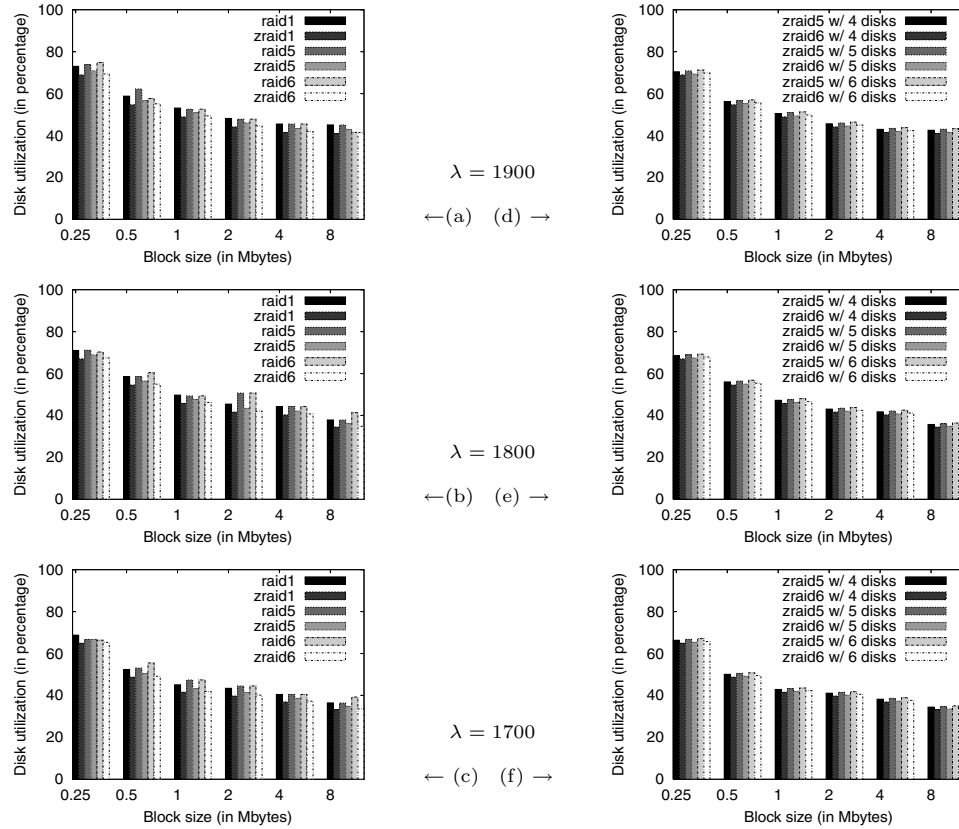


Fig. 9. Simulation results using Cheetach X15 disks. RAID vs ZRAID is in the left column, ZRAID 5 vs ZRAID 6 is in the right column.

time and rotational latency than 7200.7. However, as the block size increases, the portion of block reading time increases and becomes the dominant factor. Because 7200.7 has a higher effective data transfer rate, it provided a smaller utilization when the blocks are big.

#### 4. CONCLUSION

Our proposed Z-RAID system constrains the data block placement in RAID system utilizing the zone characteristics of multizone disk drives. The constrained data placement and retrieval incur a shorter seek time between two adjacent block retrievals, which results in a reduced overhead in block retrieval time. Moreover, because the blocks are retrieved from the faster zones of a disk, the effective data transfer rate is increased further. Our analytical and experimental results in streaming media server application demonstrate that all Z-RAID level 1, 5, and 6 outperform the traditional RAID level 1, 5, and 6, respectively. Z-RAID might need a careful allocation of blocks and a little overhead in writing. However, Z-RAID doesn't require any overhead in reading blocks and significantly increases the effective data transfer rate. Thus, Z-RAID can

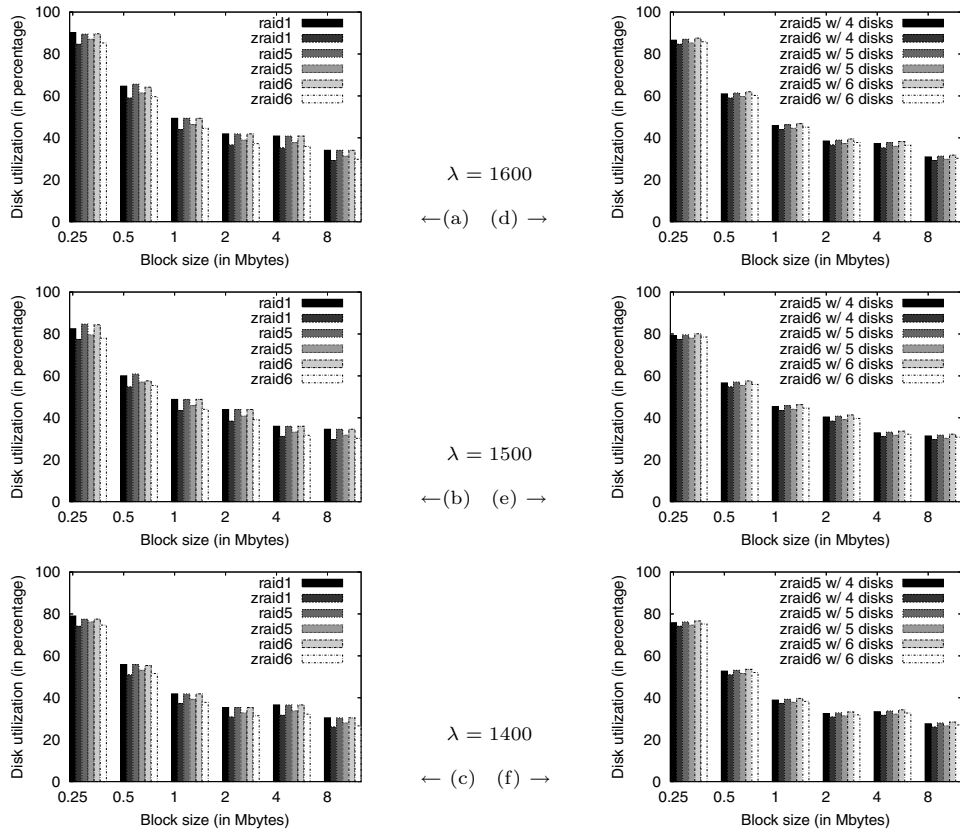


Fig. 10. Simulation results using Barracuda disks. RAID vs ZRAID in the left column, ZRAID 5 vs ZRAID 6 in the right column.

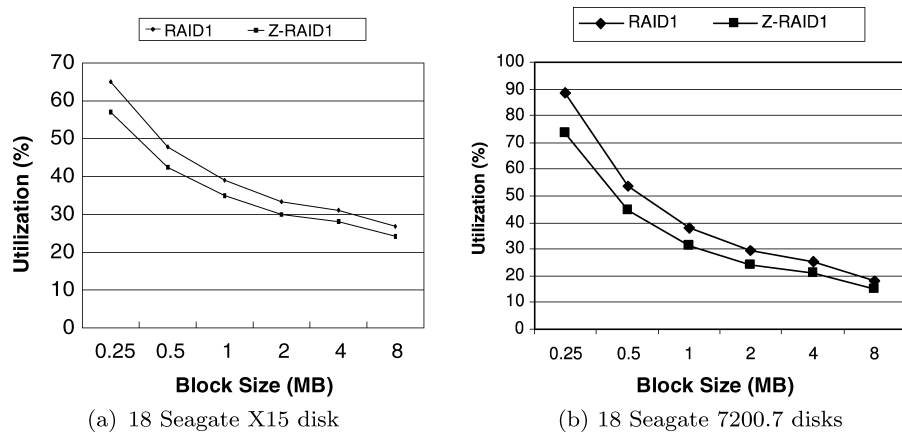


Fig. 11. Disk utilization.

enhance the performance of a read-intensive disk subsystem such as in the streaming media server application.

One different aspect of Z-RAID can result in a more cost-effective and affordable system. Typically, RAID has been constructed using high-performance disk drives such as SCSI disks. These disks, in general, provide a higher transfer rate than other inexpensive disks such as IDE disks. The drawback is a higher price. For cost effectiveness, a more economical RAID with IDE disks (IDE-RAID) has been introduced. Given that a Z-RAID system with IDE disks provides as high performance as a RAID system with high-end SCSI disks, IDE Z-RAID can achieve the performance of SCSI-RAID with a cost of IDE-RAID. Considering the recent trend that the performance gap between expensive SCSI disks and economical IDE disks is getting narrower (while the price gap still remains very significant), Z-RAID would provide an even better solution for a disk subsystem with inexpensive disks.<sup>5</sup>

#### REFERENCES

- ANDERSON, D. AND HOMSY, G. 1991. A continuous media I/O server and its synchronization. *IEEE Comput.* 24, 3.
- BERSON, S., GHANDEHARIZADEH, S., MUNTZ, R., AND JU, X. 1994. Staggered striping in multimedia information systems. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 79–89.
- BERSON, S., GOLUBCHIK, L., AND MUNTZ, R. R. 1995. A fault tolerant design of a multimedia server. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 364–375.
- BIRK, Y. 1995. Track-pairing: a novel data layout for vod servers with multi-zone-recording disks. In *IEEE International Conference on Multimedia Computing and Systems*.
- CHEN, H. AND LITTLE, T. 1993. Physical storage organizations for time-dependent multimedia data. In *Proceedings of Foundations of Data Organization and Algorithms Conference (FODO)*.
- CHEN, P., LEE, E., GIBSON, G., KATZ, R., AND PATTERSON, D. 1994. RAID: High-performance, reliable secondary storage. *ACM Comput. Surv.* 26, 2 (June), 145–185.
- NSI CONSORTIUM. 1998. Tape Roadmap.
- DAN, A., SITARAM, D., AND SHAHABUDDIN, P. 1994. Scheduling policies for an on-demand video server with batching. In *Proceedings of ACM Multimedia*. 391–398.
- DASHTI, A., KIM, S. H., SHAHABI, C., AND ZIMMERMANN, R. 2003. *Streaming Media Server Design*. Prentice Hall.
- DASHTI, A., KIM, S. H., AND ZIMMERMANN, R. 2005. Zoned-raid for multimedia database servers. In *Proceedings of the 10th International Conference on Database Systems for Advanced Applications (DASFAA '06)*, Lecture Notes in Computer Science, vol. 3453, 461–473.
- FOGG, C. 1995. DSS and MPEG technical notes. <http://www.dbs-online.com/d-mpitech.htm>.
- GAFSI, J. AND BIERSACK, E. W. 2000. Modeling and performance comparison of reliability strategies for distributed video servers. *IEEE Trans. Parallel. Distrib. Sys.* 11, 4, 412–430.
- GEMMELL, D. J., VIN, H. M., KANDLUR, D. D., RANGAN, P. V., AND ROWE, L. A. 1995. Multimedia storage servers: A tutorial. *IEEE Comput.* 28, 5.
- GHANDEHARIZADEH, S., IERARDI, D., AND ZIMMERMANN, R. 1996. An on-line algorithm to optimize file layout in a dynamic environment. *Inform. Process. Lett.* 2, 57, 75–81.
- GHANDEHARIZADEH, S. AND KIM, S. 1995. Striping in multi-disk video servers. In *SPIE Proceedings High-Density Data Recording and Retrieval Technologies*. vol. 2604, 88–102.
- GHANDEHARIZADEH, S. AND KIM, S. 1999. A comparison of alternative continuous display techniques with heterogeneous disks. In *Proceedings of the International Conference on Information and Knowledge Management*. 442–449.

<sup>5</sup>Z-RAID might revive the original name of RAID (Redundant Array of **In**expensive Disks).

- GHANDEHARIZADEH, S., KIM, S., SHI, W., AND ZIMMERMANN, R. 1997. On minimizing startup latency in scalable continuous media servers. In *SPIE Proceedings of Multimedia Computing and Networking*. vol. 3020, 144–155.
- GHANDEHARIZADEH, S., KIM, S. H., AND SHAHABI, C. 1995. On configuring a single disk continuous media server. In *Proceedings of ACM SIGMETRICS*.
- GHANDEHARIZADEH, S., KIM, S. H., SHAHABI, C., AND ZIMMERMANN, R. 1996. Placement of continuous media in multi-zone disks. In *Multimedia Information Storage and Management*, S. M. Chung, Ed. Kluwer Academic Publishers, Boston, MA.
- GHANDEHARIZADEH, S., STONE, J., AND ZIMMERMANN, R. 1995. Techniques to quantify SCSI-2 disk subsystem specifications for multimedia. Tech. rep. USC-CS-TR95-610, University of Southern California.
- GRAY, J. AND GRAEFE, G. 1997. The five-minute rule ten years later, and other computer storage rules of thumb. *ACM SIGMOD Record* 26, 4 (Dec.).
- GRAY, J. AND SHENOY, P. 2000. Rules of thumb in data engineering. In *Proceedings of IEEE International Conference on Database Engineering*.
- GROCHOWSKI, E. 1999. IBM leadership in disk storage technology. IBM Almaden Research Center. <http://www.storage.ibm.com/technolo/grochows>.
- KANG, S. AND YEOM, H. 2003. Storing continuous media objects to multi-zone recording disks using multi-rate smoothing technique. *IEEE Trans. Multim.* 5, 3, 473–482.
- KIM, S. H. 2001. Bulk prefetching with deadline-driven scheduling to minimize startup latency of continuous media servers. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'01)*.
- MUNTZ, R., SANTOS, J., AND BERSON, S. 1997. RIO: A real-time multimedia object server. *ACM Sigmetrics Perform. Evalu. Rev.* 25, 2 (Sept.).
- NG, S. W. 1998. Advances in disk technology: Performance issues. *IEEE Comput. Mag.* 75–81.
- NUSSBAUMER, J., PATEL, B., SCHAFFA, F., AND STERNBENZ, J. 1995. Network requirement for interactive video-on-demand. *IEEE Trans. Selec. Areas Comm.* 13, 5, 779–787.
- OZDEN, B., RASTOGI, R., AND SILBERSCHATZ, A. 1995. Disk striping in video server environments. In *IEEE International Conference on Multimedia Computing and System*.
- PATTERSON, D., GIBSON, G., AND KATZ, R. 1988. A case for redundant arrays of inexpensive disks (RAID). In *Proceedings of the ACM SIGMOD International Conference on Management of Data*.
- RANGAN, P. AND VIN, H. 1993. Efficient storage techniques for digital continuous media. *IEEE Trans. Knowl. Data Engin.* 5, 4 (Aug.).
- RANGAN, P., VIN, H., AND RAMANATHAN, S. 1992. Designing an on-demand multimedia service. *IEEE Comm. Mag.* 30, 7 (July).
- REDDY, A. L. N. AND WYLLIE, J. C. 1994. I/O issues in a multimedia system. *IEEE Comput. Mag.* 27, 3 (March), 69–74.
- RUEMLER, C. AND WILKES, J. 1994. An introduction to disk drive modeling. *IEEE Comput.* 27, 3.
- S.R. HELTZER AND J.M. MENON, M. M. 1993. Logical data tracks extending among a plurality of zones of physical tracks of one or more disk devices. U.S. Patent No. 5,202,799.
- TOBAGI, F., PANG, J., BAIRD, R., AND GANG, M. 1993. Streaming RAID-A disk array management system for video files. In *Proceedings of the 1st ACM Conference on Multimedia*.
- WANG, Y. C., TSAO, S. L., CHANG, R. I., CHEN, M. C., HO, J. M., AND KO, M. T. 1997. Fast data placement scheme for video servers with zoned-disks. In *SPIE Proceedings of Multimedia Storage and Archiving Systems II*. vol. 3229, 92–102.
- WORTHINGTON, B. L., GANGER, G. R., AND PATT, Y. N. 1994. Scheduling algorithms for modern disk drives. In *Proceedings of ACM SIGMETRICS*. 241–251.
- YU, P., CHEN, M.-S., AND KANDLUR, D. 1993. Grouped sweeping scheduling for DASD-based multimedia storage management. *Multimed. Syst.* 1, 1 (Jan.), 99–109.
- YU, P. S., CHEN, M. S., AND KANDLUR, D. D. 1992. Design and analysis of a grouped sweeping scheme for multimedia storage management. In *Proceedings of the 3rd International Workshop on Network and Operating System Support for Digital Audio and Video*.
- ZIPF, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading MA.

Received September 2006; accepted January 2007