DISK SHADOWING

Dina Bitton[1]

Department of Electrical Engineering and Computer Science
University of Illinois at Chicago

Jim Gray

Tandem Computers
Cupertino California

Abstract

Disk shadowing is a technique for maintaining a set of two or more identical disk images on separate disk devices. Its primary purpose is to enhance reliability and availability of secondary storage by providing multiple paths to redundant data. However, shadowing can also boost I/O performance. In this paper, we contend that intelligent device scheduling of shadowed disks increases the I/O rate, by allowing parallel reads and by substantially reducing the average seek time for random reads. In particular, we develop an analytic model which shows that the seek time for a random read in a shadow set is a monotonic decreasing function of the number of disks in the set.

1. Introduction

Disk shadowing is a technique used to enhance availability and reliability of secondary storage. It consists of dynamically creating and maintaining a set of two or more identical disk images on different disks coupled as as a *mirrored disk* (two disks) or a *shadow set* (two or more disks). One or more hosts can be connected to a shadow set, which they consider as a single disk device. When a host directs a write request to the shadow set, the data is written to all disks in the shadow set. A read request is executed by reading from any disk in the set.

The primary purpose of shadowing is to provide a fault-tolerant and highly available mass-storage system, by duplicating hardware resources and maintaining multiple copies of the data. Shadowed disks provide online backup storage, thus reducing the need for periodic offline backup procedures. They also continue to provide access to data as long as at least one disk in the shadow set is available.

A less obvious advantage of shadowing is that it can also boost I/O performance. By providing multiple paths to duplicate data, a shadow set can service multiple read requests in parallel. Furthermore, it can reduce access time for random reads by optimizing the choice of the disk to which a read is assigned. As a consequence, shadowed disks provide higher I/O service rates and lower average access times for random reads than a single disk. With proper configuration of controllers and data paths (Section 2.1), writes to all disks in a shadow set can be executed in parallel. Then writes can be serviced at a rate similar to a single disk. Thus, in spite of the hardware cost, shadowing may be a viable technique for coupling disks in systems that require both high reliability and increased I/O performance.

Other approaches that are being explored for obtaining higher I/O rates by coupling multiple disks are *disk striping* [SG86] and synchronous *disk interleaving* [Ki86]. These techniques increase the I/O bandwidth, but do not provide a fault-tolerant storage system. Another recently proposed technique consists of interleaving disks and using additional disks to store redundant information [PGK87]. This technique, termed *RAID* for Redundant Arrays of Inexpensive Disks, promises to enhance both performance and reliability in a cost effective manner. However, further investigation is needed to determine the proper balance of interleaving and redundancy in a *RAID*, and evaluate its performance.

In this paper, we concentrate on pure shadowing, which is a fully redundant scheme for coupling two or more magnetic disks. We briefly describe the functions required to maintain a shadow set, and investigate the performance advantages of shadowing. In particular, we estimate the expected seek time in shadowed disks, and show that for read requests, it decreases as the inverse of the number of members in the shadow set.

[1] This research was partially supported by the National Science Foundation under grant #8704434 and by a grant from Argonne National Laboratory.

2. How a shadowed disk works

The functions required to support shadowing can be implemented in the disk driver software on the host(s), or in hardware, in a dedicated mass-storage server. The first approach (Figure 1) was chosen in Tandem's mirrored disks [Sit86]. The second approach (Figure 2) was implemented in the DEC HSC50 server, an intelligent controller which can manage up to 24 disks in one shadow set [BT85].

2.1. Controller configuration

With both approaches, there are different possible configurations depending on the number of disk controllers and access paths. Shadowing implies added I/O overhead at 3 levels: host CPU, channel, and controller. With a single controller configuration, the controller is a single point of failure and controller contention may become a bottleneck, since every write request is interpreted as a write for each disk in the shadow set.

For reliability and performance reasons, disks should be dual ported and connected to a pair of controllers (Figure 1). A controller pair, or a server pair (Figure 2) can support one or more shadow sets.

Having multiple controllers and configuring them properly is also a major factor in the performance of a shadow set. In order to support parallel reads and writes to the disks in a shadow set, a preferred controller should be designated for each disk, or for a subset of disks. The nonpreferred controllers will be used only in case of a failure. Providing the necessary paths for parallel writes is especially critical since a write must always be duplicated to all disks. With parallel access to all disks in the shadow set, the time for a write will be the maximum of the times required by individual disks, instead of being their sum. For reads, the availability of multiple data paths provides true parallelism: Multiple read requests can be serviced in parallel, since a read need only be executed on one disk.

2.2. Recovering from failure

When a failure occurs in one of the disk drives, the shadow set continues to provide access to the data from the other disk(s) in the set. Disks can be removed from or added to a shadow set. To replace a disk that failed, a new disk can be assigned to the set, and an image of the data can be copied from another disk in the shadow set. There are two options for copying. The first is conventional offline copying, which requires losing availability of the mass-storage system during the time of the operation (typically 10 or 15 minutes). The second is online copying, which can be supported by adding a function to the disk server. During online copying, new data is written to the disks in the current set and to the new disk; Reads are made from the current shadow set or, if the data to be read has already been copied, from the new disk.

Shadowing also solves the "bad spot" problem. If a bad sector is encountered when reading from one disk, the read is reassigned to another disk in the set. The bad sector can be subsequently rewritten.

3. Two or more copies

Disk mirroring is commonly used for improving reliability. An interesting question is whether it makes sense to have more than 2 disks in a shadow set. In this section, we argue that 2 copies are sufficient to provide a very high level of reliability, but that more than 2 copies can substantially improve performance.

3.1. Reliability of a shadow set

With current technology, the mean time between failures (*MTBF*) of a disk is rated between 3 and 5 years. Assuming independent and exponential times to failure for the k disks in a shadow set, the time until the first failure has a mean equal to $MTBF/k$ (see for example [MGB74]). However, since a single disk failure does not make the shadow set unavailable, a shadow set should be considered to fail only if after the first failure, the other disks fail during the time it takes to repair or replace the first disk. This window represents the time to replace the bad disk with a new disk and "revive" the mirror. It may vary from 15 minutes, the time for a copy operation, if spare disks are kept in standby, to several hours.

For reliability purposes only, having two disks in a shadow set, or mirroring disks, is practically sufficient, since the probability of two disks with two independent controllers failing in a small time window is almost null. As an example, suppose that the failure time of a single disk is exponentially distributed with a mean of five years, and that the time to repair the mirror set *MTTR* is 3 hours. After one disk failed, the probability of the its mirror failing during the next 3 hours will only be 6×10^{-5} (see Appendix). The *MTBF* of a mirrored disk is much smaller than the time to the first disk failure. It is given by

$$MTBF_{mirror} = \frac{MTBF}{2} * \frac{MTBF}{MTTR}$$

This expression can be formally derived (see for example [MGB74]). Its intuitive meaning is that the mean time to failure of the mirror is the mean time till the first failure $MTBF/2$ multiplied by the inverse of the probability of a second failure during the repair time, which is equal to $MTTR / MTBF$. With a 5 years *MTBF* and a 3 hours *MTTR*, the mean time between failures of a mirrored disk, $MTBF_{mirror}$, will be more than 30,000 years!

o

3.2. Performance of a shadow set

From a performance point of view, it may be effective to have shadow sets with more than two disks. Having k disks in a shadow set, with a data path to each disk, may increase the I/O service rate by a factor of k for reads, while maintaining approximately the same I/O rate for writes. The actual speedup would depend on the pattern of the request arrivals, their scheduling, and the server's capabilities, and thus be lower than k.

For example, in a benchmark of a shadow set with 4 disks, supported by the DEC HSC50 server, it was found that shadowing provided a service rate of 100 I/O's per second to a VAX-11/780 host, a 3 fold increase from the I/O service rate of a single disk [BT85]. In a multiprocessor environment, it is even more likely that shadow sets with a larger number of disks can be instrumental in further increasing the number of I/O requests serviced per time unit by utilizing the disks in parallel.

Another reason for having shadow sets with more than two disks is the potential for reducing random access time. In non-sequential I/O, disk access time is a major factor limiting the performance of secondary storage. Typically, one random access takes about 30 milliseconds, with about half of this delay accounted for by seek time and the other half due to latency and channel contention. We will show that shadowing can dramatically reduce seek time, thus decrease disk access time for individual I/O requests.

4. Expected seek time of shadowed disks

The expected seek distance of a magnetic disk device is defined as the average number of tracks traversed when the actuator moves the magnetic read/write head from a random track to any other random track. This definition assumes a uniform distribution of accesses. That is, from the current track, any other track is equally likely to be accessed next. In reality, track requests may be non-uniform, depending on the way data is laid out on the disk and on the relative frequency of access to different files [STH83]. However, the assumption of uniform accesses provides a good approximation of seek time, and disk scheduling is often aimed at minimizing the expected seek time computed under this assumption [TP72].

For shadowed disks, one must differentiate between seek time for read operations and seek time for write operations, since the seek distance required in these two cases is different. For a shadow set with k disks, the distances from the current track to the requested track can be seen as k random variables $X_1, X_2, \cdots, X_k$ with identical distributions. Then the seek distance for a read from the shadow set is the random variable X_R defined as

$$X_R = \min (X_1, X_2, \cdots, X_k)$$

and the seek distance for a write is the random variable X_W defined as

$$X_W = \max (X_1, X_2, \cdots, X_k)$$

In order to obtain an approximate distribution for X_R and X_W, we will assume that the X_i are *independent*. In reality, since a write operation may drive all the disk arms to the same position, there is a certain degree of correlation between these variables. However if the load is not very low and reads are frequent enough, it is reasonable to assume that most writes are done independently on each disk and reads undo the effect of concurrent writes. Under these assumptions, we can model the seek distances on the different disks in a shadow set as independent random variables.

Let us recall what the distribution of seek distances on one (non-shadowed) disk is. Let n be the number of tracks in the data band. There are n^2 unique seeks: n seeks of length zero (one starting at each of the n tracks) and $2(n-i)$ different seeks of length i, for $i = 1,2,...,n-1$. Thus each of the X_j variables has a distribution defined by

$$P(X = i) = 2(n - i) / n^2$$

or

$$P(X \geq i) = (2/n^2) \sum_{j=i}^{n-1} (n - j)$$

$$= (n - i) (n - i + 1) / n^2$$

for i = 1, 2, ..., n - 1.

4.1. Expected seek distance for reads

To derive the expected value of X_R, we observe that

$$P[min (X_1, X_2, \cdots, X_k) \geq i] = P(X_1 \geq i) \cdots P(X_k \geq i)$$

Thus

$$E[X_R] = \sum_{i=1}^{n-1} P[min (X_1, X_2, \cdots, X_k) \geq i]$$

$$= \sum_{i=1}^{n-1} [(n - i) (n - i + 1) / n^2]^k$$

$$= (1/n^{2k}) \sum_{i=1}^{n-1} (n - i)^k (n - i + 1)^k$$

For large n, this expression is well approximated by

$$(1/n^{2k}) \sum_{i=1}^{n-1} (n - i)^{2k} = n \sum_{i=1}^{n-1} (1 - i/n)^{2k} \, 1/n$$

The sum of the right-hand side is the Riemann sum for the integral

$$\int_0^1 (1-x)^{2k} = 1/(2k+1)$$

Thus we conclude that the expected seek distance for reading from a shadowed set with k disks is approximately

$$E[X_R] \approx n / (2k+1)$$

For $k = 1$, this reduces to the known expected seek of $n/3$ tracks [TP72], and for mirrored disks, $k = 2$, we observe a substantial decrease to $n/5$ tracks. Thus disk mirroring decreases the average seek time for random reads by a factor of 1.8.

4.2. Expected seek distance for writes

To derive the expected seek distance for writes, we observe that

$$P[max(X_1, X_2, \cdots, X_k) \le i] =$$
$$P(X_1 \le i) \cdots P(X_k \le i)$$
$$= [2/n^2 \sum_{j=1}^{i} (n-j)]^k$$
$$= 2^k/n^{2k} [i(2n-i-1)/2]^k$$
$$= 1/n^{2k}\, i^k (2n-i-1)^k$$

Thus

$$E[X_W] = \sum_{i=1}^{n-1} (1 - 1/n^{2k}\, i^k (2n-i-1)^k)$$
$$= (n-1) - n\sum_{i=1}^{n-1} (i/n)^k [2-(i+1)/n]^k (1/n)$$

For large n, the sum on the right-hand side is approximately equal to the Riemann sum for the integral

$$I_k = \int_0^1 x^k (2-x)^k\, dx$$

It can be shown (see Appendix) that the I_k satisfy the recurrence formula

$$I_k = \frac{2k}{(2k} + 1)\ I_{k-1}$$

thus

$$I_k = \frac{2k}{(2k+1)} \frac{(2k-2)}{(2k-1)} \cdots \frac{2}{3}$$

and

$$E[X_w] \approx n(1-I_k)$$

Again, for $k = 1$ we obtain the known seek distance $n/3$. For mirrored disks, $k = 2$, the expected seek distance becomes much higher: $0.46\,n$, that is nearly half of the disk data band. However, as the number of disks in the shadow set is increased beyond 2, we observe that the expected seek distance for writes does not degrade as badly. In Figure 3, the upper curve representing $E[X_W]$ flattens as the number of disks increases. For $k = 10$, the expected seek distance is 0.73 of the disk data band.

4.3. Expected seek distance for combined reads and writes

If we assume that a proportion α, $0 \le \alpha \le 1$, of all I/O requests to the shadow set are read requests, then the expected seek distance will be

$$X = \alpha X_R + (1-\alpha) X_W$$

Since reads from a shadow set are serviced faster but writes may take longer than on a single disk, the higher the proportion of reads, the better the shadow set will perform. In a transaction processing system, it will usually be the case that most random accesses are for read requests. Writes to the transaction log are performed on a separate disk, and they are sequential. In Figure 3, we have plotted the expected seek time in shadow sets containing 1 to 10 disks, with proportions of reads varying from 1.0 to 0.5. The lower curves, corresponding to proportions of reads equal to 0.6 or higher, remain under the 0.33 value, which corresponds to the expected seek distance for a single disk. These curves also show that the expected seek distance decreases as the number of disks in the shadow set increases. For an equal proportion of reads and writes, the expected seek distance $E[X_{.5}]$ remains approximately equal to 0.3 of the data band, independently of the number of disks in the shadow set.

4.4. Expected seek time

4.4.1. Constant speed actuator

The *nominal access time* [STH83] is defined as

$$E[T] = a + b\, E[X]$$

where *E[X]* is the expected seek distance computed under the assumption of uniform accesses, a is the mechanical settling time, and b is a constant determined by the speed of the actuator and the track density on the magnetic media. The expected seek time is equal to the nominal access time if the speed of the disk actuator is constant (since the expected value of a random variable $a+bX$ is $a+bE[X]$). In this case the time to seek a distance of i tracks is given by

$$T(i) = a + b\, i$$

With current technology, typical values for these constants

are $a = 5$ *milliseconds* and $b = .5$ *milliseconds*. The nominal access time corresponding to these values for a disk with 100 cylinders is 23 milliseconds. With the same access time function, the nominal access time for the same disk mirrored will be equal to

$$E[T_R] = 15 \text{ milliseconds} \quad \text{for } reads, \text{ and}$$

$$E[T_W] = 28 \text{ milliseconds} \quad \text{for } writes$$

Because the seek time is a linear function of the seek distance, the graphs in Figure 3 also indicate the behavior of the expected seek time as a function of the number of disks in a shadow set.

4.4.2. Voice coil actuator

The linear model is often used to estimate the expected seek time. However, in current disk technology, actuators have non-constant speed [STH83]. In particular, for voice coil actuators, the seek time is given by a non-linear function:

$$T(i) = a + b\sqrt{i}$$

For this case, we have not been able to derive the expected seek time $E[T]$ as a function of the expected seek distance. We were able to derive $E[T_R]$ and $E[T_W]$ directly, using a method similar to the computation of the expected seek distance (Sections 4.1. and 4.2.), but only in the case of mirrored disks. A brief summary of this derivation follows.

Recall from Section 4.1 that the probability of seeking i tracks is

$$P(X = i) = 2(n - i)/n^2$$

For 2 disks, the seek distance for reads X_R is distributed as

$$P[X_R=i] = P[min(X_1,X_2) = i]$$

$$= 2P(X = i) * \sum_{j=i}^{n-1} P(X = j)$$

$$= 4(n-i)^3 / n^3$$

Thus the expected seek time for reads in a mirrored disk is

$$E[T_R] = a + b \sum_{i=1}^{n-1} \sqrt{i}\, P(X_R = i)$$

Using the approximation

$$\sum_{i=1}^{n} i^j \approx n^{j+1} / j+1$$

we obtain

$$E[T_R] = a + .4\, b\sqrt{n} = a + b\sqrt{.16n}$$

In order to compute the expected seek time for writes, we will use the relationship between the expected values of the minimum and maximum of 2 identically distributed random variables

$$E[max(X_1,X_2)] + E[min(X_1,X_2)] = 2E[X_1]$$

The expected seek time for one disk (which was previously derived in [STH83]) is equal to

$$E[T] = a + b \sum_{i=1}^{n-1} \sqrt{i}\, 2(n-i) / n^2$$

$$= a + .53\, b\sqrt{n} = a + b\sqrt{.28n}$$

Thus the expected seek time for a write in a mirrored disk is

$$E[T_W] = 2E[T] - E[T_R]$$

$$= a + .66\, b\sqrt{n} = a + b\sqrt{.43n}$$

In Table 1, we summarize these results for the expected seek time in terms of the number of tracks it corresponds to, for constant speed ($T_i = a+bi$) and varying speed ($T_i = a+\sqrt{i}$) actuators.

Table 1
Proportion of Data Band Traversed
In Expected Seek Time
Constant Vs Varying Speed Actuator

Disk Read/Write	Constant Speed	Varying Speed
1 disk read/write	0.33	0.28
mirrored disk read	0.20	0.16
mirrored disk write	0.46	0.43

Note that with varying speed actuators mirroring decreases even further the expected seek time for reads. Compared to .28 of the data band for a single disk, a mirrored disk will seek only .16 of the data band.

5. Conclusions

In addition to providing high data availability and fault-tolerance, disk shadowing can boost the performance of mass-storage systems. A shadow set increases the number of I/O requests that can be handled per second, and reduces random access time for individual read requests. We developed a model to estimate the expected seek time in a shadow set as a function of the number of disks in the set.

In particular, we showed that in a mirrored disk with n cylinders in each drive, the expected seek distance for a random read is $n/5$, as compared to $n/3$ for a single drive. This result partially explains the performance improvement that has been observed in mirrored disks [BT85,

Sit86]. Our results indicate that shadow sets with a larger number of disks will provide significantly lower access times for random reads, in addition to increasing the I/O service rate. Further investigation is needed to quantify the impact of other parameters on the performance of shadow sets with a larger number of disks: rotational latency, buffer capacity, size of I/O requests, number of actuators, and disk scheduling algorithms.

Acknowledgements

We thank Jeffrey Millman for producing the graphs and providing insightful comments on an early draft on this paper. We are also grateful to Betty Salzberg and Garth Gibson for carefully reading the paper and pointing out a number of interesting problems.

References

[BT85] Bates K.H. and TeGrotenhuis M., "Shadowing Boosts System Reliability," *Computer Design*, April 1985.

[Ki85] Kim M.Y., "Synchronized Disk Interleaving," *IEEE Transactions on Computers*, November 1986.

[MGB74] Mood A.M., Graybill F.A., and Boes D.C., *Introduction to the Theory of Statistics*, Mc Graw Hill, 1974.

[PGK87] Patterson D. A., Gibson G., and Katz R.H., "A Case for Redundant Arrays of Inexpensive Disks (RAID)," *Proceedings ACM Sigmod*, Chicago, June 1988.

[SG86] Salem K. and Garcia-Molina H., "Disk Striping," *Proceedings 1986 Data Engineering Conf*, Los Angeles, February 1986.

[STH83] Scranton R.A., Thompson D.A. , and Hunter D.W., "The Access Time Myth," *IBM Tech. Report*, RC 10197, September 1983.

[TP72] Teorey T.J. and Pinkerton T.B., "A Comparative Analysis of Disk Scheduling Policies," *Communications of ACM*, 15:3, March 1972.

[Sit86] Sitler T. et al., "Configuring Disks," Tandem Systems Review, December 1986.

Appendix

Let T be a random variable representing the time between failures of a disk. If the expected time between failures is five years, and the distribution of T is exponential, then the probability of a disk failing in a time window of 3 hours is

$$P(T \le 3) = 1 - e^{-\frac{3}{365 \times 24 \times 5}}$$

$$= 6 \times 10^{-5}$$

Because of the memoryless property of the exponential distribution, this is also the probability of a second disk failing within 3 hours after a first disk has failed. However, note that in a shadow set of k disks, each with an expected failure time *MTBF*, the expected time until one disk in the set fails is k times shorter than *MTBF*. In particular, this means that one of the two disks in a mirror is expected to fail twice sooner than a single disk.

The integral in Section 4.2. :

$$I_k = \int_0^1 x^k (2-x)^k \, dx$$

Substituting $u = 1-x$ and $\sin v = u$, we get

$$I_k = \int_0^1 (1-u)^k (1+u)^k \, du$$

$$= \int_0^{\frac{\pi}{2}} (1 - \sin^2 v)^k \cos v \, dv$$

$$= \int_0^{\frac{\pi}{2}} \cos^{2k+1} v \, dv$$

Integration by parts gives the recurrence

$$\int_0^{\frac{\pi}{2}} \cos^{2k+1} v \, dv = \frac{2k}{(2k+1)} \int_0^{\frac{\pi}{2}} \cos^{2k-1} v \, dv$$

and since

$$I_1 = 2/3$$

we obtain

$$I_k = \frac{2k}{(2k+1)} \frac{(2k-2)}{(2k-1)} \cdots \frac{2}{3}$$

FIGURE 1

A Mirrored Disk - Mirroring Supported by Host

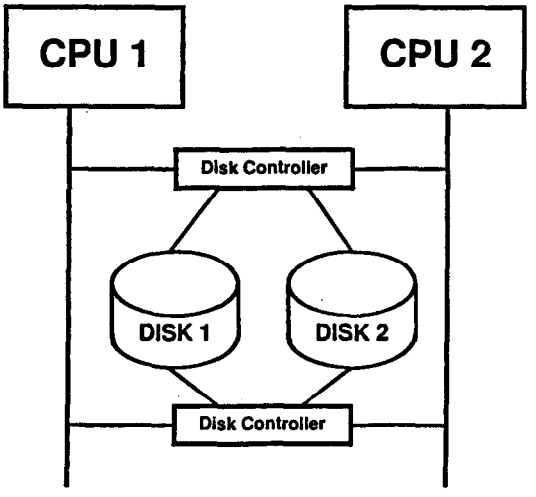

FIGURE 2

A Shadow Set - Shadowing Supported by Server

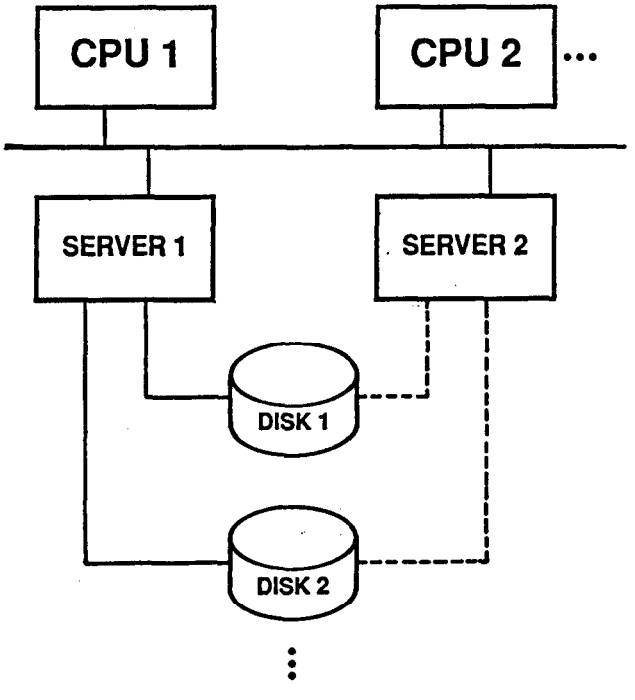

FIGURE 3

EXPECTED SEEK DISTANCE AS PROPORTION OF DATA BAND

Number of disks in shadow set: 1 to 10

Proportion of reads vs writes: 1.0 to 0.5