研 究 速 報

Storage Management Scheme of Disk Arrays Using Hot Mirroring

Hot Mirroring を用いたディスクアレイの記憶管理方式

Kazuhiko MOGI* and Masaru KITSUREGAWA**

茂 木 和 彦・喜連川 優

1. Introduction

RAID[1] utilizes a large number of commodity inexpensive drives in parallel to achieve higher performance as well as obtaining higher reliability by recording redundant information. Patterson et. al classified RAID into five levels. Among those five levels, level 1 (mirrored disk arrays) and level 5(RAID5 disk arrays) are regarded as two of the most promising approaches for providing highly reliable secondary storage systems which support concurrent access to small blocks. But there are two big problems in using RAID5 disk arrays. One is the overhead of recording redundancy information. The other is the overhead of reconstructing data after disk failure. Mirrored disk arrays pay the penalty of much smaller data capacity than that of RAID5 disk arrays, because of the data copying for redundancy. In order to get not only higher performance but also larger usable disk capacity, we consider the combination of a mirrored disk array and a RAID5 disk array with hot block separation. This storage management scheme is named "hot mirroring".

2. Hot mirroring

2.1 Concept of hot mirroring

With respect of high storage efficiency with high reliability, RAID5 is best among all the RAID levels. From the point of view of performance, mirrored disk arrays are better than RAID5 disk arrays. If it were possible to merge the characteristics of high storage efficiency from RAID5 disk arrays with the low overhead of recording redundancy information for mirrored disk arrays, it might be one of the best disk array configurations. In general, there are access localities which can be utilized for improving the performance of RAID5 disk arrays[2]. To solve the problems of mirrored disk arrays and RAID5 disk arrays, access localities are exploited. According to the access frequency, two groups of disjoined blocks are made, one group contains blocks with high access rates (hot blocks) and the other has low access rate blocks (cold blocks). With this separation, the mirror scheme and the parity encoding scheme are combined to get higher performance and larger capacity.

For load balancing, it is important to distribute the access requests for hot blocks evenly over all disks. It is also desirable that the penalty of recording redundant information on hot blocks be small. These requirements are well suited to the mirror scheme, thus hot blocks are mirrored. For the cold area, we use parity protection for redundancy to obtain higher storage efficiency. This storage manage-

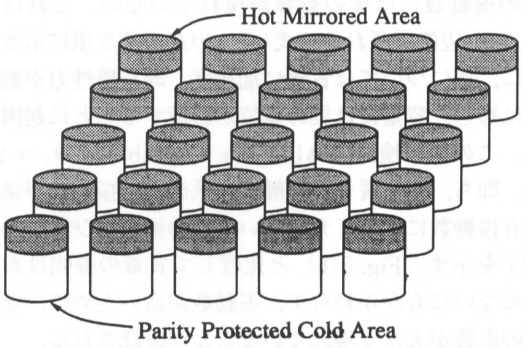

Fig. 1 Hot mirroring

*Department of Electrical Engineering and Electromics, Institute of Industrial Science, University of Tokyo

**Center for Conceptual Processing of Multi-media Information

ment scheme is named "hot mirroring" (Fig. 1).

Identification of hot blocks is the key to this proposed method. Usually almost all blocks written by write requests can be regarded as hot, since these blocks are likely to be used again because of locality. So we assume that all blocks of normal write requests are hot and perform all write accesses to the hot area. With small probability, a cold block is written into the hot area, thus consuming hot area free space. Cold blocks in the hot area need to be migrated back to the cold area. By recording the time at which each block in the hot area was last accessed, the cold blocks residing in the hot area can be found by finding the blocks with the oldest access time. If the amount of free space in the hot area falls below a threshold value, this migration is invoked. Cold block migration from hot to cold needs two extra write accesses and a read access for cold block write operations. As will be clarified in section 3, this is not a high overhead.

2.2 Data placement policy

The data placement policy of the mirrored hot area and the parity protected cold area considerably impacts performance on a disk failure because the effect of the rebuild process on the parity protected area is very large. For hot mirroring, the copy allocation of the hot area is illustrated in Fig. 2(a) and parity stripes for the cold area are as shown in Fig. 2(b). In the cold area, parity stripes are made into a disk group (vertically in the figure). In the hot area, the copy is allocated on a different disk group (horizontally in the figure), which is based on the chained declustering method[3].

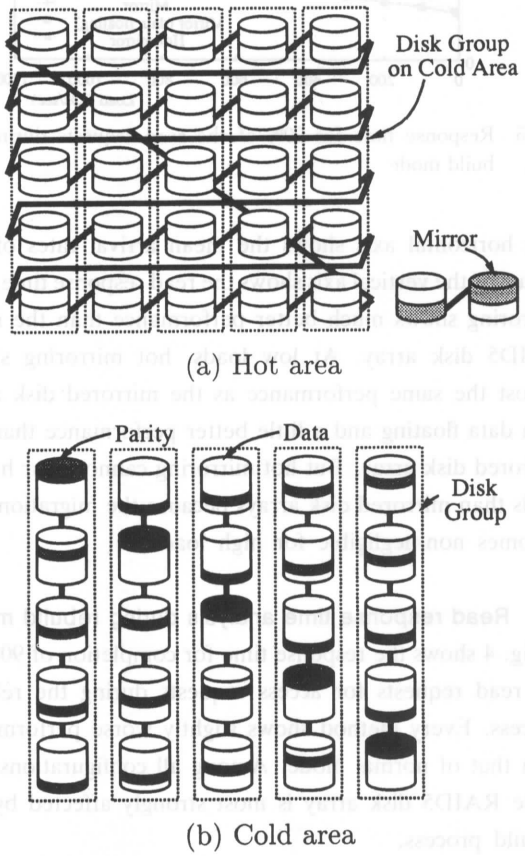

Fig. 2 Data allocation policy

The reason why we employ such orthogonal placement for parity stripe and mirroring is as follows. Since the rebuild time needs to be minimized, the disks of the parity stripe containing the broken disk should work towards rebuilding as much as possible. This means that frequent accesses against hot areas of the broken disk group should be redirected to the mirrored hot area of the other disk group. Thus parity stripe and copy allocation are orthogonal to each other. By employing this scheme, hot accesses can be served by surviving stripes, while all the drives on the broken stripe work on rebuilding.

3. Evaluations of hot mirroring

The feasibility of hot mirroring was examined through simulation.

3.1 Simulation assumptions

Simulation parameters are as follows. Table 1 shows the disk model parameters. The block size is 4KB. The striping unit is set to the block size. Access requests are fixed at 4KB. The interval of access request arrivals have a negative exponential distribution. The load is controlled by changing the mean time between access requests. 70% of the total requests are read operations and the others are write operations. For access locality, we assume that 90% of the requests are concentrated to 10% of the valid data blocks.

Table 1 Disk model parameters

capacity	318 MB
cylinders/disk	949
tracks/cylinder	14
sectors/track	6
sector size	4096 bytes
revolution time	13.9 ms
seek time model	$2.0 + 0.01 \cdot d + 0.46 \cdot \sqrt{d}$
track skew	1 sector

Table 2 Data capacity for each configuration

	Data capacity
RAID5 (4*(5D+P))	83.3 %
Mirror (naive)	50.0 %
Mirror (data floating)	49.9 %
Hot Mirroring (4*(5D+P), 20% hot area)	66.7 %

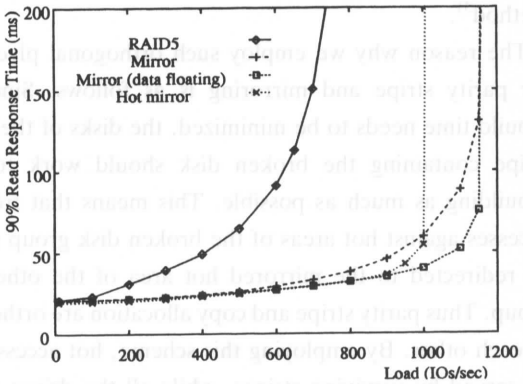

Fig. 3 Response time for 90% of the read requests in normal mode

To simplify the simulation, it is assumed that the overhead of the controller and the communication overhead between the controller and disks is negligible. All the control tables are maintained by the controller. All disk accesses are performed on first come first serve basis.

To compare performance, four configurations are examined, hot mirroring, naive RAID5, mirroring with fixed data position, and mirroring which adopts data floating and uses the same method to balance the load as hot mirroring on write operations. Naive RAID5 is the same management scheme as used by the cold area management on hot mirroring, and mirroring and mirroring with data floating disk arrays uses the same management scheme of hot mirrored area as hot mirroring does. Table 2 shows the effective data capacity of these configurations under simulation. Statistics gathering begins after an initial two millions write accesses to the hot mirroring portion and after an initial hundred thousands write accesses to the other disk arrays.

3.2 Read response time analysis in normal mode

Fig. 3 shows the response time in which 90% of the read requests have been completed for 100,000 access requests.

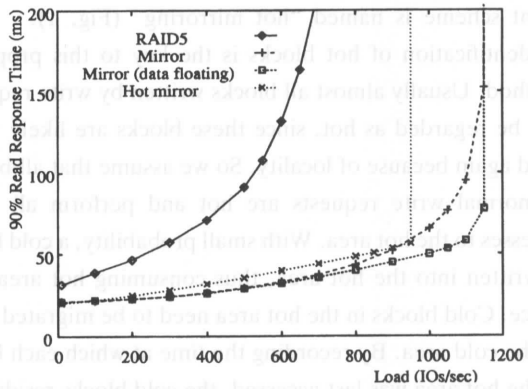

Fig. 4 Response time for 90% of the read requests during rebuild mode

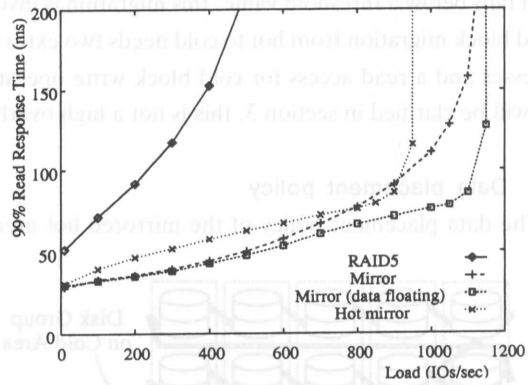

Fig. 5 Response time for 99% of the read requests during rebuild mode

The horizontal axis shows the mean arrival rates of I/O requests, the vertical axis shows the read response time. Hot mirroring shows much better performance than the naive RAID5 disk array. At low loads, hot mirroring shows almost the same performance as the mirrored disk array with data floating and a little better performance than the mirrored disk array. But hot mirroring cannot bear higher loads than mirrored disk arrays because the migration cost becomes non-negligible for high loads.

3.3 Read response time analysis during rebuild mode

Fig. 4 shows the response time for completion of 90% of the read requests for access requests during the rebuild process. Every method shows slightly worse performance than that of normal mode. Among all configurations, the naive RAID5 disk array is most strongly affected by the rebuild process.

In order to clarify the impact on the access requests which

are highly affected by the rebuild process, the response time for completion of 99% of the read requests on the same data used in Fig. 4 is also examined. Fig. 5 shows the result. The naive RAID5 disk array is strongly affected by the rebuild process. The other methods shows slightly worse performance than that of the 90% read requests case. Hot mirroring shows worse performance than that of mirrored disk arrays. In mirrored disk arrays, all data is copied. In hot mirroring, upon a request, reconstruction of the broken data in the cold area is required. This difference causes the response degradation for hot mirroring. But the performance of hot mirroring is significantly better than that of the naive RAID5 disk array since most of the read requests against the broken disk can be covered by the paired hot mirrored drive.

4. Conclusion

This paper presents the new storage management scheme named "hot mirroring" for obtaining higher performance and larger data capacity. This scheme makes use of access localities. Each disk is divided into two regions, the hot area and the cold area. In order to reduce the overhead of recording redundant information and to balance the load among all disks, all blocks in the hot area are mirrored. In the cold area, a parity encoding scheme is adopted for redundancy with low storage overhead. Hot mirroring makes the assumption that all written blocks are hot. Cold blocks in the hot area are estimated by examining the elapsed time since the last access occurred. They are migrated to the cold area according to the number of free blocks in the hot area.

The feasibility of hot mirroring was examined through simulation. Hot mirrored disk arrays show much higher performance than that of naive RAID5 disk arrays. At low loads, hot mirrored disk arrays have slightly better performance than mirrored disk arrays and almost the same performance as that of mirrored disk arrays which adopt data floating to balance the load for write operations. But hot mirroring cannot provide higher performance than mirrored disk arrays because of the overhead of separating the hot blocks. During rebuild mode, hot mirroring shows slightly worse performance than do mirrored disk arrays, but has much better performance than naive RAID5 disk arrays.

(Manuscript received, July 7, 1995)

References

1) D. A. Patterson, G. Gibson, and R. H. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID). In Proc. of ACM SIGMOD, pp. 109–116, Jun. 1988.
2) K. Mogi and M. Kitsuregawa. Hot Block Clustering for Disk Arrays with Dynamic Striping --- exploitation of access locality and its performance analysis. In Proc. of VLDB, Sep. 1995.
3) H. Hsiao and D. DeWitt. Chained Decrustering: A New Availability Strategy for Multiprocessor Database Machines. In Proc. of ICDE, pp. 456–465, Feb. 1990.