RAID 레벨 5 구조의 혼합형 데이타 복구 기법
(Hybrid Data Recovery Technique of RAID Level 5 Architecture)

전상훈 [†] 안병철 [††]
(Sang-Hoon Jeon) (Byung-Chul Ahn)

요 약 실시간 멀티미디어 응용 서비스를 위해 저장 시스템은 성능 저하 없이 데이타를 제공하여야 한다. 단일디스크 고장시 빠른 데이타 복구는 상당히 중요하므로 디스크 복구 장치가 필요하다. RAID 레벨 5 에서 단일디스크 고장시 새 디스크가 교체되기 전까지 유발되는 급격한 성능저하를 방지할 수 있는 혼합형 복구 기법을 제안한다. 이 기법은 기존 연구에서 제시된 예비 디스크 기법에 비해 추가 디스크를 사용하지 않으므로 경제적이다. 제안한 기법의 성능은 여러 가지 요구크기에 대해 RAID 레벨 5 성능과 시뮬레이션을 통하여 비교한다. 시뮬레이션 결과 제안한 기법은 실패모드에서는 20%, 재구성모드에서는 80%이상의 성능 개선을 보인다. 그리고, RAID 컨트롤러와 하드디스크를 사용해 멀티미디어 서버 시스템을 구축하여 실측과 시뮬레이션 결과를 비교한다.

Abstract For real time services of multimedia applications, storage systems should provide data without degrading their performance. Since it is very important to recover data immediately at a disk failure, the disk recovery system is required. This paper presents a hybrid recovery scheme which prevents degraded the performance on a single disk failure at RAID level 5 architecture until the failed disk is replaced with a new disk. The proposed scheme is very economical compared to previous spare disk schemes because it does not use extra disks. The performance of the proposed scheme is evaluated and analyzed with that of the RAID level 5 for various requested sizes through the simulation. The results show that the performance of the proposed scheme is improved up to 20 percents at the failure mode and 80 percents at the reconfigured mode. After a multimedia server system has been built with a RAID controller and hard disks, the data recovery performance of the propose scheme are compared with the results of simulation.

1. Introduction

The high speed networks enable to transfer multimedia data such as audio and video. For multimedia applications such as VOD systems and multimedia database systems, disk arrays are used to store and retrieve multimedia data.

Redundant disk arrays have fault tolerant characteristics, incorporating a layer of error handling which does not found in non-redundant disk systems [1]. Recovery from errors is complex because the disk array may reach to the large number of erroneous states[2]. RAID level 5 architecture provides reliability using the data protection scheme based on parity and it improves performance using the block interleaving scheme by smaller additional costs[3]. The primary weakness of RAID level 4 is to over-utilize by writing the parity disk. RAID level 5 overcomes this problem by distributing parity blocks across all of the member disks. Thus all member disks contain data blocks and parity blocks. RAID-5 spreads parity blocks for each stripe unit in successive different locations. Both data blocks and parity blocks are evenly distributed throughout the array. A variety of strategies exist to evenly

[†] 정 회 원 : 경동정보대학 인터넷정보계열 교수
shjeon@kyungdong-c.ac.kr
[††] 종신회원 : 영남대학교 컴퓨터공학과 교수
ahn@yeungnam.ac.kr

논문접수 : 1999년 8월 17일
심사완료 : 2000년 4월 18일

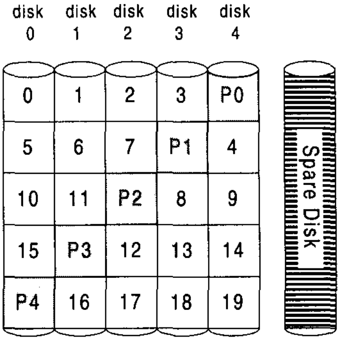
distribute data blocks and parity blocks[4]. The more the number of disks on a system increases, the more the probability of fault increases. RAID is a set of disks with redundancy to protect against data loss. Therefore, data should be recovered from a single disk drive crash in disk array systems. But if rapid restoration can not be supported, severe degradation of performance could be resulted from doubling the access rate to survived disks until the crashed disk is replaced with a new disk. It is important that single disk failures are expected to be relatively frequent in RAID systems[5]. RAID level 1, disk mirroring, is a traditional approach to improve reliability, but calls for using more than 50% of storage capacity[1]. This is the most expensive option since all disks are duplicated[1].

Since multimedia applications do not require a complete data backup, RAID-5 is used to write and retrieve multimedia data by real time. For RAID-5, it is necessary to reconstruct data from a single disk failure in a RAID system. Several spare schemes are proposed to reconstruct data from the failed disk. But they use additional redundant disks and show low performance. An efficient scheme is proposed to reconstruct data without using any redundant disks.

In Section 2, the previous related researches and strategies are discussed to rebuild a failed disk. In Section 3, the hybrid data recovery scheme is proposed. Section 4 and 5 present simulated system and experiment results respectively. Section 6 provides some concluding remarks.

2. Related Researches

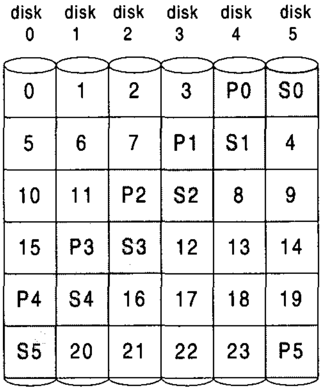
Hot stand-by disks are used to recover a failed disk immediately by adding usable area in disk arrays. They automatically rebuild data of the failed disk on the stand-by disk from the redundant information on the survived disks. One of these simple schemes is hot sparing scheme, which is locked on state of not being used during normal operation until failed disk appeared[1]. In a system with n disks, only n-1 disks are utilized during normal operation. Fig. 1 shows the hot sparing scheme, where each column corresponds to a disk and each row corresponds to the data layout for a track on the disks.



disk 0	disk 1	disk 2	disk 3	disk 4
0	1	2	3	P0
5	6	7	P1	4
10	11	P2	8	9
15	P3	12	13	14
P4	16	17	18	19

Fig. 1 Hot sparing scheme

Distributed sparing scheme uses the spare space on the disks as a part of workload processing. The spare space is distributed on all disks in the array with stored data and parity instead of locating it on a separate disk. Distributed spare space on the disks permits recovering data from a disk failure with no interruption of data availability. Fig. 2 shows spare blocks of the distributed sparing scheme. Compared to the hot sparing scheme, this scheme uses all the disks in the array during normal operation and raises the response time. Thomasian has analyzed the performance of RAID-5 with distributed sparing inthe normal mode, the degraded mode, and the rebuild mode in and OLTP environment, which implies small



disk 0	disk 1	disk 2	disk 3	disk 4	disk 5
0	1	2	3	P0	S0
5	6	7	P1	S1	4
10	11	P2	S2	8	9
15	P3	S3	12	13	14
P4	S4	16	17	18	19
S5	20	21	22	23	P5

Fig. 2 Distributed sparing scheme

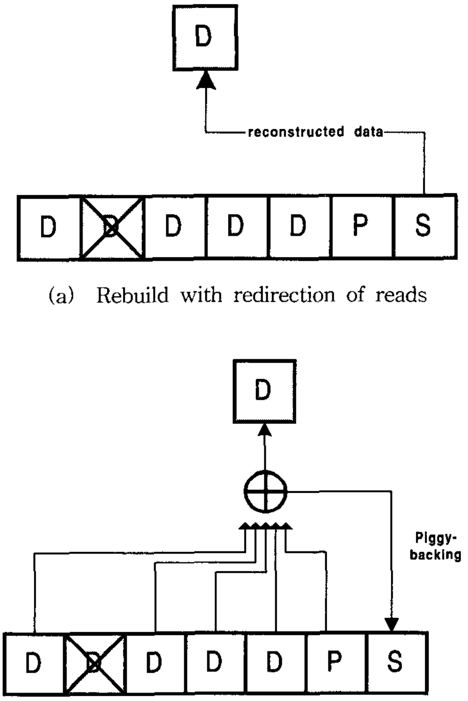
reads and writes[6].

Parity sparing scheme uses the spare space on the disks as a part of secondary parity disk and it can reduce the parity group length, which means the number of disks in a parity group. When one disk failure is detected, the parity blocks of two groups are merged to get a single larger parity. This scheme can reduce the parity group length by making effective use of the spare space during normal operation. Small size of parity group length is more efficient to construct parity blocks during normal operation and shows better performance in transaction processing applications. Compared to the RAID-5, all survived disks are used to reconstruct the data on a failed disk, if one of the disks fails. Therefore, these survived disks increase a load of 100% during a failure mode. Fig. 3 shows block locations of the parity sparing scheme.

disk 0	disk 1	disk 2	disk 3	disk 4	disk 5
0	1	2	3	PA	PB
5	6	7	PA	PB	4
10	11	PA	PB	8	9
15	PA	PB	12	13	14
PA	PB	16	17	18	19
PB	20	21	22	23	PA

Fig. 3 Parity sparing scheme

Several disk rebuild strategies for disk arrays are discussed by Muntz and Lui in 1990[5]. They propose three disk rebuild strategies termed as baseline copy procedure, rebuild with redirection of reads and piggy-backing rebuild. The baseline copy procedure simply sequentially reads blocks from the survived disks by reconstructing and writes them to the standby disk. The rebuild with redirection of reads is supported by reading the data from the spare disk rather than reconstructing the data from the survived disks again. The piggy-backing rebuild is reconstructed due to a read request that was issued as part of the normal workload. Fig. 4 shows the rebuild with redirection of reads and piggy-backing rebuild strategies.



(a) Rebuild with redirection of reads

(b) Piggy-backing rebuild on normal workloads

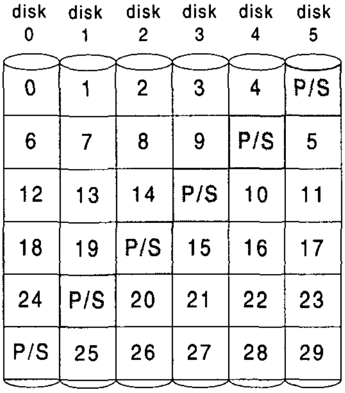
Fig. 4 Disk rebuild strategies

These strategies are used to reduce the load of survived disks in an disk array with a failed disk. Since a single disk failure is expected to be relatively frequent in disk arrays, these sparing schemes are required to have additional redundant disks. Therefore expensive cost should be supported to disk array systems. This is a great drawback in inexpensive disks array systems. A cost-effective architecture is proposed to increase the performance significantly for a single disk failure.

3. Hybrid Recovery Scheme

Sparing disk schemes are very effective on single disk failures but these schemes require additional disks to maintain each array size. A cost effective architecture without degrading performance is proposed when a single disk failure occurs at the RAID level 5 architecture. The file processing time of the hybrid recovery scheme is evaluated on various transaction file sizes with redirection of reads and piggy-backing rebuild strategies.

When a disk is failed, the hybrid recovery scheme uses the parity blocks as recovery blocks instead of using a spare disk illustrated in Fig. 5. It reconfigures data faster than sparing disk schemes do. The operation of the hybrid recovery scheme is described in five spare operation modes categorized by Menon and Matterson[7].



disk 0	disk 1	disk 2	disk 3	disk 4	disk 5
0	1	2	3	4	P/S
6	7	8	9	P/S	5
12	13	14	P/S	10	11
18	19	P/S	15	16	17
24	P/S	20	21	22	23
P/S	25	26	27	28	29

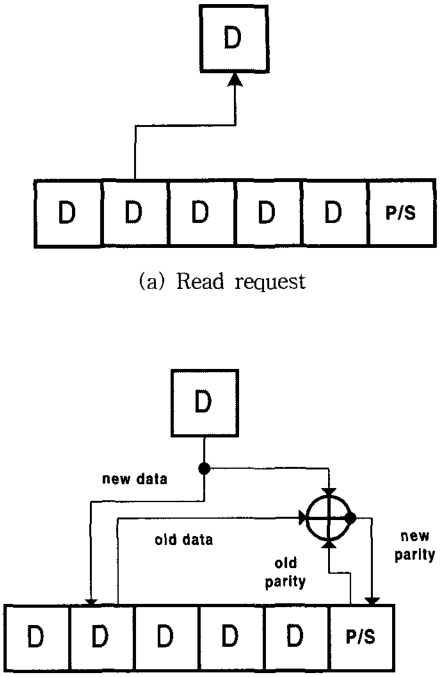
Fig. 5 Hybrid recovery scheme

3.1 Normal Mode

During the normal mode operation, the hybrid recovery scheme operates the same as a RAID level 5. The RAID system works the parity-based protection, which is operated by exclusive-OR operations. Fig. 6 shows read and write request operations on one parity group on the normal mode. In Fig. 6, D means a data block and P/S means a hybrid block.

3.2 Failure Mode

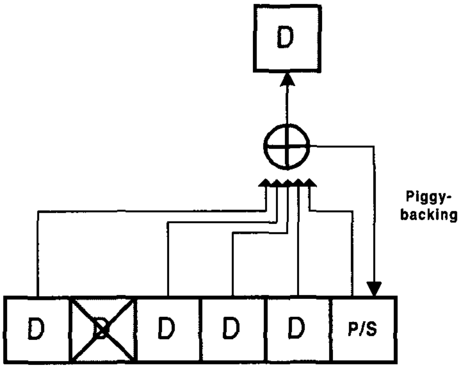
When a read request on failed blocks is supported, only one unit of access time for hybrid blocks is added to access operation for disk arrays using the piggy-backing rebuild. But this operation can help to reduce the load time of the survived disks in diskarrays during the reconstruction mode. When a write operation is requested on a failed block, only one unitof access time is needed for a hybrid block. In RAID level 5, the first unit of access time is needed on survived disks to obtain an old data and then the second unit of access time is needed on a parity block to write the new parity data. Fig. 7 illustrates various operations on the failure-mode.



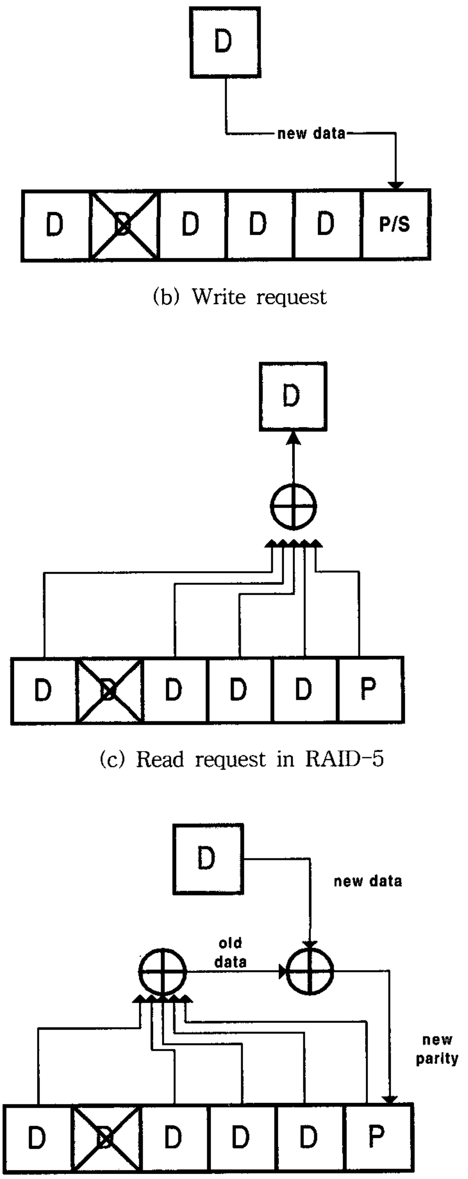
(a) Read request

(b) Write request

Fig. 6 Normal-mode operation



(a) Read request with piggy-backing



(b) Write request

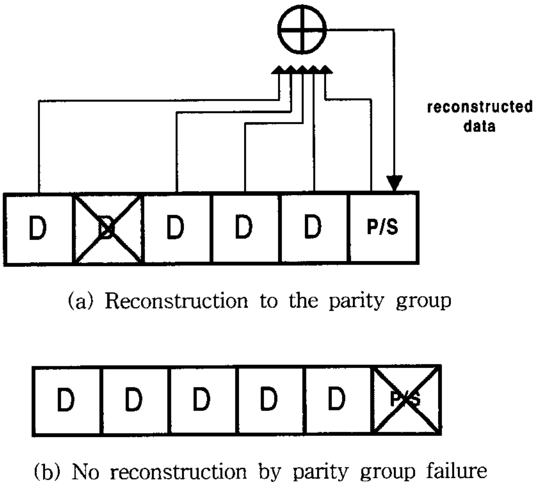
(c) Read request in RAID-5

(d) Write request in RAID-5

Fig. 7 Various operations on the failure-mode

3.3 Reconstruction Mode

For the hybrid recovery scheme, parity blocks in RAID level 5 is converted to spare space as distributed sparing scheme on a single disk failure. Fig. 8(a) shows reconstructing procedure on a data block in parity group when a single disk fails. Fig. 8(a) is similar to the hot sparing scheme and the distributed sparing scheme. And Fig. 8(b) shows reconstructing procedure on a parity block in parity group when a single disk fails. No operation is occurred, while the distributed sparing scheme reconstructs a parity block in a parity group on a spare space.



(a) Reconstruction to the parity group

(b) No reconstruction by parity group failure

Fig. 8 Reconstruction-mode operation

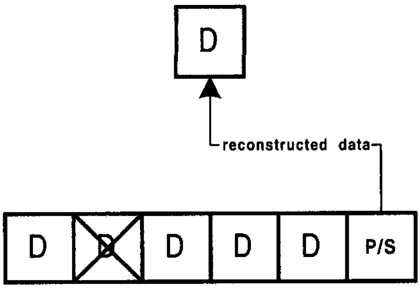


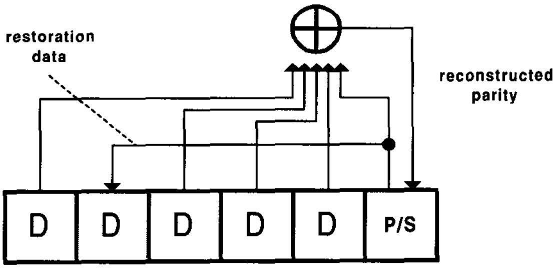
Fig. 9 Read request on the reconfigured-mode operation

3.4 Reconfigured Mode

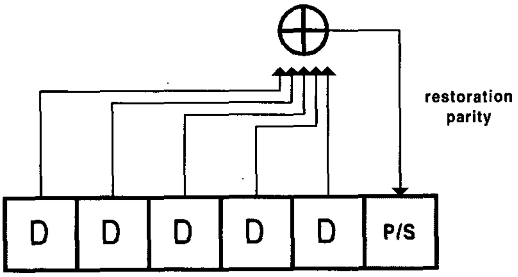
After the reconstruction process finishes reconstructing data on the failed disk, the reconfigured mode is executed when a new spare is brought into the system to replace the failed disk. In Fig. 7(c), RAID level 5 requires multiple reads to the survived disks in the same array each time. In the worst case, this can double the access time to the survived disks[5]. But, the hybrid recovery scheme has a good response time on a workload. Fig. 9 shows a read request on a failed block in the reconfigured mode operation.

3.5 Restoration Mode

After replacing a failed disk with a good disk to return the normal mode, data and parity in disks are reallocated to restore on a new disk. Fig. 10(a) shows the restoration process for a failed data block in a parity group. The first unit of access time is required on survived disks to reconstruct parity blocks. The second unit of access time is required to restore data on a new disk and write back to save reconstructed parity on hybrid blocks. Fig. 10(b) shows restoration process for a failed parity block in a parity group. The process is the same as restoration process in RAID-5.



(a) Restoring data and parity blocks



(b) Restoring to parity blocks

Fig. 10 Restoration-mode operation

4. Simulation and Performance Evaluation

4.1 Simulation

An analytical model to simulate the hybrid recovery scheme is implemented by the discrete event simulation library(SMPL) based on C language[8]. Table 1 shows parameters of the disk array simulation. Disk parameters are based on a Quantum Atlas XP34300S SCSI 4GB disk drive. As input/output data, large multimedia data are used. To improve the performance in a · disk array system means to reduce the response time or to increase the throughput[9]. Comparing with RAID level 5, the performance rate is measured on a single disk failure. In Equation (1), performance rate is obtained from speedup model.

$$\textit{Speedup} = \textit{Processing Time(old)} \,/\, \textit{Processing Time(new)} \quad (1)$$

Table 1 Disk Parameters

cylinders per disk	3,832
tracks per cylinder	20
sectors per track	110
bytes per sectors	512
disk capacity	4GB
revolution time	8.33ms
single cylinder seek time	1.0ms
average seek time	8.5ms
max stroke seek time	18ms
Rotational Speed	7,200rpm

Simulation for the disk array is constructed with six disks and parity blocks are allocated by the left-symmetric parity distribution method[9]. The left-symmetric placement is derived by left rotations of entire parity stripes from the RAID level 4 placement. The left-symmetric placement shows the best performance at RAID level 5. With the nonlinear model, seek times are calculated by Equation 2[4].

$$\textit{seekTime}(x) = \begin{cases} 0 & \text{if } x = 0 \\ a\sqrt{x-1} + b(x-1) + c & \text{if } x > 0 \end{cases} \quad (2)$$

When x is the seek distance in cylinders and a, b and c are chosen to satisfy the minimum seek time, average seek time and maximum stroke seek time constraints. If cylinders per disk are greater than approximately 200, a, b and c can be approximated using the following equations.

$$a = (-10MinSeek + 15AvgSeek - 5MaxSeek) / (3\sqrt{NumCyl}) \tag{3}$$

$$b = (7MinSeek - 15AvgSeek + 8MaxSeek) / (3NumCyl) \tag{4}$$

$$c = MinSeek = \textit{Single cylinder seek time} \tag{5}$$

For the disk parameters of Table 1, a = 0.1481, b = 0.1265 and c = 1 are used.

4.2 Performance Evaluation

The performance of the hybrid recovery scheme is analyzed by the analytical simulation model. It is assumed that a seek cost function is nonlinear and disks in a parity group are synchronized. Normal requests are assumed that read requests are 70% probability and write requests are 30% probability. Requests are assumed to arrive with an exponential distribution.

In the normal operation, response times are equal to RAID level 5 since the hybrid scheme uses the same size as parity group of RAID level 5. File processing times, during the failure operation, are shown in Fig. 11. Fig. 11 illustrates file processing times at the failure mode. Since the reliability of the disk array is quite dependent on the reconstruction time[10], the hybrid recovery scheme employs rebuilding strategies with redirection of reads and piggy-backing rebuild strategies. In the failure mode operation, the hybrid recovery scheme has less file processing time than RAID-5. The processing time on performance is more pronounced at higher loads.

RAID level 5 does not operate in the reconstruction mode because there are no spare space for reconstruction. In the hybrid recovery scheme, parity blocks in RAID-5 are converted to spare spaces as distributed sparing scheme on a single disk failure. After the reconstruction process finishes reconstructing data for the failed disk, the hybrid recovery scheme has better performance than the RAID-5 architecture shown in Fig. 12. Because of request redirection, normal requests to already reconstructed data on a hybrid block get serviced quicker than the RAID-5 architecture.

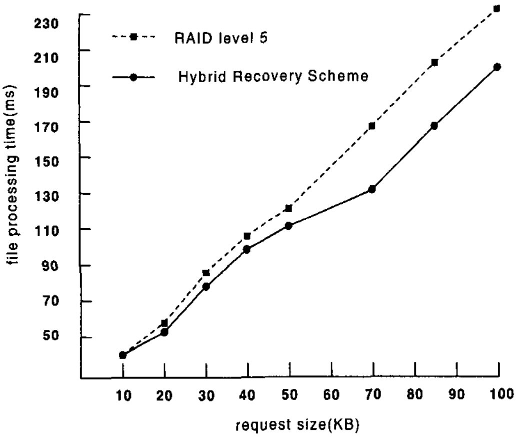


Fig. 11 File processing times during failure-mode

Larger transaction files in simulation are used in reconstructing data from the failed disk. At RAID level 5, it requires several access to reconstruct data from the survived disks. Hence, converted spare disk with hybrid blocks is very useful on a single disk failure. Since the reliability of the disk array is quite dependent on the reconstruction time, a single disk

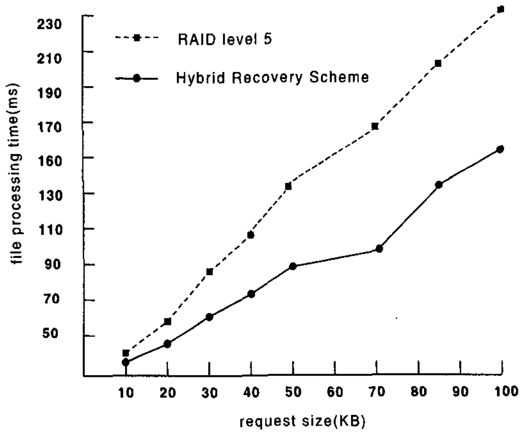


Fig. 12 File processing time during reconfigured-mode

failure incurs worse performance in the disk array system if the replacement time is long[11].

5. Experiment with a VOD System

A VOD system with a disk array is designed to evaluate the hybrid recovery scheme. A RAID controller implements a linear address space. The array is appeared to the host as a linear sequence of data units, numbered from 0 to $N \cdot B-1$, where N is the number of disks in the array and B is the number of units of user data on a disk. The RAID controller translates linear addresses into physical disk locations.

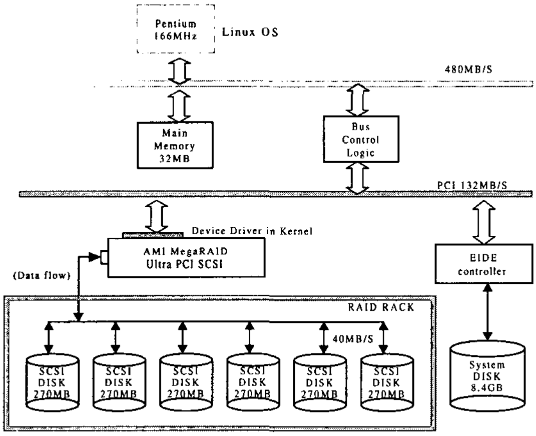
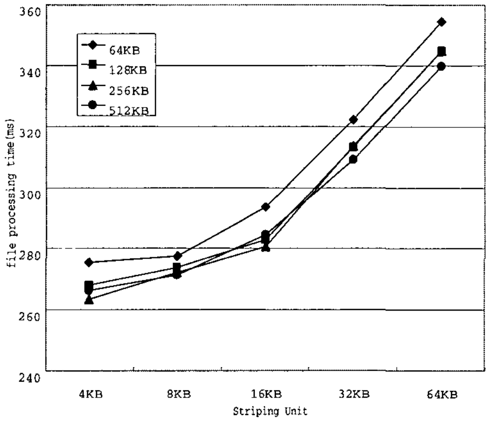


Fig. 13 VOD system for the hybrid recovery scheme

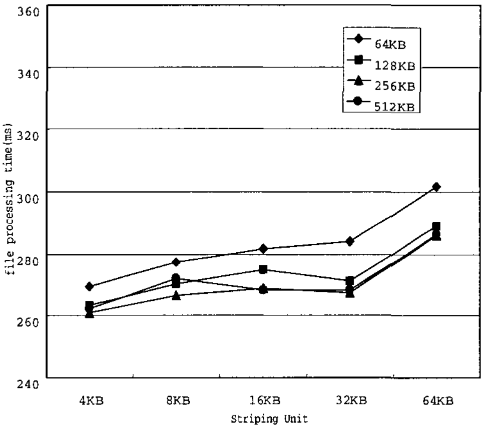
User data is striped to consecutive units across the disks of the array. The striping unit size can be theoretically as small as a single bit, byte, or as large as an entire disk.

Fig. 13 shows a block diagram of the VOD system implemented the hybrid recovery scheme using a RAID controller. PCI MegaRAID Ultra SCSI from American Megatrends is used for the disk array controller and LPS270S SCSI hard disks by QUANTUM. The RAID controller BIOS is configured for each experiment.

To remove the measurement errors, experiments are executed three times and averaged the measured results. Read requests between 64KB and 512KB are used. Fig. 14 shows file processing time during the failure mode when a 500MB MPEG file is read completely during reconfigured mode with various striping unit sizes. In Fig. 14(a) file processing time of RAID 5 is rapidly increased and degraded performance compared with that of hybrid recovery scheme in Fig. 14(b) as the request size and the striping unit size are increased.



(a) RAID level 5



(b) Hybrid recovery scheme

Fig. 14 File processing time

Most multimedia applications request larged striping unit for storage systems and buffer management policies[12]. To achieve load balance in concurrent workloads, the striping units size and the request size are required to increase as the concurrency is increased. For these reasons, it is very important to recover data immediately at a single disk failure for real time multimedia applications such as VOD systems, multimedia databases and etc.

6. Conclusions

The RAID-5 architecture is very effective, but on a failed disk, it shows very slow response because of reconstruction process every access.

The hybrid recovery scheme has spare spaces, which provide for reconstructing the failed data during the reconstruction process without additional redundant disks. It's performance is evaluated in various transaction file sizes with rebuild strategies. The results show that the performance of the hybrid recovery scheme shows up to 20 percents at failure mode and 80 percents at reconfigured-mode compared with RAID-5 architecture.

The VOD system implemented on the hybrid recovery system shows that the performance has improved as the request size and the striping unit size are increased.

References

[1] D. A. Patterson, G. A. Gibson, and R. H. Katz, "A Case for Redundant Arrays of Inexpensive Disks(RAID)," *International Conference on Management of Data(SIGMOD)*, pp. 109-116, June. 1988.

[2] W. V. Courtright II, G. A. Gibson, "Backward Error Recovery in Redundant Disk Arrays," *Technical Report REF42170, Carnegie Mellon University*, 1994.

[3] D. Stodolsky, G. A. Gibson, and M. Holland, "Parity Logging Overcoming the Small write Problem in Redundant Disk Arrays," *Proceeding of the 20th Annual International Symposium on Computer Architecture*, pp. 190-199, May. 1993.

[4] E. K. Lee, "Performance Modeling and Analysis of Disk Arrays," *PhD Thesis, Carnegie Mellon University*, 1994.

[5] R. R. Muntz and J. Lui, "Performance Analysis of Disk Arrays Under Failure," *Proceedings of 16th VLDB Conference*, pp. 162-173. 1990.

[6] Alexander Thomasian and Jai Menon, "RAID5 Performance with Distributed Sparing," *IEEE Trans. on Parallel and Distributed Systems*, Vol.8, No.6, pp. 640-657. June. 1997.

[7] J. Menon and R. Mattson. "Comparison of sparing alternatives for disk arrays," *Proceeding of International Symposium on Computer Architecture*, May. 1992.

[8] M. H .MacDougall, "Simulating Computer Systems," *MIT Press*, 1987.

[9] E. K. Lee, R. H. Katz, "Performance Consequences of Parity Placement in Disk Arrays." *International Conference on Management of Data(SIGMOD)*, pp. 190-199. 1991.

[10] J. Chandy and A. L. Narasimha Reddy, "Failure Evaluation of Disk Array Organization," *Proceedings of the International Conference on Distributed Computing Systems, IEEE Computer Society, Washington D.C.*, 1993.

[11] E. K. Lee, "Software and Performance Issues in the Implementation of a RAID Prototype," *Technical Report UCB/CSD 90/573, University of California at Berkeley*, May. 1990.

[12] Peng Cheng, Hai Jin, Jiangling Zhang, "Design of High Performance RAID in Real-Time System," Computer Achitecture News, Vol.27, N.3 pp. 10-17, Jun. 1999.

[13] St. Peter, "The RAIDBook: A Source Book for RAID Technology," RAID Advisory Board, 1996

[14] P. M. Chen, E. K. Lee, "Striping in a RAID level 5 Disk Array," *Technical Report University of Michigan*, 1993.

전상훈

1992년 2월 영남대학교 전산공학과(공학사). 1994년 2월 영남대학교 전산공학과(공학석사). 1998년 8월 영남대학교 컴퓨터공학과 박사수료. 1999년 3월 ~ 현재 경동정보대학 인터넷정보계열 전임강사. 관심분야는 멀티미디어시스템, 정보통신, 컴퓨터구조, 차세대 인터넷

안병철

1976년 영남대학교 전자공학과 졸업. 1986년 오레건 주립대 전기 및 컴퓨터 공학과 석사학위취득. 1989년 오레건 주립대 전기 및 컴퓨터 공학과 박사학위 취득. 1978년 ~ 1984년 국방과학연구소 연구원. 1989년 ~ 1992년 삼성전자 컴퓨터 부문 수석 연구원. 1992년 ~ 현재 영남대학교 공과대학 컴퓨터공학과 부교수. 관심분야는 컴퓨터구조, 그래픽스, 멀티미디어 및 실시간 운영체제