Redundant Optical Storage System Using DVD-RAM Library

Takaya Tanabe, Makoto Takayanagi, Hidetoshi Tatemiti, Tetsuya Ura, and Manabu Yamamoto
tanabe@ilab.ntt.co.jp
NTT Integrated Information & Energy Systems Laboratories, Tokyo, Japan

Abstract

A Digital virtual Disk (DVD) random access memory (RAM) Redundant Array of Inexpensive Libraries (RAIL) optical storage system has been developed and tested at NTT Integrated Information & Energy Systems Laboratories. The RAIL storage system incorporates multiple DVD libraries that consist of dual DVD-RAM drives. Each DVD library utilizes a single mechanical robot picker for media loading and unloading. The current capacity of the single sided and single layered DVD optical media used in that system is 2.6 gigabytes. To increase the reliability of stored data and at the same time to eliminate the need for read after write verification, a process that can double the recording time, a RAID 4 algorithm has been implemented in the control unit of the RAIL storage system. Data sent by the host are transferred to a control unit, that stripes data over five data groups plus one parity unit. The striped and parity data are sent to individual libraries and written to DVD media. This system writes and retrieves storage data with a transfer rate of approximate 6 MB/sec, using write and read control methods that minimize data striping overhead. Other performance factors that affect the transfer rates are the striping size and the number of drives used in the RAIL system. Experimental results indicate that stripe sizes of 32 to 64 KB are sufficient to achieve high throughput. In addition, the transfer rates showed no further increase when the number of drives exceeded eight. This RAIL optical storage system which offers data redundancy can be used for networked multimedia applications.

1. Introduction

The proliferation of internet based systems, digital networks, cellular phones, and ISDN have increased the demand for multimedia network systems[1,2]. The technological progress made with optical technology and network systems have facilitated larger volumes of data to be transferred at faster speeds. This in turn has stimulated the expansion of electronic trading, business transactions, and the use of digital medical images (especially for medical care in remote areas). Highway traffic control, local administration, and other educational systems have contributed and benefited from this evolution but have increased network congestion. The implementation of digital broadcasting, which requires highly advanced server systems to achieve excellent performance, will further increase the use of digital images. To satisfy these growing needs, large capacity file systems must be developed to produce high performance server systems.

Optical disk systems can be used for quick random access storage systems with large capacity at relatively low cost. With their large capacity, they are widely used in applications such as backup for data processing systems, archival for long term storage of data, and multi-media service files to store photos and images. The demand for these optical systems requires an optical cataloguing system (disk library) capable of storing and retrieving reliably large volume of data at high speed. A mass storage system [3] based on magneto-optical disks had been previously developed and introduced to satisfy the need of database applications. However, due to the increasing volume of data to be processed and the demand for faster processing speed, these systems are now facing the inherent limits of optical drives.

To improve the overall performance throughput of a disks system while providing increased reliability, RAID has been proposed [4]. A similar concept has also been applied to an array of magnetic tapes [5] to improve speed and storage capacity. However, despite several recent studies [6] on Redundant Arrays of Inexpensive Libraries, very little has been reported.

Consequently, a Redundant Optical Storage system that consists of a RAIL has been developed. This optical storage system based on DVD-RAM drives, makes multiple libraries behaves like a single library (a “virtual” single library). This system uses high-speed transfer technology based on data flow control to reduce recording time and cost, and to increase the performance of optical storage system.

The focus of this paper is to report the results obtained with a RAIL system. First, the design of this system is described in section 2. Then the buffering and the high-speed transfer mechanisms are described in detail in section 3. Finally the performance of a RAIL system is presented in section 4.

2. System Architecture

The RAIL system is composed of 6 DVD libraries (one is used as parity), and a controller unit as shown in Fig. 1. Each DVD library consists of 2 DVD drives, holds 150 pieces of recording media, and has a robot mechanism (picker) to move the DVD media. The recording media is

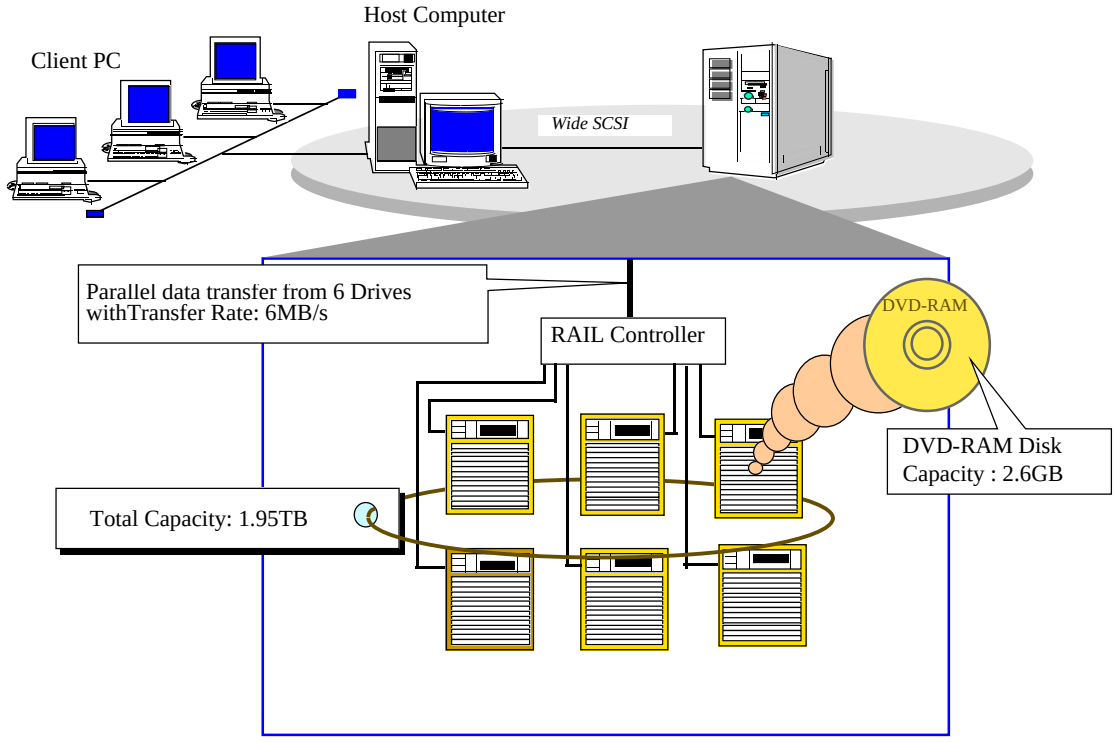

Figure 1. RAIL system architecture

single sided single layered and has a total capacity of 2.6 GB. To control each individual DVD library, a command is issued to move the DVD media tray from the shelf by the picker and to insert it into the DVD drive. After a read/ write process is finished, the picker removes the tray from the drive and places it back on the shelf. Because the pickers transport the media simultaneously from the same location of each library, the time delay to move media is kept to a minimum. In order to improve the overall data transfer rate of the RAIL system data is striped over five drives. To provide reliability, the 6th DVD drive is used as parity drive. Thus, through this design several libraries are combined to form a single virtual library that can hold up a total capacity of 1.95 TB (150 * 13 GB) and achieve a theoretical cumulative transfer rate five times faster than the one of a single DVD drive.

To enhance reliability, and to improve data availability the following protocols are followed:

(1) Errors occurring during media recording
Problems may occur during both the read and write processes. When a read error occurs, data is reconstructed using the correction algorithm of RAID-4. When a write error occurs, the data is recorded into an alternate sector. To further minimize the occurrence of errors, all media have been certified before use. When more than three PIDs (Physical Identification Data) out of the four PIDs are not accessible during recording or when more than eight data EEC blocks have more than four errors in their rows, data is recorded to new sectors. The DVD media life is guaranteed by the manufacturer to be over 10 years and has a corrected BER (bit error rate) of less than 10^{-12}.

(2) Faulty drives
If a drive failure occurs, a recovery process is automatically started and the second DVD drive of the same library is used. The new DVD media is then moved to the second drive. The failed drive is then marked as faulty to prevent it from being used. The faulty drive can be repaired or replaced during maintenance. The failure of both drives from the same library should be very rare; however, should it occur, data is reconstructed using the RAID algorithm.

(3) Faulty Robot pickers
When a picker problem occurs, the corresponding DVD library is disabled and data is reconstructed using the RAID algorithm. The faulty DVD library is then removed for repair. After being fixed, the DVD library is reinstalled and the RAID recovery process is executed to reconstruct data on the failed DVD library.

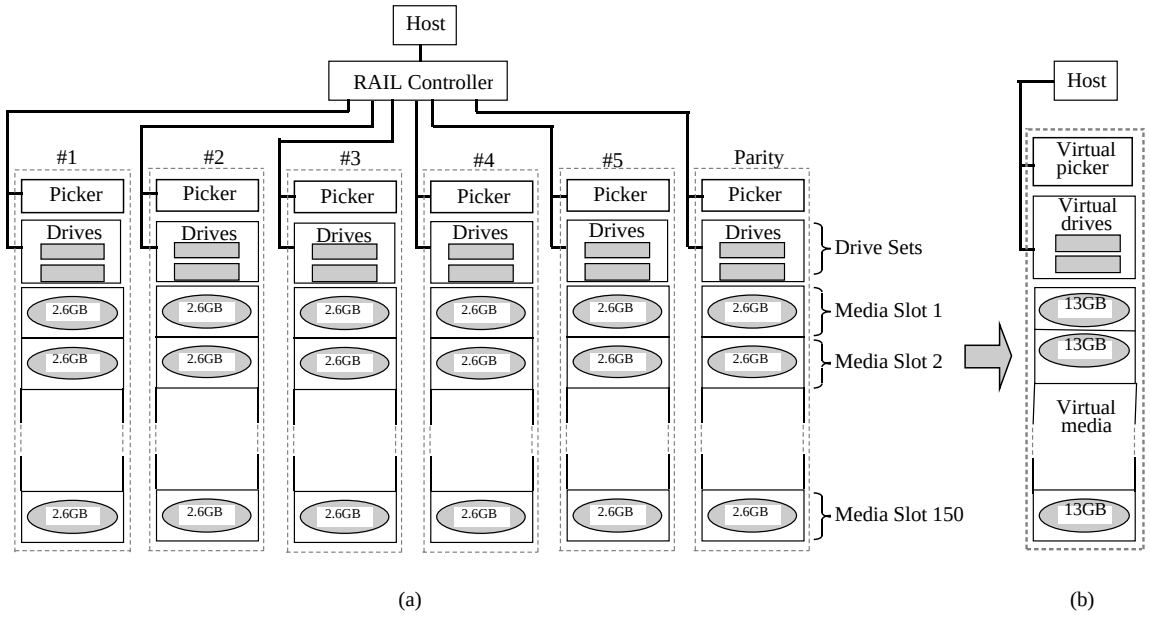

Figure 2. RAIL-4 system for parallel control of DVD libraries.
(a) Physical Library, (b) Virtual Library.

(4) Power failure
Both commercial power supplies and emergency batteries are used to prevent the loss of data in the RAIL controller and to improve the level of reliability of the entire RAIL system.

The architectural diagram of the RAIL system that controls multiple DVD libraries is shown in Fig 2. The key factor in performing parallel read and write operations managed by the RAIL controller is to make the drives function as efficiently as possible.

The RAIL system is connected to the host via a SCSI interface and responds as if it were a single library. When a media transfer command is issued, all DVD media located on the same slot of each library are transferred to their associated drives simultaneously. To achieve this, a command issued by the host is passed to each individual library. When a media read/write command is issued, the DVD in each library is loaded into the drive of each library. The host views the system as a single virtual drive and when it issues a command the RAIL controller distributes it to each individual drive. During a write operation, the controller divides the data according to a stripe size, creates a parity data, and sends the data to each drive. During a read operation, data collected from each drive is rearranged and then forwarded to the host. In the event of a read error, the controller attempts to recreate the data using the parity data stored in parity library. In the event of a picker failure the parity data is used to recreate the data and avoid a potential system shutdown. Note only one library can be down at a given time. Using these techniques the RAIL system can provide a high level of reliability.

3. RAIL Simulation

A device driver was written to help evaluate the effect the total number of DVD drives has on the RAIL system. The main function of the device driver is to stripe and to redirect data between the RAIL controller and the appropriate drives. We have defined several read/write protocols (or modes) that control how data flows between the host, the RAIL controller and the DVD drives. This is illustrated the simplified diagram in Figure 3. During a read/write operation, data from the host is first sent to the main buffer located within the RAIL controller. Then, the RAIL controller distributes the data to each of the drives according to the sequence defined below. Different modes have different data flow sequences as well as data stripe sizes. The device driver has been designed to measure the cumulative transfer rate when the number of drives and the size of the stripes are changed. The RAIL controller has a single processor that is equivalent to a 200 MHz Pentium Pro processor.

In the following paragraph we describe the different modes available to read and write data between the DVD drives and the host and then examined the relationship between system performance, the number of drives used, and the striping size. The different read and write modes and their significance related to the throughput of the data transfer are also introduced.

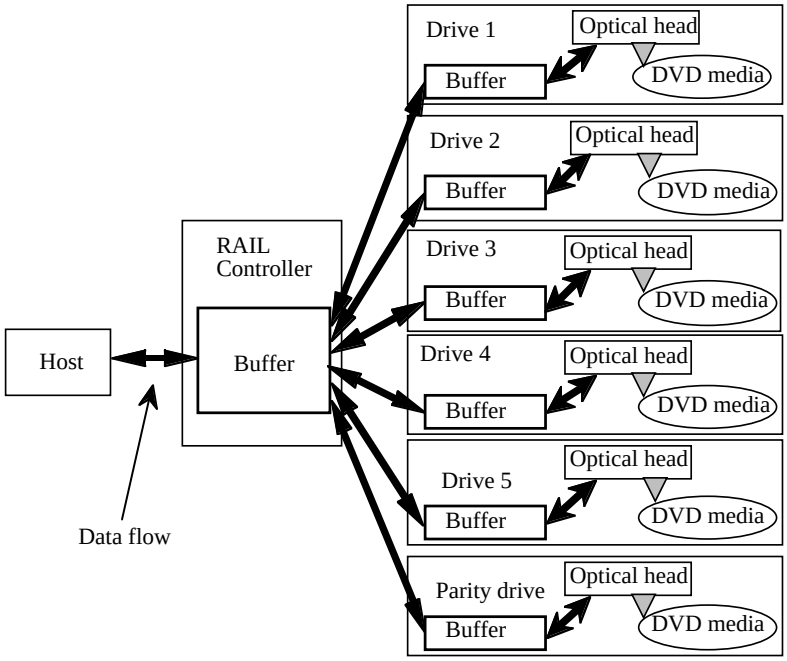

Figure 3. Diagram of the data flow within the RAIL system.

3.1 Write and Read modes

To determine the optimum number of libraries in a RAIL system, we studied the relationship between the total number of drives used and the aggregate transfer rate. In the tests conducted in read and write modes, we have connected a maximum of 2 DVD drives per SCSI bus. This system has 2 read modes and 3 write modes:

Write modes

1) Write mode 1: The following sequence of events occur in this mode:
   - Data sent to the RAIL controller is acknowledged by sending a receipt to the host
   - RAIL controller divides data for each drive
   - Data is accumulated in the controller buffer
   - Data is sent to the respective DVD drives
   - Drives report receipt of the data to the RAIL controller
   - Drives record data on the media

   When data transferred from host to the controller is completed, the host has to issue a synchronized cache command to the controller. Then, all the data is transferred to the drive and recorded on the media. After the controller receives the completion of the recording data at the drive, the controller answers the completion of the synchronized cache command to the host. In this mode there is no read after write and media reliability is increased by certifying each media before using it. Any error detected later on is corrected by relying on the ECC.

2) Write mode 2: This mode is the same as write mode 1 except that a verification is executed after the data is recorded on the media.

3) Write mode 3 (bufferless write mode): This mode does not use the drive buffers. Upon receiving data at the RAIL controller, the RAIL controller divides and sends it to the appropriate drive. The drive records the received data on the media, verifies the data, and reports the completion of the write process to the RAIL controller. After that, the controller reports its receipt to the host.

Read modes

1) Read mode 1: Upon receiving a read request from the host, the controller requests that each drive reads the data from the media and transfers the data to the RAIL controller. Then the controller collects the data from drives and send them to the host. In this case data that belong to the next consecutive sector on the media are also read and stored in the drive buffer. During a read request at each drive the buffer is first searched. When

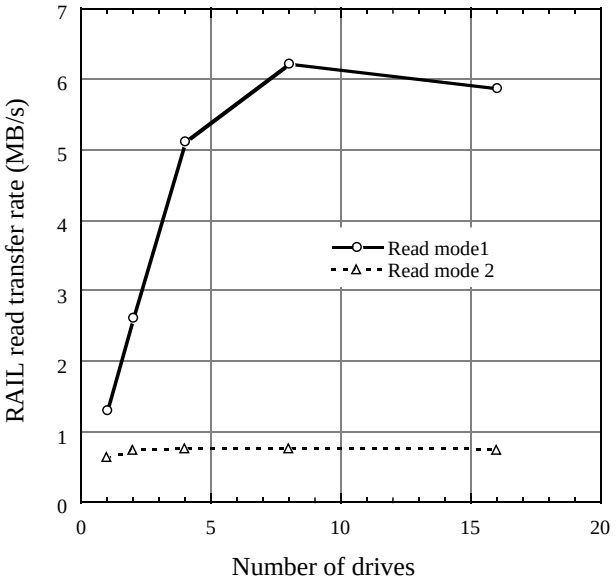

Figure 4. Relationship between the number of drives and the RAIL read transfer rate.

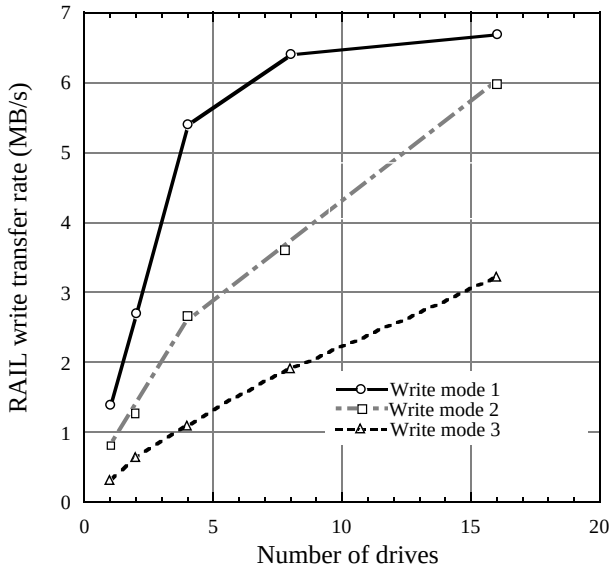

Figure 5. Relationship between the number of drives and the RAIL write transfer rate.

data are identified to be in the buffer they are retrieved directly from the buffer. Otherwise the data have to be read from the DVD media.

2) Read mode 2 (bufferless read mode): This mode does not use the drive buffers. Upon receiving a read request at the RAIL controller, then the controller requests that each drive reads the data from the media and transfers the data to the RAIL controller. Then the controller collects the data from drives and send them to the host. In this case, the data always have to be read from the DVD media.

3.2 Number of drives used

Figure 4 shows the relationship between the aggregate read transfer rate for different read modes, and the number of DVD drives used. The stripe size of each drive is set to 64 KB. In the tests conducted, data volumes of 128 MB were read in increments of 64 KB to measure the average transfer rates. In a Read mode 2 (bufferless read mode), the RAIL read rate shows no change because each drive request is executed sequentially. Conversely, in the normal read mode 1, when the number of drives is less or equal to four, the transfer rate increases almost linearly with the number of drives. However with more than eight drives, no further improvement in the transfer rate was observed. When many drives are involved, the time required to transfer data between the controller and the drive SCSI bus increases. Because of the substantial gain in read transfer rate, the RAIL system should be used in normal read mode 1.

The relationship between the number of drives used and the RAIL aggregate write transfer rate for various write modes is given in Fig 5. The conditions used to measure the throughput were the same as those used in the read modes. Write mode 1, shown in solid lines, exhibits the highest transfer rate for any number of drives. Write mode 2 does not show a high transfer rate compared to write mode 1 due to the need for the write verification process.

For the all write modes, the larger the number of drives used, the higher the RAIL transfer rate. However the transfer rate in normal write mode 1 shows only a small increase when the number of drives exceeds eight. In the write mode 3 (bufferless write mode) the transfer rate increases linearly even with a large number of drives. This can be attributed to the fact that during the bufferless write mode, all DVD drives can be operated at the same time. However, the transfer rate between the RAIL controller buffer and the DVD drive buffer needs to be sufficiently larger than the DVD drive write speed. Under bufferless write mode, the RAIL controller can dump the striped data to each DVD drive simultaneously without waiting for other DVD drives to finish writing the data onto the media.

It is important to keep the data stripe size sufficiently smaller than the DVD drive buffer capacity in order to prevent excess overflow of striped data at each DVD drive. However the transfer rate in write without buffer is always lower than the transfer rate in normal write modes. We note that the normal write mode 1 and the normal read mode 1 are completed with a single pass and the transfer rates in these 2 cases are very similar.

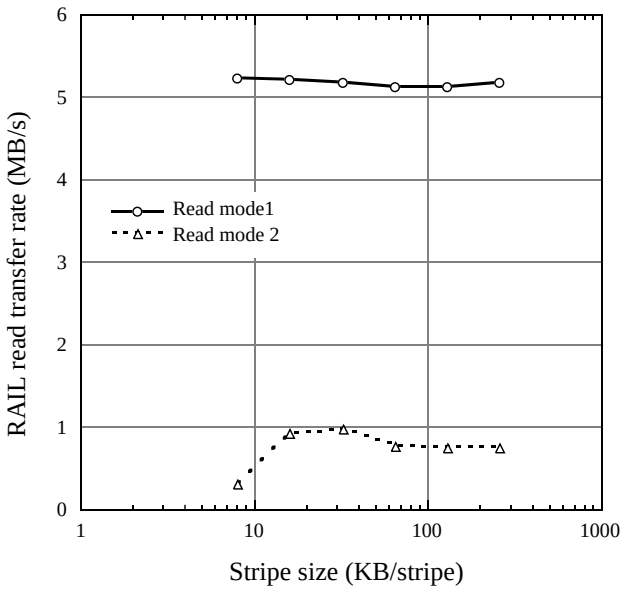

Figure 6. Relationship between tne stripe size and the RAIL read transfer rate.

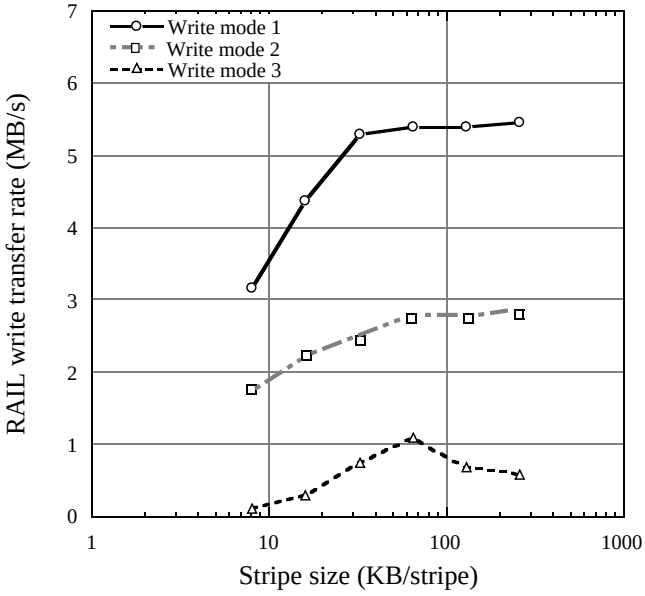

Figure 7. Relationship between the stripe size and the RAIL write transfer rate.

3.3 Stripe size

We studied the effect of the stripe size by using a configuration similar in 3.2 but connecting four drives to the RAIL controller. The relationship between the stripe size and the RAIL transfer rate for different read modes is illustrated in Fig 6. To measure the transfer rates, 128 MB of data were read from each drive using the same stripe size. When the stripe is larger than 64 KB, The data size transferred to each drive is divided into units of 64 KB. This 64 KB unit is inherent to the test software loaded on the RAIL controller.

In the normal read mode 1, the RAIL transfer rate is constant and appears to be independent of the stripe size. In the read mode 2 (bufferless read mode), the RAIL transfer rate is almost the same for stripe sizes greater than 16 KB but drops significantly for stripe sizes smaller than 10 KB. Typically, with optical technology, the Error Correction Code (ECC) and the data sectors have the same size, which is usually 512 Bytes, 1024 Bytes, or 2048 Bytes. However, in a DVD-RAM disk, the sector size is 2048 Bytes, and the error correction blocks are placed in block of 16 sectors. This implies that when small amount of data is requested (e.g., 8 KB), at least 32 KB has to be read. This in turns reduces the transfer rates for small volumes of data.

The relationship between RAIL transfer rate and the stripe size for varying write modes is shown in Fig 7. The same configuration used in read modes was applied to the write modes. Data transferred to each drive is divided in units of 64 KB. The RAIL transfer rate in a write mode 3 (bufferless write mode) peaks for stripe size of 64 KB but declines for larger stripe sizes. This resulted from the lowered efficiency of the parallel data transfer, which was caused by the sequential multi-accessing of each drive. In normal write mode 1 as well as in normal write mode 2, the RAIL transfer rate reaches its saturation level around 32 KB. For the write mode, the transfer rate remained the same even when the stripe size was larger. Therefore, we suggest that units of 32 KB to 64 KB are sufficiently large for both modes. Moreover, this stripe size range is better suited for small data size than using excessively large stripe size.

4. System evaluation

A RAIL system has been developed based on the results presented in section 3. This RAIL system incorporates a RAIL controller connected to 6 DVD libraries. One of these libraries is used exclusively as a parity library. Each library is connected to the RAIL controller with separate SCSI cable (8-bit). A host is connected to the RAIL controller with a wide SCSI cable (16-bit). The RAIL system has similar hardware implemented in section 3. However, its software was redesigned to implement RAID-4 using a parity library. Upon receiving a read/write command from the host, DVD media located at the same slot number of each library were transported simultaneously to the drives. In this RAIL system, data were striped over 5 drives in 64-Kbyte units to achieve good performance on the read/write transfer rate.

The RAIL system has been evaluated with respect to three write modes and a single read mode summarized in table 1. Here, write mode 3 does not use the RAIL and

Table 1. The buffer conditions and RAIL transfer rate of write and read modes.

Write/Read mode	RAIL controller buffer	Drive buffer	Write verification at drive	RAILtransfer rate (MB/s)
Write mode 1	Use	Use	No	6.2
Write mode 2	Use	Use	Yes	3.4
Write mode 3	No use	No use	Yes	0.26
Read mode 1	Use	Use	-	6.6

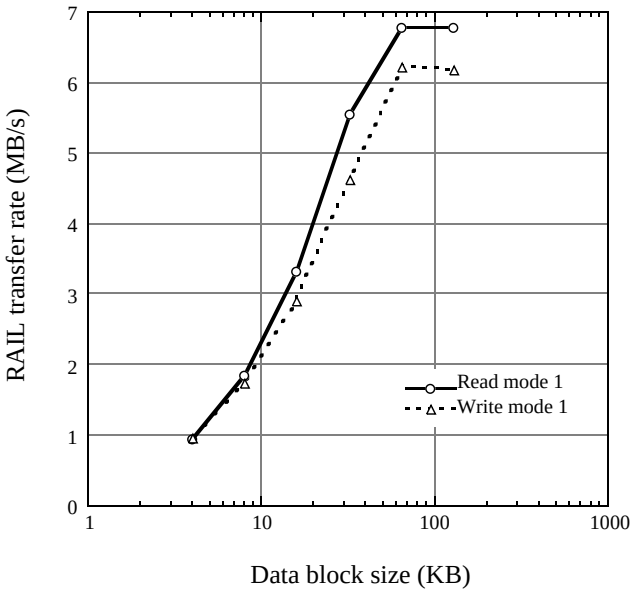

Figure 8. Relationship between the data block size and the transfer rate using write mode 1 and read mode 1.

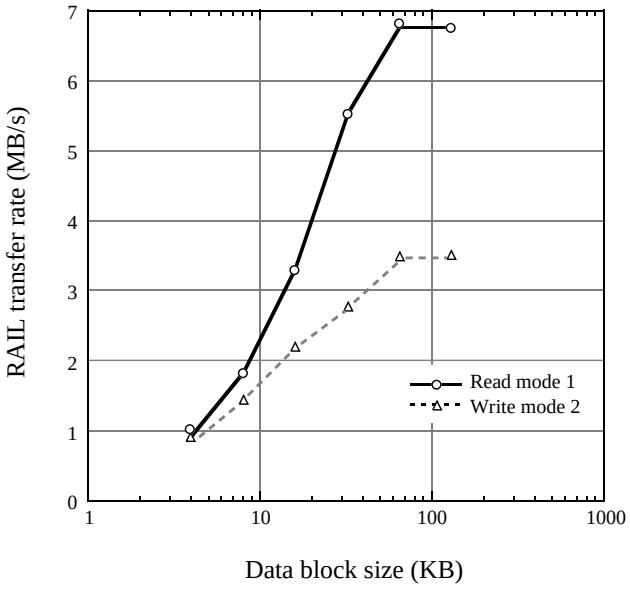

Figure 9. Relationship between the data block size and the transfer rate using write mode 2 and read mode 1.

drive buffers. Read mode 2 (bufferless read mode) is deleted because it shows a low read performance of the transfer rate.

The relationships between the block size and the transfer rate using the write mode 1, 2 and 3 are given in Figs. 8, 9 and 10. In Fig. 8, the transfer rate increases with the size of the blocks until 64 KB is reached and this for both read and write modes.

With write mode 2 in Fig. 9, data verification is applied to the drives resulting in a decrease in the transfer rate equal to almost half the write transfer of the read mode 1. In write mode 3 in Fig. 10, the buffers of the RAIL controller and disks were not used but read after write operation was executed. This resulted in extremely low write throughput for all block sizes. Therefore this write mode will not be used in general.

In this situation, data from the faulty disk are recovered using the data from the parity disk. As expected the transfer rate decreased but not substantially (See Fig 11).

5. Conclusion

A high reliable and low cost RAIL system consisting of 6 DVD libraries connected to a single controller has been developed at the NTT Integrated Information & En-

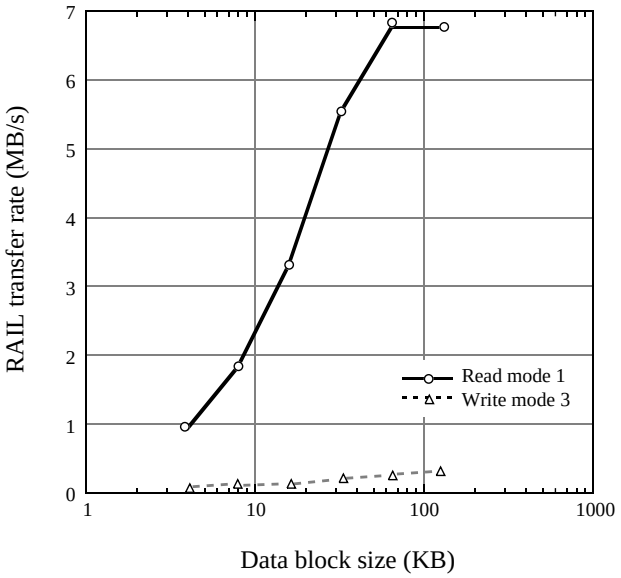

Figure 10. Relationship between the data block size and the transfer rate using write mode 3 and read mode 1.

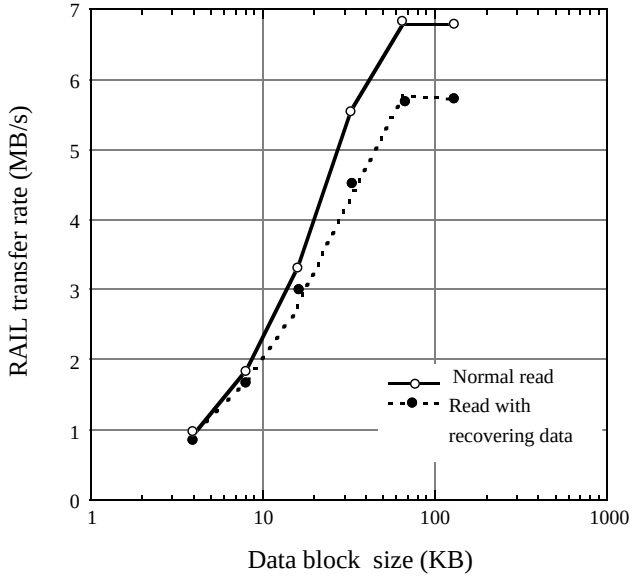

Figure 11. Read performance when one disk error occurred.

ergy systems laboratories. In this system, one of the DVD libraries is used as parity to provide redundancy and disaster recovery. DVD media holds 2.6 GB each, for a total RAIL capacity of 1.95 TB. The RAIL system has been found to have optimum performance when the stripe size is between 32 KB to 64 KB. The speed of the RAIL system developed is remarkably high (6 MB/sec) as compared to conventional system using a single optical drive. In addition the RAIL transfer rates show no further increase when the number of DVD drives exceeded eight. This system can be used in applications such as backup of database that requires large volumes of data to be stored.

References

1. R. Arai, and K. Itao, “Network File for ISDN,” Proceedings of Tenth IEEE Symposium on Mass Storage Systems, pp. 140-153, 1990.
2. M. Kawarasaki, T. Saito, H. Koyano, and N. Shigeta, “Service Strategies for Next-Generation Computer Networks,” NTT Review, Vol. 10, No. 4, pp. 100-109, 1987.
3. I. Yamada, M. Saito, A. Watabe, K. Itao, “Automated Optical Mass Storage Systems with 3-beam Magneto-optical Disk Drives,” Proceedings of Eleventh IEEE Symposium on Mass Storage Systems, pp. 149-154, 1991.
4. D. A. Patterson, G. A. Garth and R. H. Katz, “A Case for Redundant Arrays of Inexpensive Disks (RAID),” Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 109-116, 1988.
5. A. L. Drapeau and R. H. Katz, “Striped Tape Arrays,” Proceedings of Twelfth IEEE Symposium on Mass Storage Systems, pp. 257-265, 1993.
6. D. A. Ford, R. J. T. Morris and A. E. Bell, “Redundant Arrays of Inexpensive Libraries (RAIL): A Tertiary Storage System,” Proceedings of COMPCON ’96, pp. 280-285, 1996.