Design and Implementation of Scalable Tape Archiver

Toshihiro Nemoto, Masaru Kitsuregawa, Mikio Takagi
Institute of Industrial Science, University of Tokyo
7-22-1, Roppongi, Minato-ku, Tokyo, Japan
{nemoto,kitsure,takagi}@tkl.iis.u-tokyo.ac.jp
Tel: +81-3-3402-6231
Fax: +81-3-3423-2834

Abstract

In order to reduce costs, computer manufacturers try to use commodity parts as much as possible. Mainframes using proprietary processors are being replaced by high performance RISC microprocessor-based workstations, which are further being replaced by the commodity microprocessor used in personal computers. Highly reliable disks for mainframes are also being replaced by disk arrays, which are complexes of disk drives.

In this paper we try to clarify the feasibility of a large scale tertiary storage system composed of 8-mm tape archivers utilizing robotics. In the near future, the 8-mm tape archiver will be widely used and become a commodity part, since recent rapid growth of multimedia applications requires much larger storage than disk drives can provide. We designed a scalable tape archiver which connects as many 8-mm tape archivers (element archivers) as possible. In the scalable archiver, robotics can exchange a cassette tape between two adjacent element archivers mechanically. Thus, we can build a large scalable archiver inexpensively. In addition, a sophisticated migration mechanism distributes frequently accessed tapes (hot tapes) evenly among all of the element archivers, which improves the throughput considerably. Even with the failures of some tape drives, the system dynamically redistributes hot tapes to the other element archivers which have live tape drives. Several kinds of specially tailored huge archivers are on the market, however, the 8mm tape scalable archiver could replace them.

To maintain high performance in spite of high access locality when a large number of archivers are attached to the scalable archiver, it is necessary to scatter frequently accessed cassettes among the element archivers and to use the tape drives efficiently. For this purpose, we introduce two cassette migration algorithms, foreground migration and background migration. Foreground migration transfers a requested cassette from an element archiver whose drives are busy to another element archiver whose drives are idle. Background migration transfers cassettes between element archivers to redistribute frequently accessed cassettes, thus balancing the load of each archiver. Background migration occurs the robotics are idle. Both migration algorithms are based on access frequency and space utility of each element archiver. To normalize these parameters according to the number of drives in each element archiver, it is possible to maintain high performance even if some tape drives fail. We found that the foreground migration is efficient at reducing access response time. Beside the foreground migration, the background migration makes it possible to track the transition of spatial access locality quickly.

I. Introduction

Recently, large scale tertiary storage systems are becoming is more and more desired for multimedia applications or scientific data such as satellite images. Today, magnetic disks have become cheap with large capacity, however, this capacity is still not enough to archive multimedia data or satellite images. To archive huge data sets, magnetic tape archivers which have some tape drives and robotics for management of the tapes in them, are often used, but most of the current commercial tape archivers are not scalable and there is no way to migrate data from one archiver to another except by copying the data. Accordingly it takes a long time to redistribute data.

To address these issues and aiming towards scalable commodity archivers, we have been developing a scalable tape archiver for satellite images. It consists of some commodity element archivers and tape migration units between two adjacent element archivers. We believe a reasonable size tape robotics will become a commodity component in the near future. It is easy to add or remove element archivers to the scalable tape archiver at any time and any number of element archivers can be attached. To redistribute data, a cassette is transferred from one element archiver to another through the tape migration unit instead of copying data.

In this paper, we present a cassette migration mechanism for the scalable archiver and its performance evaluation. The scientific data such as satellite images are characterized not only by their size but also by access locality. Accordingly, when storing these data, efficient utilization of the tape drives and the proper positioning of frequently accessed cassettes substantially affects the performance. In order to achieve high performance in spite of changing access locality, two load balancing mechanisms, foreground migration and background migration, are introduced to the scalable tape archiver. The foreground migration transfers a requested cassette from an element archiver whose drives are busy to another element archiver with idle drives. Background migration transfers a cassette to redistribute frequently accessed cassettes between idle element archivers. The foreground migration is efficient at reducing access response time. Beside the foreground migration, the background migration makes it possible to track the transition of spatial access locality quickly.

I. Design of The Scalable Tape Archiver

The scalable tape archiver is composed of any number of small size tape archivers (element archivers) and cassette migration units connecting any two adjacent element archivers. Figure 1 shows the organization of the experimental scalable tape archiver using an 8mm tape jukebox, NTH-200B, as the element archiver. The NTH-200B has two Exabyte 8505 tape drives, a tape handler robot and a cassette rack with 200 slots. It also has a controller for its own tape handler robot and for the tape migration unit on its right. The host computer sends commands for holding, releasing and moving a tape and so on to the controller and receives the status of the element archiver through an RS-232C port. The tape handler robot takes a cassette from a slot in the rack or from the drives and places it in another slot or drive according to the command received. The tape drives are normal Exabyte 8505's and are connected to the host computer through a SCSI bus. The tape migration unit has a wagon to migrate a cassette tape to another element archiver. Cassette tape migration is executed as follows.

1. The tape migration unit brings the wagon back into the source element archiver, if the wagon is not currently in the source element archiver.
2. The tape handler robot in the source element archiver takes the cassette to migrate from a slot or a drive.
3. The tape handler robot places the cassette in the tape migration unit's wagon.
4. The tape migration unit sends the wagon from the source element archiver to the destination element archiver.
5. The tape handler robot in the destination element archiver picks up the cassette tape from the wagon, and places the tape into the appropriate slot or drive.

These steps are coordinated so that the counterweight of the tape handler robot does not interfere with the movements of the tape migration unit.

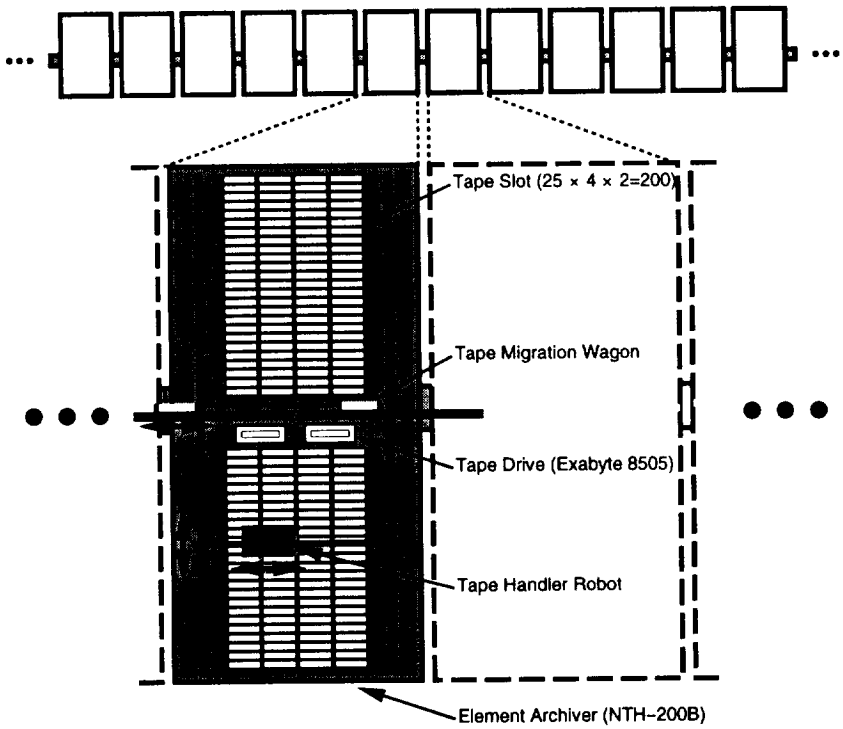

Figure 1:Organization of Experimental Scalable Tape Archiver using NTH-200B

I. Cassette Migration Strategy

A. Access locality

First, we describe the heat and temperature metrics [1]. The heat is the access frequency of a cassette or an archiver over some period of time. The heat of a cassette is the sum of its access frequencies and the heat of an archiver is the accumulated heat of the cassettes in it. Temperature is defined as the heat of a cassette or as the heat of the archiver divided by the number of tapes it contains.

High access locality hinders the efficient use of the archivers. If hot cassettes are concentrated on a few element archivers, the hot element archivers may receive too many

tape access requests leaving the cold element archivers idle. To reduce the concentration of accesses and to improve efficient use of the resources, it is necessary to scatter the frequently accessed cassettes around the scalable tape archiver. For this purpose, two load balancing mechanisms, foreground migration and background migration, are introduced to the scalable tape archiver.

A. Foreground migration

When a new access request is issued for a tape in an element archiver where all drives are currently in use, moving the tape to another element archiver and using a free drive can reduce the response time of the request. We call such migration foreground migration. When there are several element archivers which can accept a new cassette, that is, which have one or more empty slots and idle drives, and whose tape handler robot is idle, the following four strategies are used to select which element archivers to migrate the cassette to.

- **Random**: In the random strategy, the destination element archiver of the migrated cassette is selected at random.

- **Space Balancing**: When an element archiver is full, it cannot accept a new cassette to migrate into even if it has idle drives. Therefore, the destination to migrate is the element archiver in which the number of cassettes is smallest in the space balancing strategy.

- **Heat Balancing**: Selecting an element archiver whose heat is lowest to migrate a cassette to balance the heat of each element archiver.

- **Distance Minimizing**: The nearest element archiver is selected as the destination of migrated cassette. The intermediate element archivers between source and destination element archivers cannot serve any request while they relay the migrated cassette.

A. Background migration

When it is possible to migrate a cassette between two element archivers, that is, when all of the tape handler robots and migration units between the source and destination element archivers are idle, migrating a cassette can balance the number of cassettes or heat of each element archiver. We call such migration background migration. In background migration, the cassette is always migrated from the element archiver which has more cassettes to the one holding fewer cassettes. A migrated cassette is selected to balance the element archivers the most. For example, a hot cassette is selected for migration when the heat of the source element archiver is larger than the destination's and a cold one is selected in the opposite case. When more than two background migration can be executed at the same time, the following two basic strategies are used to determine the source and destination archivers.

- **Space Emphasizing**: The space difference minimizing strategy selects the pair of element archivers whose number of cassettes differs the most.

- **Heat Emphasizing**: The heat difference minimizing strategy selects the pair of element archivers whose heat difference is largest.

I. Performance Evaluation

A. Description of the simulation

To evaluate the basic performance of the scalable tape archiver, we execute computer simulations to measure average response time. The simulation parameters shown in Table 1 are based on the real scalable tape archivers using the NTH-200B. We assume that each cassette tape has fifty data in it and the size of each data is 100 MB. Accordingly, the read/write time of one data is always 200 seconds. The minimal cycle time is 487 seconds[1] on average. The interval time of request arrival depends on a negative exponential distribution. Because the destination element archiver should have a vacant slot for the migrated cassette, we selected 95% as the load factor. The scalable tape archiver consists of sixteen element archivers and the initial distribution of the cassette tapes in the scalable tape archiver is shown in Table 2. The access locality follows an 80/20 rule, that is 80% of the accesses are to 20% of the cassettes.

A. Simulation results

Figure 2 shows the average response time after 50,000 accesses from the initial cassette distribution. Compared to the result of no migration, response time is significantly reduced when only foreground migration is introduced into the scalable tape archiver. Furthermore, using background migration can produce in addition more improvement. Figure 3 shows the average response time at intervals of even 2,000 accesses where the request arrival rate is 0.045 requests per second. Between the two background migration strategies, there is no difference, but using background migration makes it possible to track the changing of access locality quickly.

Table 1: Simulation Parameters

Element archiver	
Number of element archivers	16
Maximum number of cassettes in an element archiver	200
Number of drives in an element archiver	2
Drive	
Drive setup time	35sec
Seek speed	25 MB/sec
Read/Write speed	0.5 MB/sec
Tape eject time	20 sec
Tape hander robot and tape migration unit	
Robot move time	2 sec
Robot move time with holding and placing cassette	14 sec
Wagon unit move time	9 sec

[1] Robot move time + robot move time with holding and placing cassette + drive setup time + average seek time + read/write time + average seek time (for rewinding) + tape eject time + robot move time + robot move time with holding and placing cassette

Table 2: Initial cassette tape distribution

Elem. Archiver No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Hot Cassettes	8	8	8	8	8	88	88	88	88	88	88	8	8	8	8	8
Cold Cassettes	182	182	182	182	182	102	102	102	102	102	102	182	182	182	182	182
Total	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190

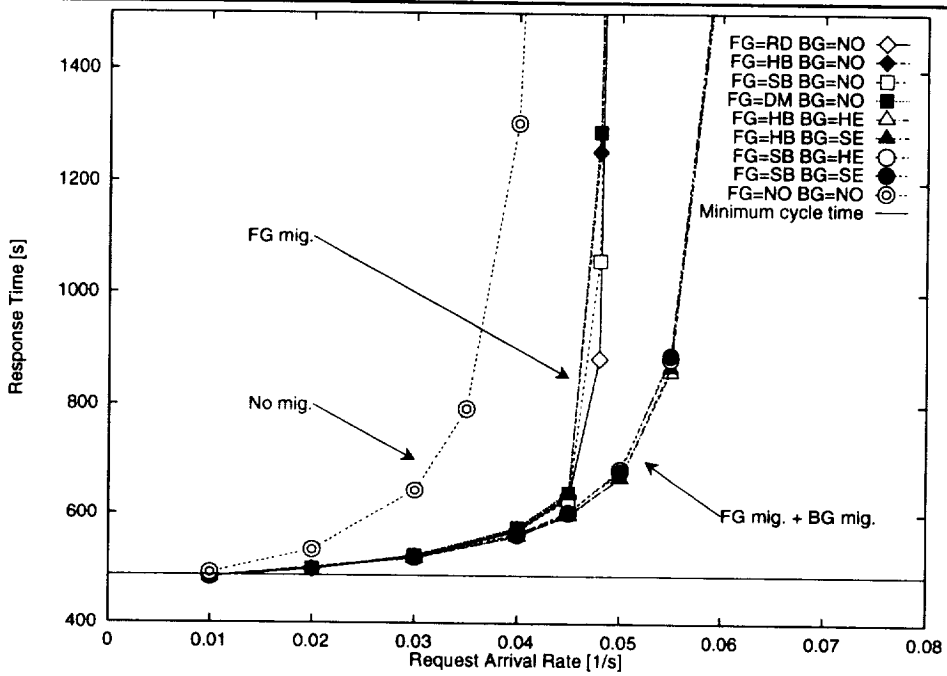

Figure 2: Average response time of initial 50,000 accesses

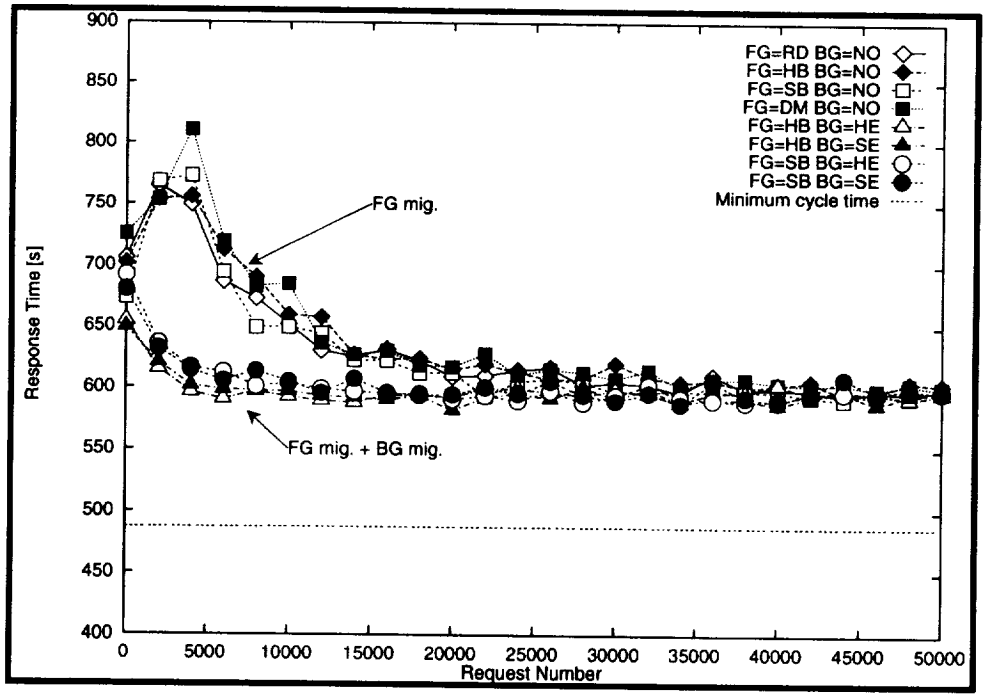

Figure 3: Average response time for intervals of 2,000 accesses

Figure 4 shows the average response time after 50,000 accesses from the initial cassette tape distribution when all tape migration wagons in the scalable tape archiver move slower. It takes 30 second to move the slower wagon from an element archiver to another, while it takes 9 second to move our experimental scalable tape archiver's. The cassette tape migration achieves better performance than using no migration even if the tape migration wagons are slower. The slower tape migration units do not deteriorate the performance very much. Therefore it is not necessary to use expensive high speed tape migration units. To connect some inexpensive small size commercial tape archivers with low cost tape migration mechanisms improves the performance significantly.

Figure 5 shows the average response time at intervals of 2,000 accesses from the initial state. In this simulation, a drive in the eighth element archiver fails when the scalable tape archiver receives 10,000 requests and it recovers after receiving 20,000 more requests. Figure 6 also shows the average response time at intervals of 2,000 accesses from initial state where both drives in the eighth element archiver fail and recover. The heat balancing strategy and the space emphasizing strategy are selected as foreground migration and background migration respectively. The average response time of the scalable tape archiver is not affected by the single drive failure significantly. Two drives failure deteriorates the average response time when the request arrival rate is 0.055. However, the scalable tape archiver can serve requests for the tape in the eighth element archiver, which has no drive in it, while ordinary archivers do not work in this situation.

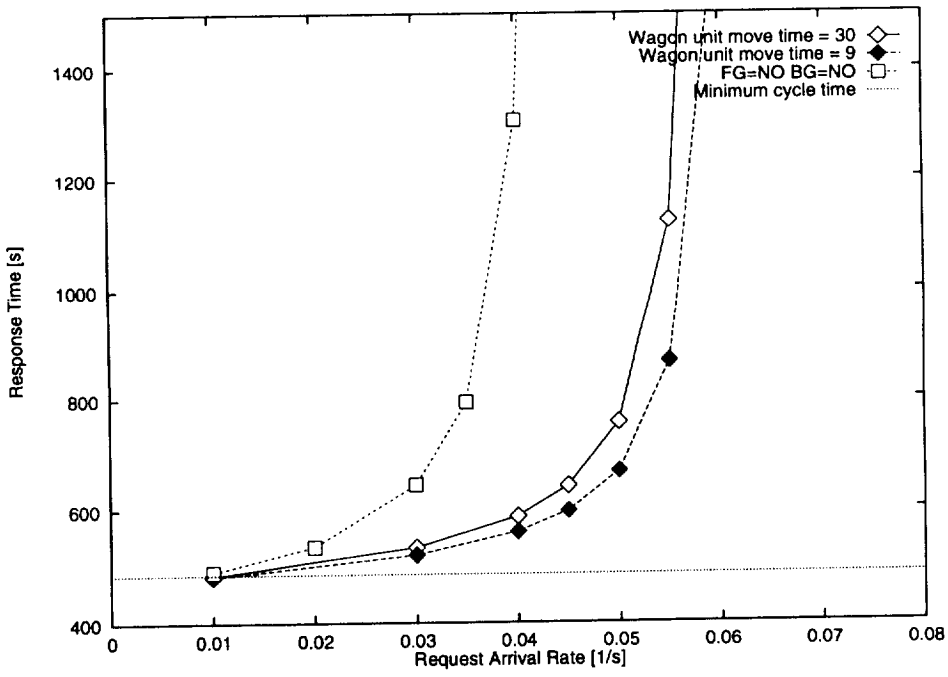

Figure 4:Average response time of slow migration wagon archiver

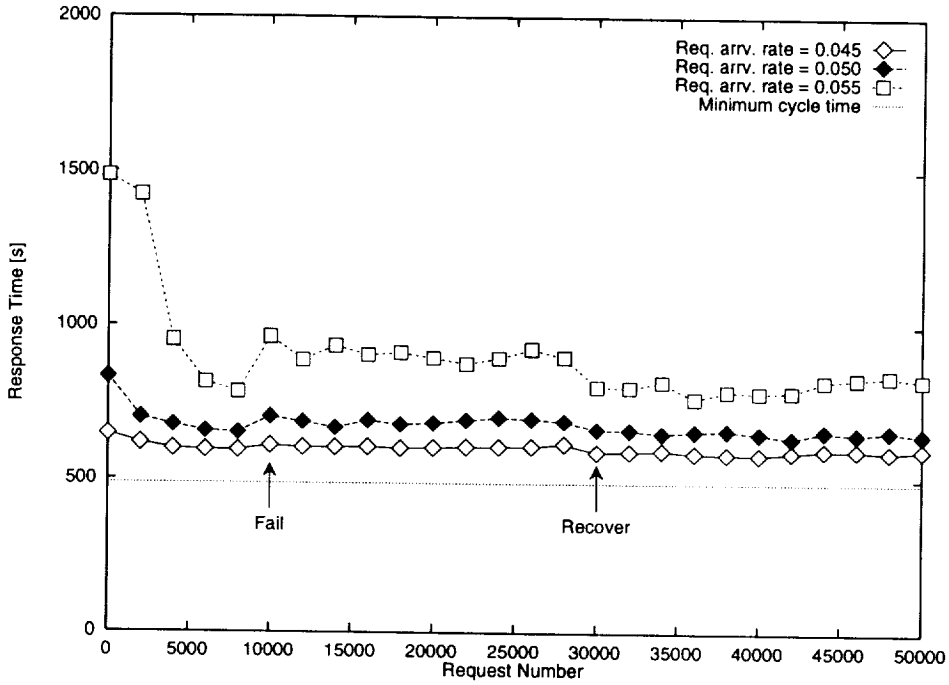

Figure 5: Average response time measured at intervals of 2,000 requests with single drive failure

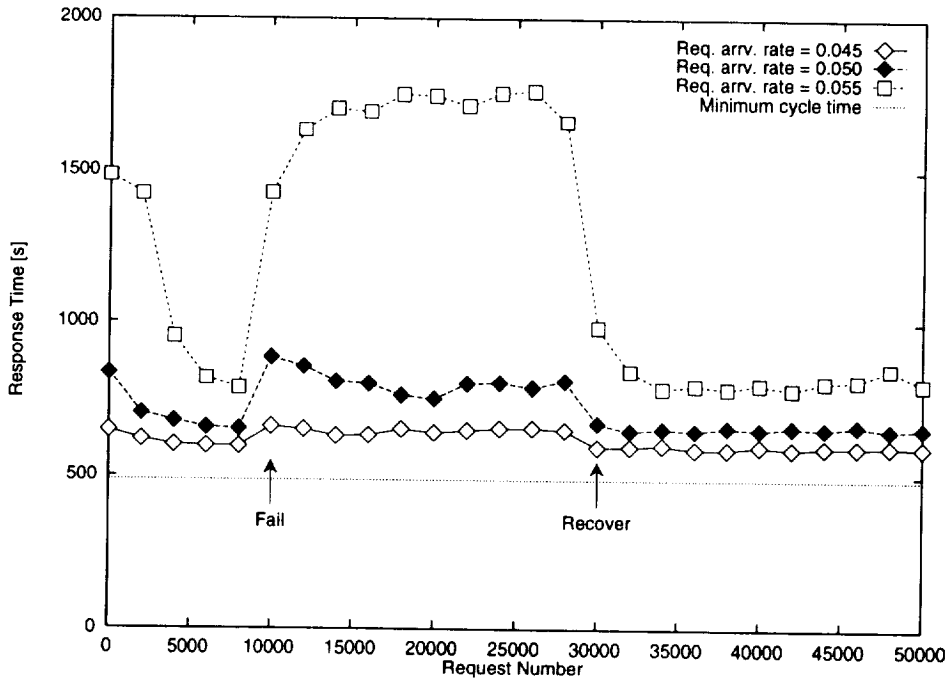

Figure 6: Average response time measured at intervals of 2,000 requests with both drives failure

I. Conclusion

In this paper, we described the design of a scalable tape archiver and the cassette migration algorithms for it. Only using foreground migration led to a large improvement in performance of the scalable archiver. In addition to foreground migration, background migration can improve performance even move. Using background migration together with foreground migration, the scalable tape archiver can continue to serve the requests, even when some drives fail.

We have already finished designing and developing the hardware of the scalable tape archiver and are now developing software for the scalable tape archiver. In the future we will examine the behavior of the scalable tape archiver with data striping.

References

1. Copeland, W. Alexander, E. Boughter, and T. Keller. "Data placement in bubba." Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data, pp. 99-109, 1088

2. Weikum, P. Zabback, and P. Scheuermann. "Dynamic file allocation in disk arrays". Proceedings of the 1991 ACM SIGMOD International Conference on Management of Data, pp. 406-415, 1991.

3. Nemoto, Y. Sato, K. Mogi, K. Ayukawa, M. Kitsuregawa, and M. Takagi. "Performance evaluation of cassette migration mechanism for scalable tape archiver". SPIE Proceedings, "Digital Image Storage and Archiving System", vol. 2606, pp. 48-58, SPIE, 1995.