# Technologies of ETERNUS Virtual Disk Library

● Shigeo Konno ● Tadashi Kumasawa

(Manuscript received September 26, 2005)

**In today's dramatically changing business environment, the extensive broadband environment and digitization of information enable huge amounts of information to be distributed over a network. The volume of fixed content such as multimedia data and intra-company data such as e-mail archives is rapidly increasing. Because the value of this data changes over time, the most suitable storage system for the value should be selected. The ETERNUS Virtual Disk Library answers this need by autonomously controlling data according to its value and reduces not only the hardware cost of a storage system but also the operational data management cost. This paper outlines the ETERNUS VD800 virtual disk controller and describes how it performs autonomous data control in the ETERNUS Virtual Disk Library.**

## 1. Introduction

The amount of data being handled in business activities is increasing at an annual rate of 60 to 70%. When choosing and operating a storage system, it is necessary to consider the data lifecycle management in terms of the data's value. Efficient, high-speed access, high-availability, large-capacity storage systems have been introduced in mission-critical areas. However, the need for low-cost, mass-storage systems that can store data more efficiently is also increasing. For instance, in a typical storage system, more than 90% of the data has not been accessed in the last 90 days, which is very inefficient. To reduce the operating costs of storage systems, frequently accessed data and updated data should be stored on a high-performance storage system such as a disk array, and infrequently or seldom accessed data should be stored on a low-cost storage system such as a tape library. To meet these requirements, Fujitsu has developed the ETERNUS VD800 virtual disk controller (hereafter VD800). The VD800 can be combined with the ETERNUS series of LTO tape libraries to form an ETERNUS Virtual Disk Library in which tape libraries can be accessed as virtual disks. This paper outlines the configuration of the VD800 and the ETERNUS Virtual Disk Library. It then describes the features of the ETERNUS Virtual Disk Library for realizing high-efficiency data storage and operational data management.

## 2. Overview of ETERNUS virtual disk controller and ETERNUS Virtual Disk Library configurations

The ETERNUS Virtual Disk Library consists of the VD800 and ETERNUS LT270 LTO tape libraries (hereafter LT270). As shown in **Table 1**, there are three models of the VD800.

The VD800 consists of a high-performance primary storage, storage processors, and Fiber Channel switches (**Figure 1**). The primary storage is an ETERNUS disk array, and the virtual disk engine (VDE) feature is implemented in the

Table 1
ETERNUS VD800 lineup.

| ˜ | Model 60˜ | Model 250˜ | Model 850˜ |
|---|---|---|---|
| Capacity of virtual disk˜ | Max. 549.6 TB˜ | Max 2, 198.4 TB˜ | Max 8, 192.0 TB˜ |
| Number of logical volumes˜ | Max. 1024˜ | Max. 4096˜ | Max. 4096˜ |
| Capacity of primary storage˜ | Max. 60 TB˜ | Max. 120 TB˜ | Max. 120 TB˜ |
| Number of LT270 tape˜ libraries supported˜ | Min.: 1˜<br>Max.: 2˜ | Min.: 2˜<br>Max.: 8 | Min.: 2˜<br>Max.: 32 |
| Host interface | Fibre Channel (2 Gb/s) | | |

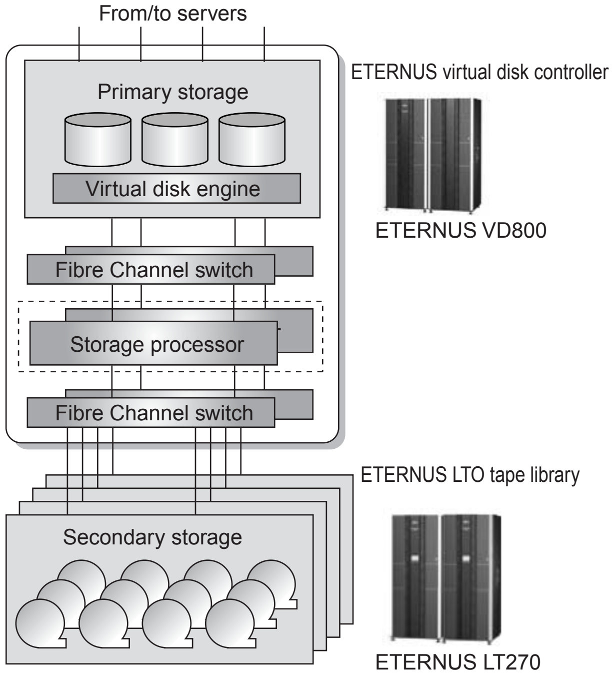

Figure 1
Configuration of ETERNUS Virtual Disk Library.

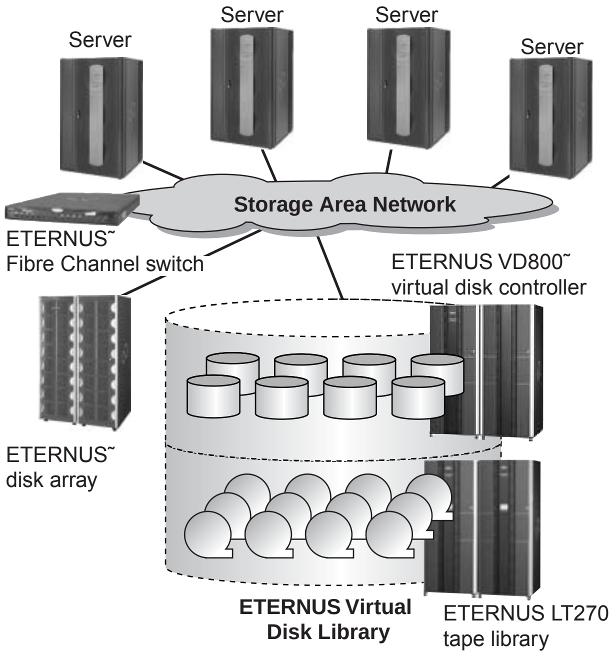

Figure 2
Overview of ETERNUS Virtual Disk Library.

disk array control firmware. The secondary storage is an LT270, which provides high-reliability, large-capacity data storage. The storage processor is an ETERNUS network storage server. A hierarchical storage control software is installed in the storage processors. This software controls data movement between the primary and secondary storages in cooperation with the VDE of the disk array. All components are connected by a high-speed 2 Gb/s Fibre Channel through the VD800's Fibre Channel switches, and all data access paths are duplicated. Furthermore, two storage processors are clustered using the high-reliability PRIMECLUSTER clustering software. If one of the storage processors fails, the other one continues to provide the services.

When combined with LT270s, the Model 850 provides a huge virtual disk space of up to 8 peta bytes and supports up to 32 LT270s.

The VD800 is connected to the Storage Area Network (SAN) environment, and the host application programs can access the entire space in the ETERNUS Virtual Disk Library as a real disk (**Figure 2**). The VD800 conceals accesses to the tape library, so applications running on the host server do not need to consider them. As a result, it can be used as a regular external disk device and there is no need to change the current operational design. The VD800 is therefore easy to introduce to an existing environment.

Up to 4096 virtual logical volumes can be defined in the ETERNUS Virtual Disk Library. All virtual volumes are managed in a fixed block unit called the Migration Recall Block (MRB). Data is controlled and transferred between the primary and secondary storages in MRBs.

There are three kinds of data movements between the primary and secondary storages: SYNC, MIGRATION, and RECALL. Data in the primary storage is copied (using SYNC) or moved (using MIGRATION) to the secondary storage based on the specified policy (described later). SYNC copies the original data on the primary storage to the secondary storage and synchronizes the data between the two. After SYNC is completed, the original data remains on the primary storage. A subsequent access to the same data by the host will therefore result in a disk hit at about the same high speed as an access to a regular disk array. MIGRATION copies the original data to the secondary storage, and then the original data is deleted from the primary storage. When the host tries to access the migrated data, a disk miss occurs and the data on the secondary storage is copied to the primary storage (RECALL) and then accessed.

## 3. Features

The ETERNUS Virtual Disk Library provides the following features.

### 3.1 Primary storage

1) Virtual disk engine (VDE)

The VDE dynamically associates the virtual volume and the physical disk space on the primary storage in MRB units and converts the virtual volume address of host accesses into an address on a physical disk (**Figure 3**). The necessary space for the MRBs is reserved on the physical disk. To prevent converged access to disks, the VDE efficiently allocates MRBs over the entire space of physical disks. To realize a virtual disk space that is larger than the space available on the physical disks, the VDE determines which MRB should

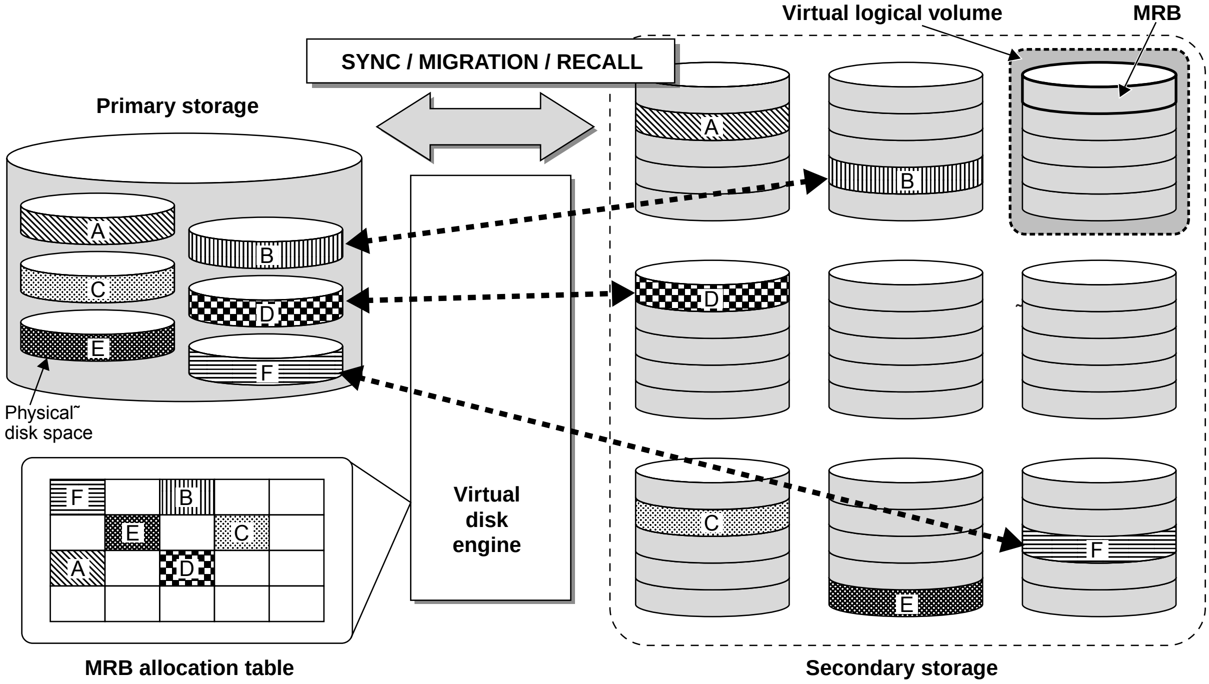

Figure 3
MRB allocation between primary and secondary storages.

be made the target of SYNC and MIGRATION. The target MRB is then reported to the storage processor, and copying from the primary storage to secondary storage is scheduled. Because of the need for fast-forwarding/reversing and the relative slowness of a library's robot mechanism, it usually takes much longer to access data on a tape than it does on a disk. Therefore, to speed up tape accesses, the VDE monitors whether the host is accessing contiguous data blocks. When contiguous data accesses are detected and a disk miss occurs, the VDE pre-fetches the next data block from the secondary storage in advance.

2) Water mark control

In order to efficiently manage the usable space of the physical disks, the user can define a High Water Mark (HWM) and Low Water Mark (LWM) in the primary storage as a policy setting. When updated data that resides only on the primary storage exceeds the HWM, the VDE invokes a SYNC, and the data is copied from the primary storage to the secondary storage until its level in the primary storage reaches the LWM. The data is copied starting with the least recently used data.

3) Disk binding

When the host issues a request to access data in the virtual volume, the data is assigned to the primary storage. When data is updated, a SYNC is executed at the timing indicated by the SYNC policy settings. Normally, primary storage data that has not been updated or is already synchronized with the secondary storage becomes a deletion target. The VDE deletes this data, starting with the least recently used data. The disk bind feature enables specific virtual volumes to reside on the primary storage until they are resigned. The data on the virtual volume that is established in the disk bind group of virtual volumes will not be deleted by the VDE. Disk binding is useful for storing meta information about the file system and index file of a DBMS in order to ensure sufficient access performance.

4) Policy control

The VD800 supports the policies shown in **Table 2**. These policies can be specified using a dedicated Web-based tool called Virtual Disk Service Console.

## 3.2 Secondary storage

1) Tape media pool

The tapes installed in the secondary storage are registered to the tape pool and managed by the storage processor. When a virtual volume is created, the required number of tapes are selected from the tape pool and assigned to the virtual volume. For example, if 400 GB of tapes are registered and an 800 GB virtual volume is created, only one tape is initially selected and assigned. When the tape becomes full and more space is needed, another tape is assigned from the tape pool.

2) Tape mirroring

This feature duplicates the data on one tape onto another tape when a SYNC or MIGRATION

Table 2
Supported policies.

| Policy type˜ | Details˜ | Unit˜ |
|---|---|---|
| RECALL˜ | Specifies RECALL target virtual logical volume and the start time.˜ | Each VLU˜ |
| SYNC˜ | Specifies SYNC target virtual logical volume and the start time. ˜<br>Specifies elapsed time from the last access. ˜ | Each VLU˜ |
| Disk bind˜ | Specifies virtual logical volume to reside on primary storage.˜ | Each VLU˜ |
| HWM/LWM˜ | Specifies percentage of disk space on primary storage for SYNC start (HWM) ˜<br>and SYNC stop (LWM).˜ | Subsystem˜ |
| Tape global spare˜ | Specifies time period for executing tape global spare.˜ | Subsystem˜ |
| Tape garbage collection | Specifies time period for executing tape garbage collection. | Subsystem˜ |

˜

from the primary storage to secondary storage is performed. The duplication is controlled by the storage processor. If multiple ETERNUS LTO tape libraries are connected, the data is written almost simultaneously to a tape in another library. If only one library is connected, the data is duplicated within that library. The duplication process is completed when the data has been successfully written to both tapes. When a RECALL is issued for mirrored data, the data is read from whichever of the two tapes is available at that time. The user can freely specify which virtual logical volumes are to be mirrored. We recommend tape mirroring to ensure safe, long-term storage of data.

3) Tape global spare

The storage processor manages the tape access information, such as the number of mount/unmounts, for all the tapes assigned to the virtual logical volumes. When the number of mount/unmounts of a tape exceeds a specific number, the storage processor invokes data copying to a new tape registered in the tape pool. After the data duplication, the new tape replaces the original tape in the virtual logical volume and the original tape is registered as obsolete. The user can specify when the data copying is done as a policy so it does not affect regular daytime operations. When the data is being copied, the host can still access the corresponding virtual logical volume.

4) Tape rebuilding

If an uncorrectable error occurs on a tape that contains mirrored data, the data is copied to a new tape that is registered in the tape pool. After the data duplication, the new tape replaces the original tape in the virtual logical volume and the original tape is registered as obsolete. When the data is being copied, the host can still access the corresponding virtual logical volume.

5) Tape garbage collection

Because of the characteristics of LTO technology, updated data is added to a tape after the last data that was written to it. Consequently, as data is updated and deleted over time, the amount of wasted data space on the tapes increases, and the data must be reconfigured in order to eliminate wasted data space and acquire more usable data space. The storage processor determines when the amount of wasted space on a tape exceeds the specified value and invokes data copying to a new tape that is registered in the tape pool. Only the valid data is copied to the new tape. After the copying, the new tape replaces the original tape in the virtual logical volume and the original tape is registered as unused. When the data is being copied, the host can still access the corresponding virtual logical volume.

## 4. Conclusion

This paper outlined the ETERNUS Virtual Disk Library, which autonomously controls data according to its value. Hierarchical storage management is performed by the virtual disk engine of the disk array and the control software of the storage processors.

The amount of data in companies is rapidly increasing because of the maturity of the broadband network environment and advancement of information digitization. Preserving these large amounts of data for a prolonged period more efficiently requires storage systems that follow the concept of the information lifecycle (generation, utilization, reference, preservation, and deletion). Such storage systems store data in the appropriate storage according to its value and reduce the total storage cost. The ETERNUS Virtual Disk Library offers an appropriate platform base for information life cycle management.

**Shigeo Konno** received the B.E. degree in Information and Computer Sciences from Toyohashi University of Technology, Aichi, Japan in 1987. He joined Fujitsu Ltd., Kawasaki, Japan in 1987, where he has been developing high-end disk controllers and RAID subsystems.

**Tadashi Kumasawa** received the B.S. and M.S. degrees in Information Processing Engineering from Iwate University, Morioka, Japan in 1983 and 1985, respectively. He joined Fujitsu Ltd., Kawasaki, Japan in 1985, where he has been developing high-end disk controllers and RAID subsystems. He is a member of the Information Processing Society of Japan (IPSJ).