

# Technologies of ETERNUS6000 and ETERNUS3000 Mission-Critical Disk Arrays

● Yoshinori Terao

(Manuscript received December 12, 2005)

Fujitsu has developed the ETERNUS6000 and ETERNUS3000 disk arrays for mid-range and high-end IT market segments. These segments have been growing because IT systems have shifted from server-centric data processing architectures to information-centric architectures. This paper introduces the system architecture; high-reliability, high-performance, and high-availability features; and technologies of the ETERNUS6000/3000. Section 2 describes the controller modules, channel adapters, device adapters, routers, and device enclosure features. Section 3 describes the high-reliability features such as cyclic mirroring, block check codes, DB Guard, and redundant copying. Section 4 describes the high-availability RAID migration and logical device expansion features. Lastly, Section 5 describes the high-performance technologies of these RAID subsystems.

## 1. Introduction

Fujitsu introduced its first ETERNUS6000 and ETERNUS3000 disk arrays in 2003 and then improved their performance in 2004 and 2005. Recently, IT systems have been shifting paradigm from architectures focused on server-centric data processing to information-centric architectures. This shift is making data storage systems even more important in the storage, distribution, and utilization of information. The result is an ongoing process of storage system integration and networking that will continue into the foreseeable future. To meet the growing requirements for IT systems, Fujitsu provides the ETERNUS6000 storage systems for high-end, multi-platform environments and the ETERNUS3000 storage systems for mid-range IT systems. Both subsystems are designed to form part of an advanced mission-critical IT system.

## 2. System architectures

**Figures 1 and 2** show the architectures of

the ETERNUS6000 and ETERNUS3000.

The ETERNUS6000/3000 consist of a controller enclosure (CE) and two or more drive enclosures (DEs). The ETERNUS3000 can have up to 16 DEs, and the ETERNUS6000 can have up to 68.

### 2.1 Controller enclosure

#### 2.1.1 Controller modules of ETERNUS6000/3000

The controller modules (CMs) are the main components of the controller enclosure. Each CM has a processor and cache and controls the entire storage system. To achieve a suitable performance and reliability for high-end RAID systems, the ETERNUS6000 can have up to four CMs.

#### 2.1.2 Channel adapters of ETERNUS6000/3000

The channel adapters (CAs) provide connections between the ETERNUS6000/3000 and servers. The ETERNUS6000 supports a wide

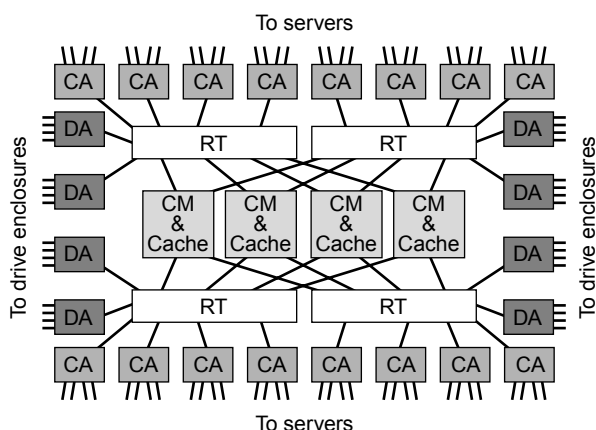


Figure 1  
ETERNUS6000 architecture.

variety of host interfaces. Its CAs are available in five types: 4 Gb/s Fibre Channel, 1 Gb/s iSCSI, FCLINK, OCLINK, and BMC. The CA types can be freely combined, and up to 16 CAs can be installed. When 4-port Fibre Channel CAs are used, up to 64 ports are available for connection to the host.

Each CA of the ETERNUS6000 has a micro-processor to distribute the I/O workload from servers between the CMs and CAs.

The ETERNUS3000 supports 4 Gb/s Fibre Channel and 1 Gb/s iSCSI CAs. These CAs have no microprocessors and are controlled directly by the CMs.

### 2.1.3 Device adapters of ETERNUS6000/3000

The device adapters (DAs) are modules that provide connections between the CMs and hard disk drives (HDDs). Each DA has a 2 Gb/s Fibre Channel interface.

Each DA of the ETERNUS6000 has a micro-processor to distribute the many HDD operations between the CMs and DAs. The ETERNUS6000 can have up to eight DAs.

Each CM of the ETERNUS3000 has a DA controller LSI. These LSIs have no microprocessor and are controlled directly by the CMs.

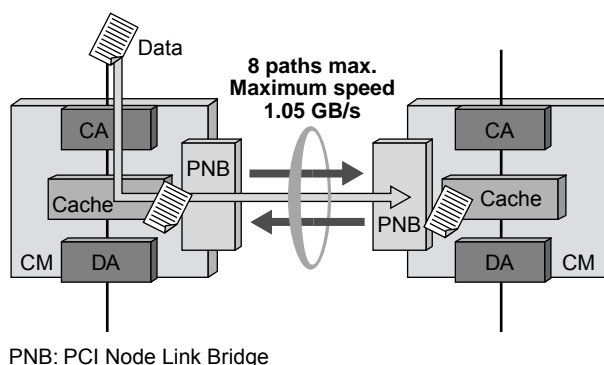


Figure 2  
ETERNUS3000 architecture.

### 2.1.4 Routers of ETERNUS6000

The routers (RTs) of the ETERNUS6000 organically interconnect the CMs, CAs, and DAs and provide high-speed communication between them. To achieve system redundancy, each RT is connected to each CM. Up to four routers can be installed.

### 2.2 Drive enclosures of ETERNUS6000/3000

The DEs can hold up to 15 HDDs plus two backend FC connectivity cards. They also have dual power supply units and fans.

## 3. High-reliability features

### 3.1 Component redundancy and hot swap of ETERNUS6000/3000

The most important requirement in integrated storage systems is robust reliability that ensures the safe use of stored data. The ETERNUS6000/3000 meet these requirements with full-component redundancy, hot-swappable hardware/firmware, and a wide variety of high-reliability and high-availability features.

All major components included in the controller and drive enclosures are redundant as standard to ensure continuous operation even when one of them fails. Moreover, these major components are all hot-swappable. This enables on-line recovery from component failures to ensure ongoing redundancy and system continuity.

### 3.2 Cyclic mirroring of ETERNUS6000 CMs

The ETERNUS6000 uses a high-reliability controller system called Cyclic Mirroring, which becomes operational when three or more CMs are installed. This system minimizes the impact on performance and redundancy when a CM or a CM's cache fails. When four CMs are installed as shown in **Figure 3**, cache write data is mirrored on two CMs in case one of them fails. If one of the CMs fails, the Cyclic Controller Module system takes it offline and reconfigures the remaining three so that all write data is mirrored among them. Data is assigned cyclically across the three CMs in a round-robin approach that does not rely on fixed pairs. This means that if a CM fails, the overall system performance is only degraded by one-quarter or less.

### 3.3 Enhanced reliability RAID-5 of ETERNUS6000

When a failure occurs in a conventional RAID-5 configuration consisting of HDDs in a drive enclosure, the entire drive enclosure becomes unusable and the data cannot be accessed. However, in the ETERNUS6000, the HDDs of a RAID-5 group are located in different drive enclosures to ensure robust RAID-5 reliability. As shown in **Figure 4**, four interfaces belonging to

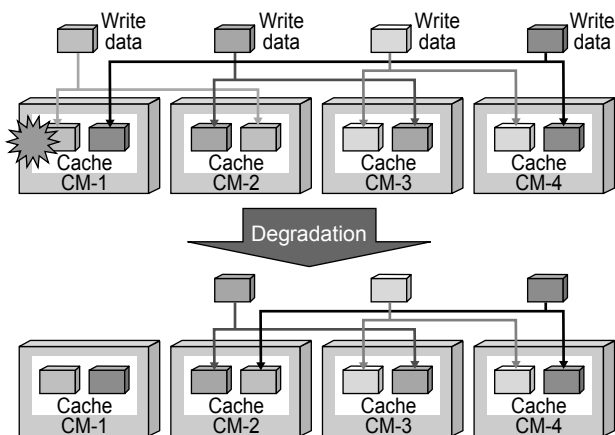
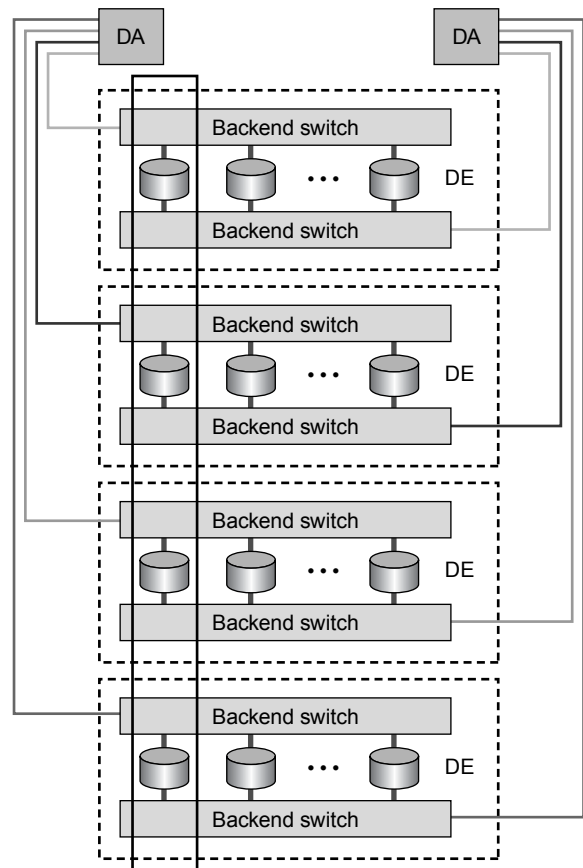


Figure 3 Cyclically mirrored CMs.

one DA are connected to four different drive enclosures. The ETERNUS6000 forms a RAID-5 group using four selected HDDs from the four drive enclosures, which means the RAID-5 configuration is always fixed at 3+1. However, the ETERNUS6000 can keep the RAID-5 group accessible even when an entire drive enclosure becomes unusable, while limiting the impact of the failure to a single HDD. This enables the RAID's data restoration feature to be used.

### 3.4 Block check code of ETERNUS6000/3000

When receiving data from a server, the ETERNUS6000/3000 generate a proprietary block checking code (BCC) from calculations made on the data block. **Figure 5** shows the BCC process of the ETERNUS6000. In this process, the BCC



DE: Drive enclosure

Figure 4 RAID-5 with enhanced reliability.

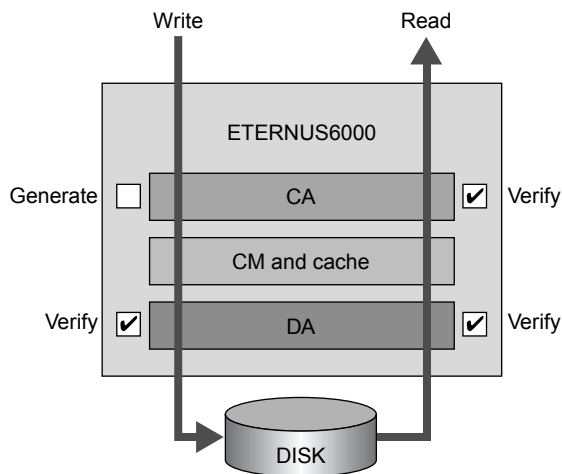


Figure 5  
Block check code feature of ETERNUS6000.

component is recorded together with the data in cache memory and on the HDDs. For writing, the BCC is used for data integrity verification when data is sent from the controller to the HDDs. For reading, the BCC is used to verify data integrity when data is sent from HDDs to the controller and when data is sent from the controller to a server. Although the BCC does not provide a data correction feature such as ECC, BCC can verify the location and order of data; for example, it can verify whether data stripped across multiple HDDs has been rebuilt in the correct order.

### 3.5 DB Data Guard of ETERNUS6000

BCC is not perfect because it does not cover data corruption that occurs between the storage system and server. This is because the ETERNUS6000 will only generate BCCs based on the corrupted data. This is a critical issue, especially for databases because data corruption can seriously impact the integrity of a database. To solve this issue, we developed DB Data Guard, which is a data assurance feature that combines the storage system's hardware and database software. When writing data to a storage system, the database generates check codes and adds them to the data (**Figure 6**). The storage system and the database share information, including the check

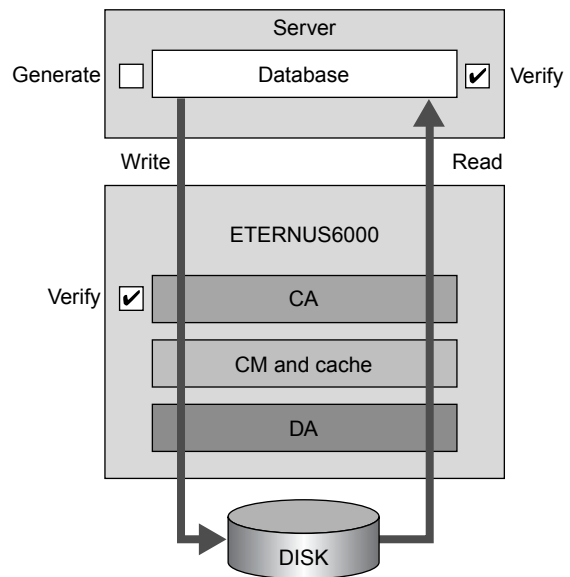


Figure 6  
DB Data Guard.

code generation logic and the positions where the check codes are added. The ETERNUS6000 has a special ASIC in its CAs that calculates and verifies this check code. If a CA detects that data received from the server is corrupted, it reports the detection and discards the data so it does not overwrite any data in the CM's cache. This prevents the use of corrupted data and protects the integrity of the database.

### 3.6 Redundant Copy feature of ETERNUS6000/3000

The ETRNUS6000 and ETERNUS3000 constantly monitor all HDDs. If an HDD generates errors at more than a specified rate, the controller judges that the HDD is about to fail. In this case, the HDD is swapped using the Redundant Copy feature (**Figure 7**). This feature maintains data redundancy even during HDD replacement and minimizes the data restoration time. In RAID configurations, when an HDD fails, data is restored from the other HDDs to prevent data loss. Although data redundancy cannot be maintained during data restoration (even though it only takes a short time), the Redundant Copy feature does not isolate the HDD in pre-failure status until it

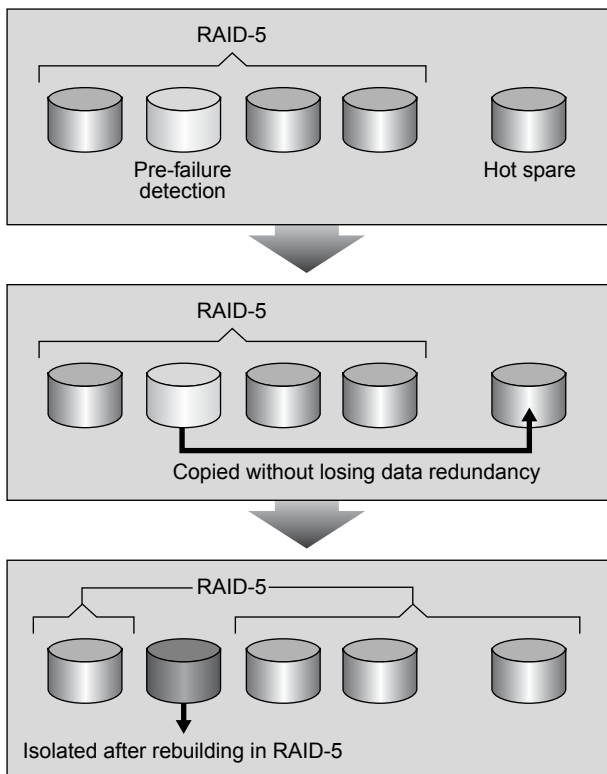
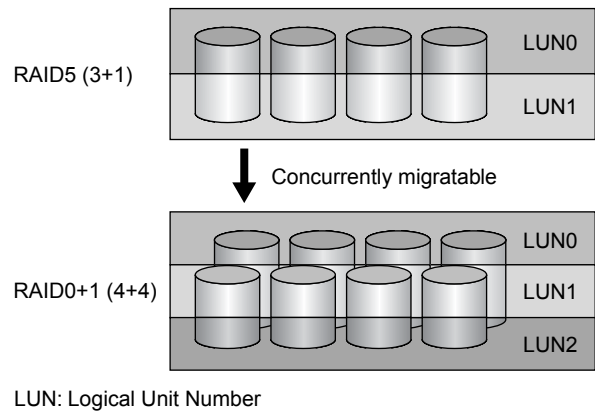


Figure 7  
Redundant Copy.

has been copied to a hot-spare HDD. This ensures data redundancy even during HDD replacement. If an HDD fails in a conventional RAID-5 configuration, its data is rebuilt from the other HDDs before it is copied to a hot-spare HDD. However, Redundant Copy just copies the pre-failure HDD directly to a hot-spare HDD, which reduces the time needed to replace an HDD and the associated system load.

#### 4. High-availability features

The ETERNUS6000/3000 include a wide variety of new features that enable flexible system reconfiguration and capacity expansion. Conventionally, a capacity expansion of a RAID group requires a cumbersome procedure: the operator must save the data, add a new disk to the RAID group, redefine the RAID configuration, and finally rewrite the data back to the revised disk configuration. Moreover, the saved data cannot be accessed during this procedure.



LUN: Logical Unit Number

Figure 8  
RAID migration.

#### 4.1 RAID migration of ETERNUS6000/3000

In some systems, RAID migration may need to be done to a differently configured RAID (**Figure 8**). For example, this may be necessary for migration from a RAID-5 group to a RAID-1 group or from a 10 000 rpm HDD RAID group to a 15 000 rpm HDD RAID group in order to improve performance. Conventionally, such operations interrupt system operation and require a cumbersome procedure including data backup, configuration of the new RAID area, and data restoration to the new RAID group. The ETERNUS6000 provides a RAID Migration feature that performs all of these procedures without interrupting system operation. Moreover if the destination RAID has free space, it can concurrently define a new Logical Unit Number (LUN).

#### 4.2 Logical device expansion of ETERNUS3000

Logical device expansion is a function that automatically performs the complex processing required to expand the RAID group (**Figure 9**). By using this function, the existing data is automatically relocated, and a new LUN can be defined using the newly generated space. The figure shows an example of adding a new 73 GB disk to a RAID5 (3+1) group to make a new RAID5 (4+1) group.

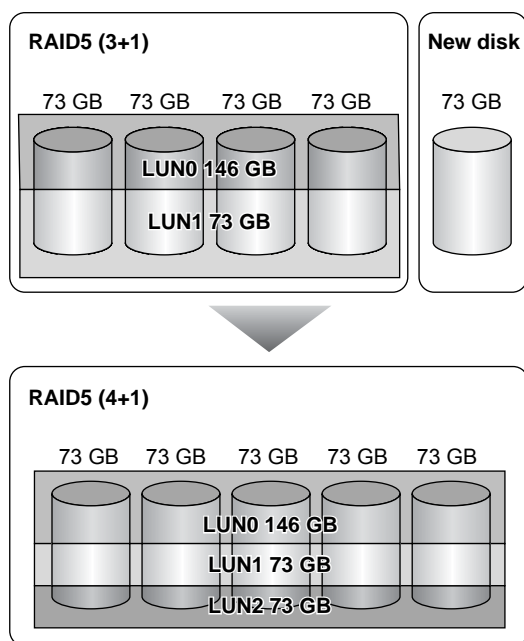


Figure 9  
Logical device expansion.

## 5. High-performance features

The ETERNUS6000/3000 achieve high performance in their RAID subsystems because their CMs employ high-speed CPUs. Also, the CMs are connected to fast Interconnect technology and fast HDDs to achieve high-speed data movement.

### 5.1 ETERNUS6000 technology

#### 5.1.1 Processors

The ETERNUS6000 uses high-performance processors in each of its CMs, CAs, and DAs and supports a maximum cache of 32 GB. The CMs, for example, have leading-edge 3.2 GHz processors.

The CAs and DAs also have high-performance RISC processors to control the interface hardware (i.e., the Fibre Channel controller).

#### 5.1.2 High speed, multi-channel interconnection

The ETERNUS6000 features ASIC RTs that were developed using Fujitsu's original technology. The RTs provide faster interconnects between the CMs and CAs/DAs. Each RT has a total band-

width of 4.2 GB/s. To improve data duplication from the CAs to CMs, the RTs duplicate the data they receive and send it to two CMs at the same time (data forking). Generally, in RAID subsystems, the write throughput performance is less than that for read operations. The data forking feature prevents a performance reduction during write operations. The ETERNUS6000 can simultaneously transfer up to 96 data blocks between the adapters and the cache, which guarantees highly autonomous operation of the system's processors.

### 5.2 ETERNUS3000 technology

#### 5.2.1 Processors

The ETERNUS3000 provides unparalleled performance by using cache sizes of up to 16 GB and using 1.26, 2.8, or 3.2 GHz processors.

#### 5.2.2 DTC

In addition to the basic hardware advantages described above, the ETERNUS3000 achieves even faster and more stable performance by using innovative Fujitsu technologies such as Fujitsu's unique Dynamic Disk Traffic Control (DTC) feature (**Figure 10**). This feature optimizes HDD access on ETERNUS3000 storage systems based on the method of access being used at the time. This ensures a quick response to individual server requests and minimum processing times for background serial access operations such as backup and recovery. When both sequential and random accesses occur simultaneously, for example, when data is being restored while random requests are still being processed from the server, they are separated into optimum extents and processed alternately on a time interval basis. This ensures that the optimum access methods are used, even during high-traffic periods.

#### 5.2.3 CM interconnection

The ETERNUS3000 also makes use of a Fujitsu technology innovation called the PCI Node Link Bridge (PNB) (**Figure 2**), which provides

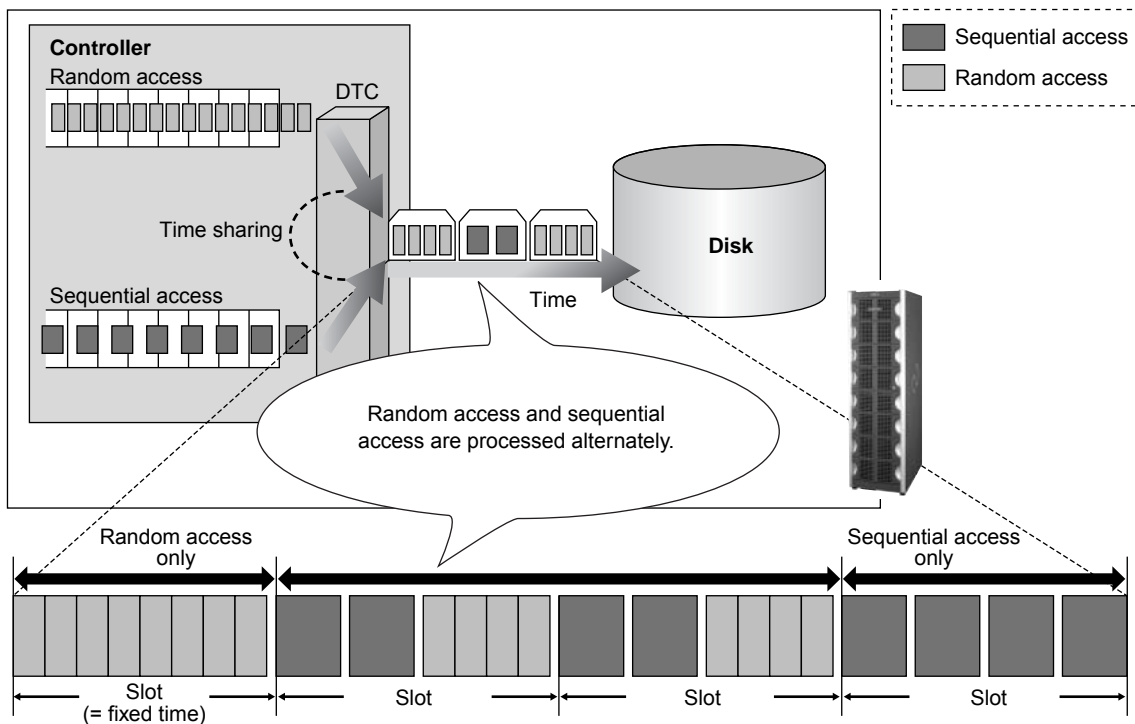


Figure 10  
Dynamic Disk Traffic Control (DTC).

faster inter-cache data duplication, resulting in improved update processing. When a controller receives a write request from a server, the controller uses PNB hardware to duplicate the data in another controller without data access interruption. This method reduces the overhead that is usually caused by inter-controller communication for such data duplication. PNB supports data transfer rates of up to 1.05 GB/s and up to 8 inter-controller communication paths for ultra-fast cache duplication.

## 6. Conclusion

This paper described Fujitsu's ETERNUS6000 and ETERNUS3000 disk arrays, which achieve high performance and high availability while giving priority to reliability.

These two subsystems use common technologies such as hot swapping, BCC, and redundant copying to improve their reliability. The ETERNUS6000 uses RTs, CAs with processors, DB Data Guard, and RAID migration to provide

the performance and reliability required in high-end products. The ETERNUS3000 achieves high performance by using DTC and PNB interconnection.

In the future, customers will request RAID subsystems with a wider variety of configurations to meet their individual IT needs. Therefore, a theme for future development at Fujitsu will be the fusion of mid-range and high-end products while maintaining scalability.



**Yoshinori Terao** received the B.S. degree in Applied Physics from Okayama University of Science, Okayama, Japan in 1989. He joined Fujitsu Ltd., Kawasaki, Japan in 1989, where he has been engaged in development of storage system products.