Remote Copy Technology of ETERNUS6000 and ETERNUS3000 Disk Arrays

● Tsutomu Akasaka

(Manuscript received July 5, 2005)

This paper gives an overview of a storage-system remote copy function and the implementation of such a function in Fujitsu's ETERNUS3000 and ETERNUS6000 disk arrays. Recently, disaster recovery features have become important for mission-critical systems. A storage-system remote copy function is often used by high-end disaster recovery systems. The ETERNUS3000 and ETERNUS6000 series both have a remote copy function called Remote Equivalent Copy (REC). The ETERNUS6000 series REC operates in both synchronous and asynchronous mode. The ETERNUS3000 series REC operates only in synchronous mode. Both RECs provide the performance and functionalities required to construct various kinds of disaster recovery systems.

1. Introduction

Recently, disaster recovery has become much more important. It has been recognized that without disaster recovery systems, critical data might be lost during a disaster, which might cause big problems for company activities. Also, the costs of constructing and operating disaster recovery systems has been decreasing, especially the cost of communicating with remote sites.

A remote mirroring function for storage devices is one of the key components of high-end disaster recovery systems. It can transfer data on a storage system from a primary site to a secondary site. It can also achieve high-bandwidth data transfer and transfer data without using host CPU system resources.

This paper describes a mechanism for a remote copy function and how it is used for disaster recovery. It also describes the remote copy function of Fujitsu's ETERNUS3000 and ETERNUS6000 disk arrays. These disk arrays offer a remote copy function called Remote Equivalent Copy (REC) that can operate in various modes for use in disaster recovery systems.

2. Overview of remote copy function

Most remote copy functions for storage systems use a mirroring method. The storage boxes (storage devices of ETERNUS3000 and ETERNUS6000) used for copy operations are connected by cables running a standard protocol such as Fibre Channel. A write from the local host first reaches the local storage boxes, which then forward the write to remote storage boxes (**Figure 1**). The copy-on-write technique[1),2)] is another method of implementing remote copying. However, this paper does not describe this technique because most remote copying is implemented using mirroring.

For remote copying, data transfer can be done synchronously or asynchronously. When a write is issued to local storage boxes, it propagates to remote storages boxes either synchronously or asynchronously.

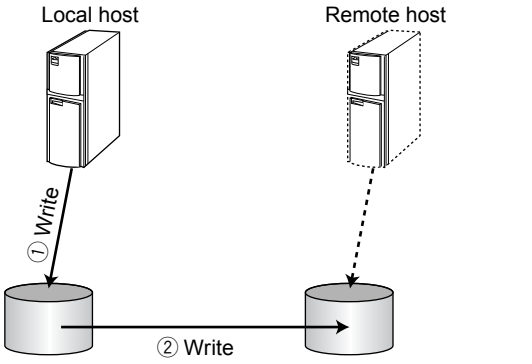

Figure 1
Write operation of remote copy.

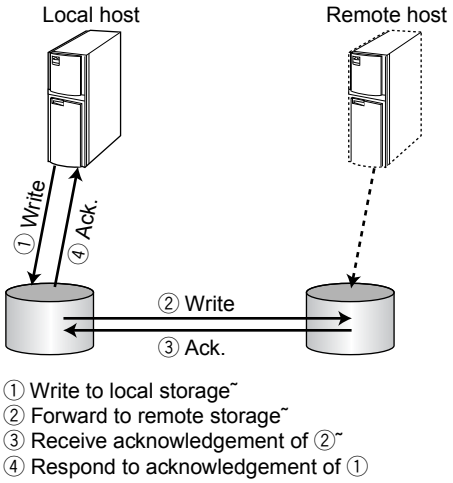

Figure 2
Write operation in synchronous mode.

2.1 Synchronous mode

A write from a local host is not acknowledged until the write has been completed at both the main and remote sites (**Figure 2**). A write that is acknowledged to the application that issued the write request is always guaranteed to reach the remote site. However, the write response time may be degraded because acknowledgement of the write has to wait until the data has propagated to the remote site. Synchronous propagation over long distances may cause a serious performance degradation.

2.2 Asynchronous mode

A write from a local host may be acknowledged before the write has been completed on a remote site (**Figure 3**). Write acknowledgement does not wait for remote site access, so the write response time is not affected by transmission delays. However, write data that is acknowledged to the application might not reach the remote site, in which case data will be lost.

3. Disaster recovery with storage system based copy

There are two types of remote copy functions for a disaster recovery storage system. One is remote backup, which makes a point-in-time copy of a working system at a primary site[note 1)] to a secondary site[note 2)] by using a remote copy function. When a disaster occurs, the backup system can restart from the point the snapshot was taken, but data that was generated after the snapshot is lost. The typical recovery point objective (RPO)[note 3)] of remote backups is 12 to 24 hours. The other type is remote mirroring, which always keeps consistent copies of primary volumes[note 4)] in secondary volumes.[note 5)] A consistent copy means a copy that enables software to restart and continue processing. Remote mirroring is not always applicable because whether the data is restartable often depends on the software. Remote mirroring is focused on database mirroring because for most mission-critical systems, the most important data is stored as a database.

note 1) Primary site refers to the main site of a production system.
note 2) Secondary site refers to the remote site of the backup system.
note 3) RPO (Recovery Point Objective) refers to the point from which the backup system can restart after failover.
note 4) Primary volume refers to the source volume from which data is copied by a remote copy function.
note 5) Secondary volume refers to the destination volume to which data is copied by a remote copy function.

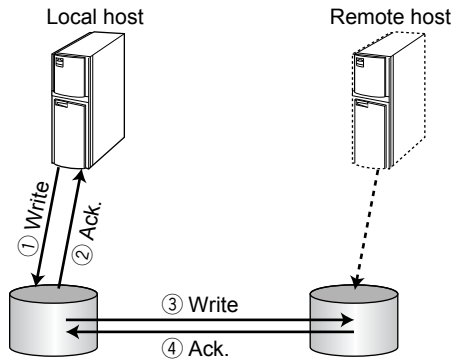

Figure 3
Write operation in asynchronous mode.

3.1 Remote mirroring

For remote mirroring, the remote copy function requires the following features:

1) Preserving the write order

Data must be written to the secondary box (storage box at secondary site) in the same order it was written to the primary box (storage box at primary site) so the correct write order is preserved.

2) Automatic mirror breaking

When a remote mirroring automatically breaks a mirror, propagation of write data must be stopped at the same point in time across all mirrored volumes.

If a remote copy function has these two requirements, it can be used for a database mirroring system. There are two main ways to use a remote copy function for database mirroring: whole database mirroring and database log mirroring.

Whole database mirroring is a simple method. It uses a storage-system remote copy function to copy an image of a whole database from the primary box to the secondary box. When a disaster subsequently occurs, the database can be restarted with the mirrored data by crash recovery.[note 6)] This method does not require a host system at the secondary site before failover occurs. However, it requires a broad-bandwidth communication path to transfer a whole database, which is sometimes very expensive.

Database log mirroring mirrors only database logs from the primary box to the secondary box. The host system should update the database in the secondary site by referencing the mirrored database logs. Therefore, this method requires a host system at the secondary site, even when failover has not occurred. However, the required bandwidth of the communication path may be much less than with the whole database mirroring method.

When the mirrored data is not a database, the data is not always meaningful, even if the remote copy function satisfies the two requirements described above.

3.2 Remote backup

For a remote backup, compared to remote mirroring, there are fewer requirements for the remote copy function because it only requires data consistency when mirrors are split. Remote backup involves the following sequence:

1) Keep primary volumes consistent until the backup is finished.
2) Split the mirror. Then, keep consistent backup data in the secondary box.
3) After splitting the mirror, use of the primary volumes can be restarted.

After a backup operation, the remote copy function receives a command from the system to split the mirror. After the command is received, data writing to the primary volume must not propagate to the secondary volume. However, data writing before the command is received must propagate to the secondary volume.

3.3 Requirement of remote copying for disaster recovery

The choice of method for a disaster recovery system depends on the following requirements.

note 6) Crash recovery refers to database recovery from a server crash.

1) Performance

Remote mirroring always transfers data, and it may cause an I/O performance degradation. The remote backup method might even stop transferring data during periods of high communication traffic.

2) Expense

Mirroring a whole set of data requires high bandwidth in the connecting path, which may be very expensive.

3) Recovery Time Objective (RTO)

This depends on what data is mirrored or backed up and when it is backed up. Mirroring a whole set of data may lead to faster recovery than other methods.

4) Recovery Point Objective (RPO)

When remote mirroring is used, the RPO depends on whether synchronous or asynchronous data transfer is used. When a disaster occurs, the amount of data that is lost with the remote backup method is higher than with remote mirroring and can be very large depending on the backup interval.

5) Simplicity

The procedure for mirroring a whole set of data is simpler than that for other methods. Remote backup requires an operation to periodically take a point-in-time copy. To mirror logs, logs must be copied to the database at the secondary site.

4. Remote copy on ETERNUS3000 and ETERNUS6000 series

The ETERNUS3000 and ETERNUS6000 series are Fujitsu disk array subsystems. The ETERNUS6000 series are high-end subsystems, and the ETERNUS3000 series are mid-range subsystems. Both series have a remote copy function called REC, which has various operation modes and can be used in disaster recovery systems. REC uses mirroring and the following commands:

1) Start

Makes a mirrored primary and secondary volume and starts synchronization.

2) Stop

Splits a mirror and discards the mirrored pair. All the data written to the primary volume before this command is issued must be transferred to the secondary volume until the command is finished.

3) Suspend

Splits a mirror but keeps the mirrored pair, which can then be resynchronized by the resume command. All the data written to the primary volume before this command is issued must be transferred to the secondary volume until the command is finished.

4) Forced stop

Splits a mirror and discards the mirrored pair. This command might not transfer write data that is written to the primary volume before the command is issued.

5) Forced suspend

Splits a mirror but keeps the mirrored pair, which can be resynchronized by the resume command. This command might not transfer write data that is written to the primary volume before the command is issued.

6) Resume

Starts resynchronization after the suspend command. Resynchronization is done by transferring the data written during the suspend state.

These commands are issued to the copy session, which is the copy operation specified by the primary and secondary volumes. Both the primary and secondary volumes are either a Logical Unit Number (LUN) or continuous extents of a LUN.

The state transitions that can occur during an REC copy session are shown in **Figure 4**.

The ETERNUS6000 and ETERNUS3000 series of disk array subsystems are connected by 2 Gb/s Fibre Channel cables. For long-distance connections, Fibre Channel can be converted to another protocol suitable for long-distance data transfer (e.g., IP or SONET) by another box. The ETERNUS3000 and ETERNUS6000 series cannot make such a conversion.

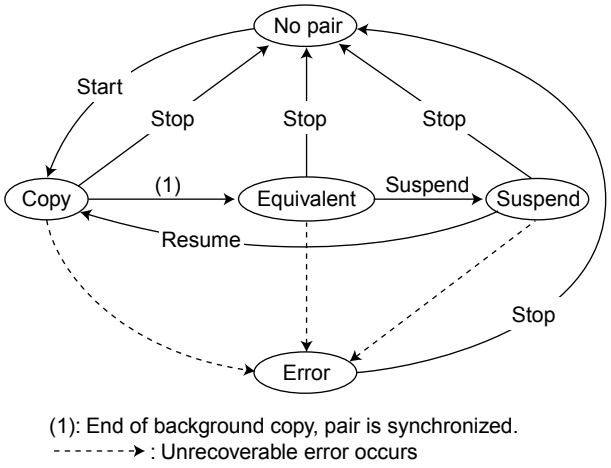

Figure 4
State transition of REC.

4.1 ETERNUS6000 series

The ETERNUS6000 REC can operate in synchronous and asynchronous mode to satisfy the system requirements.

4.1.1 Synchronous mode

The ETERNUS6000 synchronous mode is used for remote mirroring. It degrades system performance more than asynchronous mode. However, when a mirror is broken by a disaster or other event, all the write data that an application has acknowledged at the primary site must be propagated to the secondary site.

To avoid degrading the write performance, we recommend that the distance between the primary and secondary sites is less than 100 km. This distance depends on the application that runs on the host server system and the network that connects the two sites.

For a database, transactions executed at the primary site are guaranteed to propagate to the secondary site. There are two types of synchronous mode: automatic split mode and manual split mode. In automatic split mode, if write data cannot propagate to the secondary box, the primary and secondary volumes are automatically split and the primary volume can then be accessed. In the manual split mode, the primary and secondary volumes are also automatically split, but the primary volume cannot be accessed after the split (**Table 1**). Whether REC is in synchronous or manual split mode, the primary and secondary volumes are always synchronized, even if data cannot be transferred to the secondary site. When REC is in automatic split mode, the primary volume can be accessed even if data cannot be transferred. However, if a disaster occurs during network trouble, the data that cannot be transferred to the secondary volume will be lost in automatic split mode.

Table1
Trade off between automatic and manual split.

Mode	Data lost note	Availability of primary
Automatic split	Possible	Available when secondary is unavailable.
Manual split	No data loss.	Unavailable when secondary is unavailable.

note: This column shows possibility of data loss if disaster occurs while secondary volume is unavailable.

4.1.2 Asynchronous mode

There are two types of asynchronous mode: consistency mode and stack mode. Consistency mode is for remote mirroring; it can preserve the write order even when there are multiple sessions. Stack mode is for remote backup; it cannot preserve the write order, but it causes less performance degradation than consistency mode.

1) Consistency mode

Consistency mode is intended for use with remote mirroring. It preserves the write order and breaks a mirror automatically. To implement consistency mode, the ETERNUS6000 series use a part of their cache memory as a data transfer buffer. Data to be written to the primary volume is copied to this buffer, which is called the REC buffer, and then transferred to the secondary box. In the secondary box, the transferred data is sent to the REC buffer and then copied to the secondary volumes.

The primary and secondary boxes can have multiple REC buffers. The REC buffer that copies the write data is switched to another one at checkpoints that the hardware periodically

generates (**Figure 5**).

When the write data is copied to the REC buffer, the LUN of the secondary volume and location of the data are also written to the REC buffer. This information is used to specify where the data should be stored when it is copied from the REC buffer in the secondary box to the secondary volumes.

Data transfer from the REC buffer in the primary box to the REC buffer in the secondary box is performed automatically. When all of the buffer contents are transferred successfully, the data received in the REC buffer of the secondary box is copied to the secondary volumes. If data transfer is stopped by a disaster, network trouble, or other cause before all of the buffer contents are transferred, the contents of the REC buffer in the secondary box are discarded to keep the contents of the secondary volumes consistent (**Figure 6**).

The interval between checkpoint generations can be selected by the user. If the REC buffer becomes full before the next checkpoint, a new checkpoint is generated before the interval is finished and the incoming write data is sent to a new REC buffer.

We measured the overhead for generating checkpoints. The generation halts all I/O operations to synchronize the processors of the controllers (ETERNUS6000 can have up to four controllers). The I/O operation halts are usually very short. We estimated the overhead for switching the REC buffers to be less than 0.1%, even when the checkpoint generation interval is short. The interval should be as short as possible to minimize the amount of data that is lost when a disaster occurs and also minimize the required REC buffer size.

2) Stack mode

Stack mode is intended for use with remote backup. This mode cannot preserve the write order, so it cannot be used for remote mirroring. However, this mode usually degrades the write response time less than consistency mode. Stack mode works as follows:

When a write command is issued to the primary volumes, the controller tracks the location that is updated.

Another process, called the transfer engine, transfers the updated data from the primary vol-

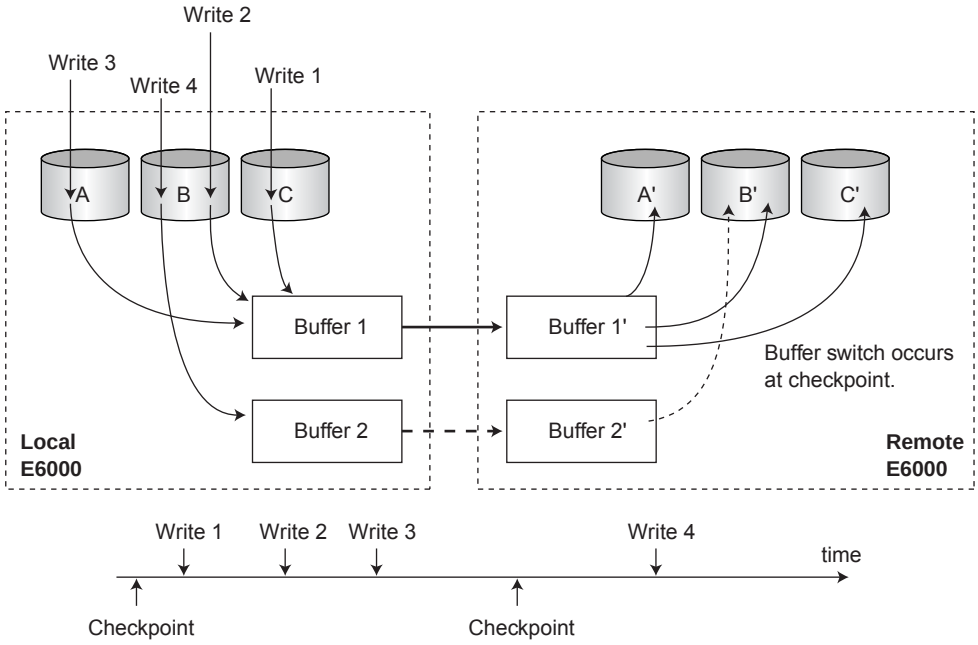

Figure 5
Data transfer operation in consistency mode.

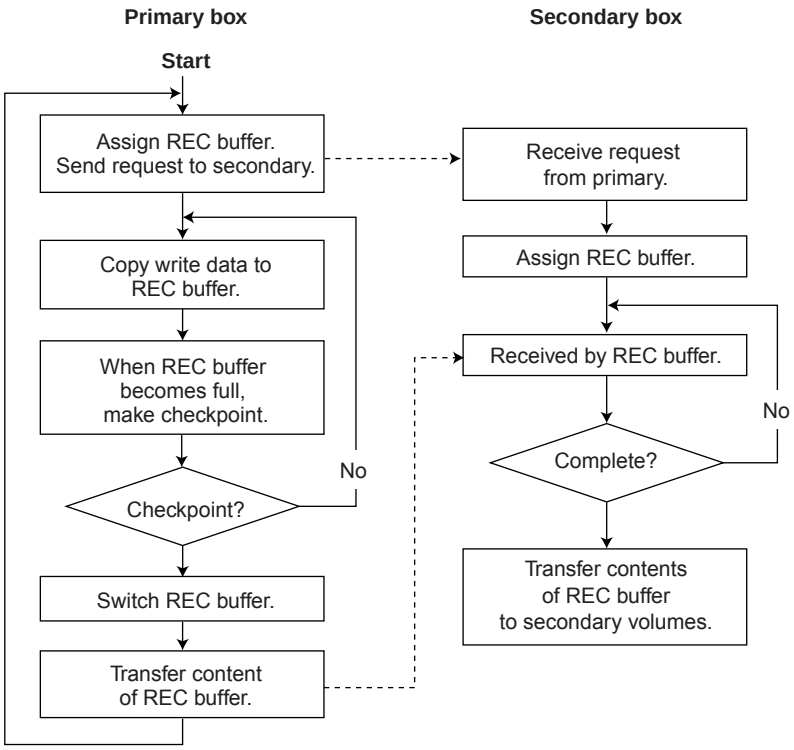

Figure 6
Consistency mode flowchart.

umes to the secondary box.

3) Performance

We estimated the write operation overhead of REC consistency mode to be less than 0.1 ms per write command. The overhead of stack mode is less than that of consistency mode. These overheads are independent of the transmission delay.

The data throughput from the primary box to the secondary box depends on the configuration of the primary box and exceeds 200 MB/s, even for the entry model of the ETERNUS6000 series.

4.2 ETERNUS3000

The ETERNUS3000 REC operates only in synchronous mode. It can use remote mirroring and remote backup; however, it is difficult to use the ETERNUS3000 REC for long-distance remote mirroring or remote backup because it does not support asynchronous mode.

Like the ETERNUS6000 series, the ETERNUS3000 series can operate in automatic split mode or manual split mode.

5. Conclusion

The ETERNUS6000 and ETERNUS3000 series provide a remote copy function called REC. The ETERNUS6000 REC can operate in synchronous and asynchronous mode, and the ETERNUS3000 REC operates only in synchronous mode.

The consistency mode of asynchronous mode REC can always preserve the write order and breaks a mirror automatically when a disaster occurs. In this mode, REC can be used in a remote mirroring disaster recovery system. The performance degradation in consistency mode is very small and is independent of the transmission delay. This degradation might cause data loss, but the amount of loss is usually less than several seconds. Stack mode is intended for use with remote backup; it cannot preserve the write

order, but it provides better performance.

The performance degradation of synchronous mode REC depends on the transmission delay; however, a zero data loss disaster recovery system can be constructed using synchronous mode.

References

1) N. Osorio and B. Lee: Guidelines for Using Snapshot Storage Systems for Oracle Databases. October 2001.
http://www.oracle.com/technology/deploy/availability/pdf/oscp_snapshot_use.pdf
2) B. Lee: Guidelines for Using Remote Mirroring Storage Systems for Oracle Databases. November 1999.
http://www.oracle.com/technology/deploy/availability/pdf/oscp_remote_mirror_use.pdf

Tsutomu Akasaka received the B.S. degree in Mathematical Engineering from the University of Tokyo, Tokyo, Japan in 1984. He joined Fujitsu Ltd., Kawasaki, Japan in 1984, where he has been developing Global Server products. He is currently developing storage system products.