Chapter 14
Remote Replication

Remote replication is the process of creating replicas of information assets at remote sites (locations). Remote replicas help organizations mitigate the risks associated with regionally driven outages resulting from natural or human-made disasters. Similar to local replicas, they can also be used for other business operations.

KEY CONCEPTS

- Synchronous and Asynchronous Replication
- LVM-Based Replication
- Host-Based Log Shipping
- Disk-Buffered Replication
- Three-Site Replication
- Data Consistency

The infrastructure on which information assets are stored at the primary site is called the *source*. The infrastructure on which the replica is stored at the remote site is referred to as the *target*. Hosts that access the source or target are referred to as *source hosts* or *target hosts,* respectively. This chapter discusses various remote replication technologies, along with the key steps to plan and design appropriate remote replication solutions. In addition, this chapter describes network requirements and management considerations in the remote replication process.

14.1 Modes of Remote Replication

The two basic modes of remote replication are synchronous and asynchronous. In *synchronous remote replication*, writes must be committed to the source and the target, prior to acknowledging "write complete" to the host (see Figure 14-1). Additional writes on the source cannot occur until each preceding write has been completed and acknowledged. This ensures that data is identical on the source and the replica at all times. Further writes are transmitted to the remote

site exactly in the order in which they are received at the source. Hence, write ordering is maintained. In the event of a failure of the source site, synchronous remote replication provides zero or near-zero RPO, as well as the lowest RTO.

However, application response time is increased with any synchronous remote replication. The degree of the impact on the response time depends on the distance between sites, available bandwidth, and the network connectivity infrastructure. The distances over which synchronous replication can be deployed depend on the application's ability to tolerate extension in response time. Typically, it is deployed for distances less than 200 KM (125 miles) between the two sites.

In *asynchronous remote replication*, a write is committed to the source and immediately acknowledged to the host. Data is buffered at the source and transmitted to the remote site later (see Figure 14-2). This eliminates the impact to the application's response time. Data at the remote site will be behind the source by at least the size of the buffer. Hence, asynchronous remote replication provides a finite (nonzero) RPO disaster recovery solution. RPO depends on the size of the buffer, available network bandwidth, and the write workload to the source. There is no impact on application response time, as the writes are acknowledged immediately to the source host. This enables deployment of asynchronous replication over extended distances. Asynchronous remote replication can be deployed over distances ranging from several hundred to several thousand kilometers between two sites.

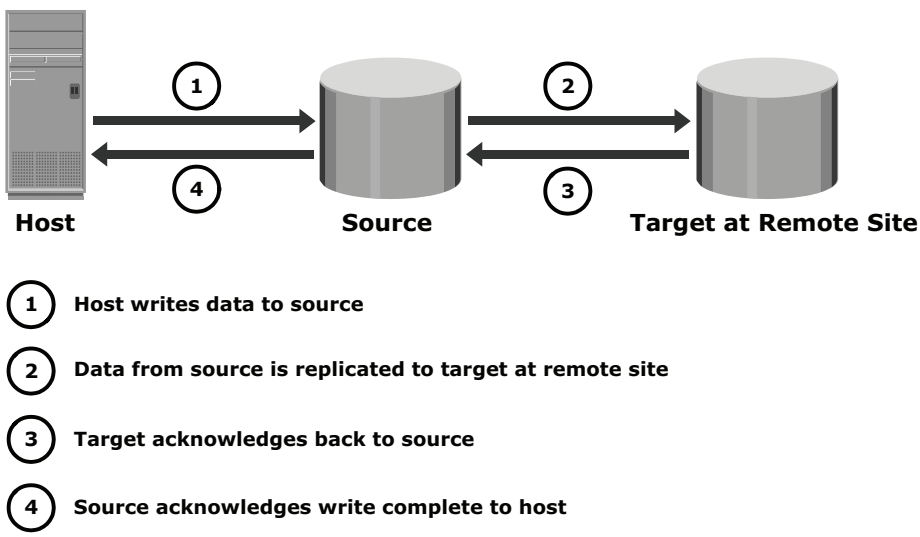
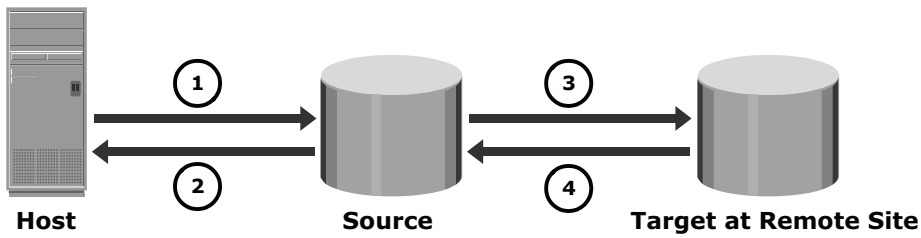


Figure 14-1: Synchronous replication

14.2 Remote Replication Technologies

Remote replication of data can be handled by the hosts or by the storage arrays. Other options include specialized appliances to replicate data over the LAN or the SAN, as well as replication between storage arrays over the SAN.



1. Host writes data to source
2. Write is immediately acknowledged to host
3. Data is transmitted to the target at remote site later
4. Target acknowledges back to source

Figure 14-2: Asynchronous replication

14.2.1. Host-Based Remote Replication

Host-based remote replication uses one or more components of the host to perform and manage the replication operation. There are two basic approaches to host-based remote replication: *LVM-based replication* and *database replication via log shipping*.

LVM-Based Remote Replication

LVM-based replication is performed and managed at the volume group level. Writes to the source volumes are transmitted to the remote host by the LVM. The LVM on the remote host receives the writes and commits them to the remote volume group.

Prior to the start of replication, identical volume groups, logical volumes, and file systems are created at the source and target sites. Initial synchronization of data between the source and the replica can be performed in a number of ways. One method is to backup the source data to tape and restore the data to the remote replica. Alternatively, it can be performed by replicating over the

IP network. Until completion of initial synchronization, production work on the source volumes is typically halted. After initial synchronization, production work can be started on the source volumes and replication of data can be performed over an existing standard IP network (see Figure 14-3).

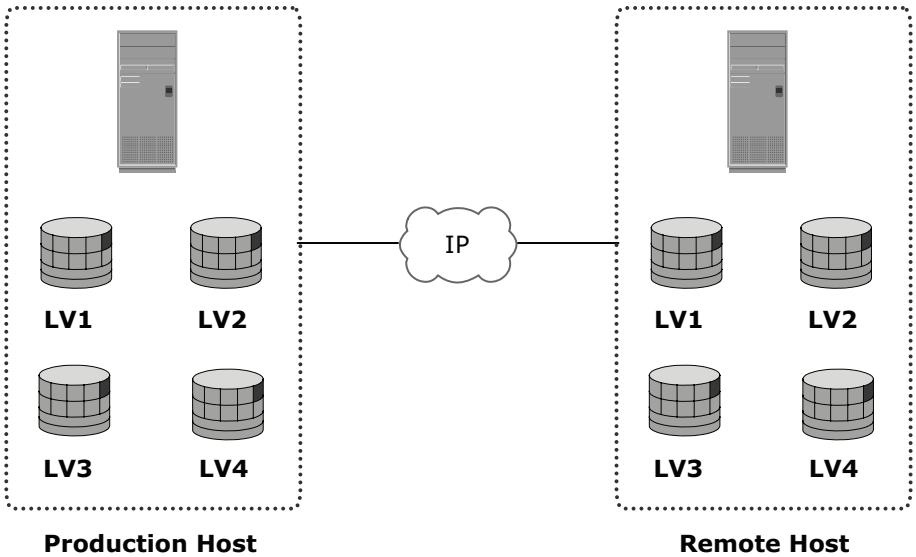


Figure 14-3: LVM-based remote replication

LVM-based remote replication supports both synchronous and asynchronous modes of data transfer. In asynchronous mode, writes are queued in a log file at the source and sent to the remote host in the order in which they were received. The size of the log file determines the RPO at the remote site. In the event of a network failure, writes continue to accumulate in the log file. If the log file fills up before the failure is resolved, then a full resynchronization is required upon network availability. In the event of a failure at the source site, applications can be restarted on the remote host, using the data on the remote replicas.

LVM-based remote replication eliminates the need for a dedicated SAN infrastructure. LVM-based remote replication is independent of the storage arrays and types of disks at the source and remote sites. Most operating systems are shipped with LVMs, so additional licenses and specialized hardware are not typically required.

The replication process adds overhead on the host CPUs. CPU resources on the source host are shared between replication tasks and applications, which may cause performance degradation of the application.

As the remote host is also involved in the replication process, it has to be continuously up and available. LVM-based remote replication does not scale well, particularly in the case of applications using *federated databases*.

Host-Based Log Shipping

Database replication via log shipping is a host-based replication technology supported by most databases. Transactions to the source database are captured in logs, which are periodically transmitted by the source host to the remote host (see Figure 14-4). The remote host receives the logs and applies them to the remote database.

Prior to starting production work and replication of log files, all relevant components of the source database are replicated to the remote site. This is done while the source database is shut down.

After this step, production work is started on the source database. The remote database is started in a standby mode. Typically, in standby mode, the database is not available for transactions. Some implementations allow reads and writes from the standby database.

All DBMSs switch log files at preconfigured time intervals, or when a log file is full. The current log file is closed at the time of log switching and a new log file is opened. When a log switch occurs, the closed log is transmitted by the source host to the remote host. The remote host receives the log and updates the standby database.

This process ensures that the standby database is consistent up to the last committed log. RPO at the remote site is finite and depends on the size of the log and the frequency of log switching. Available network bandwidth, latency, and rate of updates to the source database, as well as the frequency of log switching, should be considered when determining the optimal size of the log file.

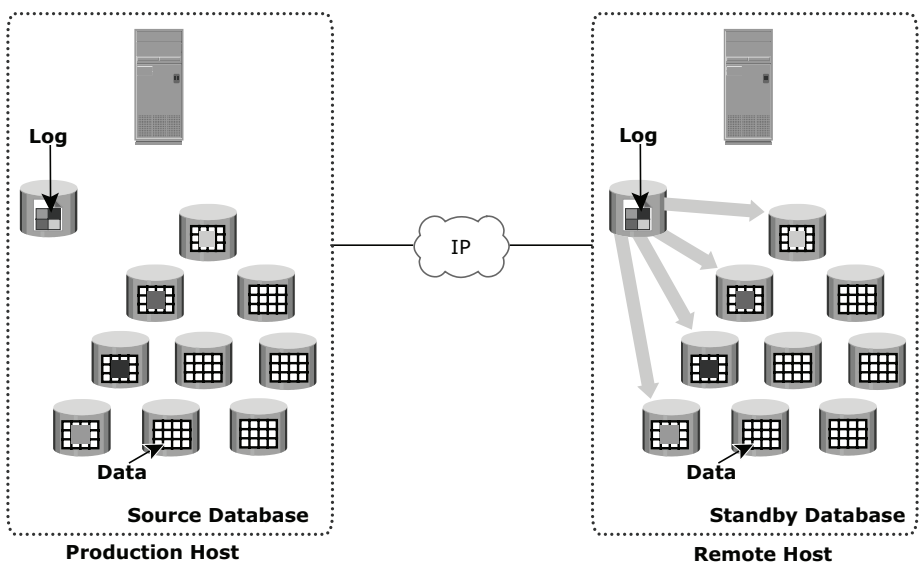


Figure 14-4: Host-based log shipping

Because the source host does not transmit every update and buffer them, this alleviates the burden on the source host CPU. Similar to LVM-based remote replication, the existing standard IP network can be used for replicating log files. Host-based log shipping does not scale well, particularly in the case of applications using federated databases.

14.2.2 Storage Array-Based Remote Replication

In *storage array-based remote replication*, the array operating environment and resources perform and manage data replication. This relieves the burden on the host CPUs, which can be better utilized for running an application. A source and its replica device reside on different storage arrays. In other implementations, the storage controller is used for both the host and replication workload. Data can be transmitted from the source storage array to the target storage array over a shared or a dedicated network.

Replication between arrays may be performed in synchronous, asynchronous, or disk-buffered modes. Three-site remote replication can be implemented using a combination of synchronous mode and asynchronous mode, as well as a combination of synchronous mode and disk-buffered mode.

Synchronous Replication Mode

In array based synchronous remote replication, writes must be committed to the source and the target prior to acknowledging "write complete" to the host. Additional writes on that source cannot occur until each preceding write has been completed and acknowledged. The array-based synchronous replication process is shown in Figure 14-5.

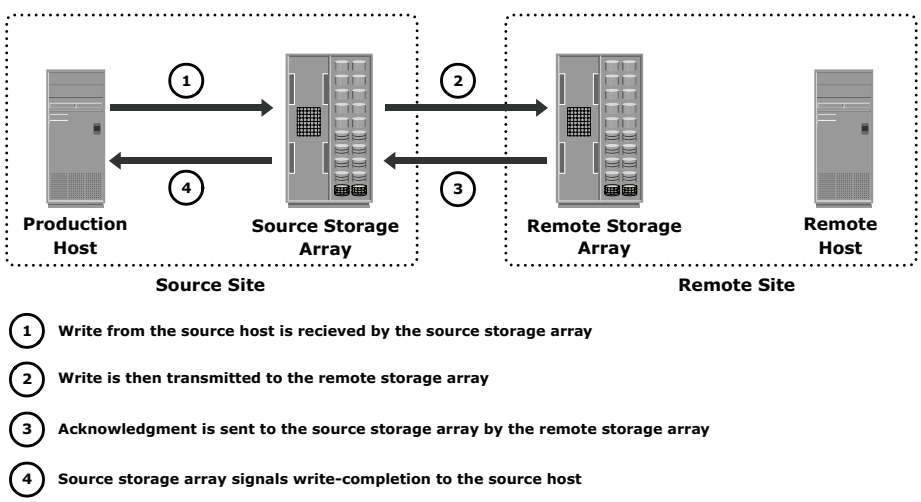


Figure 14-5: Array-based synchronous remote replication

In the case of synchronous replication, to optimize the replication process and to minimize the impact on application response time, the write is placed on cache of the two arrays. The intelligent storage arrays can de-stage these writes to the appropriate disks later.

If the network links fail, replication is suspended; however, production work can continue uninterrupted on the source storage array. The array operating environment can keep track of the writes that are not transmitted to the remote storage array. When the network links are restored, the accumulated data can be transmitted to the remote storage array. During the time of network link outage, if there is a failure at the source site, some data will be lost and the RPO at the target will not be zero.

For synchronous remote replication, network bandwidth equal to or greater than the maximum *write workload* between the two sites should be provided at all times. Figure 14-6 illustrates the write workload (expressed in MB/s) over time. The "Max" line indicated in Figure 14-6 represents the required bandwidth that must be provisioned for synchronous replication. Bandwidths lower than the maximum write workload results in an unacceptable increase in application response time.

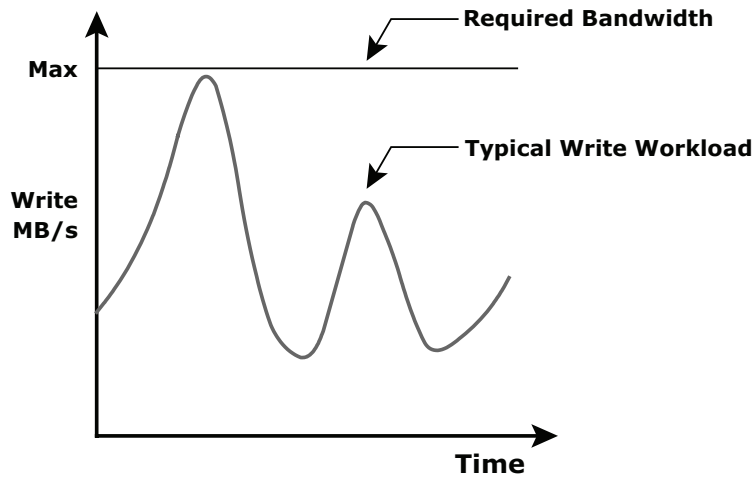


Figure 14-6: Network bandwidth requirement for synchronous replication

Asynchronous Replication Mode

In array-based *asynchronous remote replication mode*, shown in Figure 14-7, a write is committed to the source and immediately acknowledged to the host. Data is buffered at the source and transmitted to the remote site later. The source and the target devices do not contain identical data at all times. The data on the target device is behind that of the source, so the RPO in this case is not zero.

Similar to synchronous replication, asynchronous replication writes are placed in cache on the two arrays and are later de-staged to the appropriate disks.

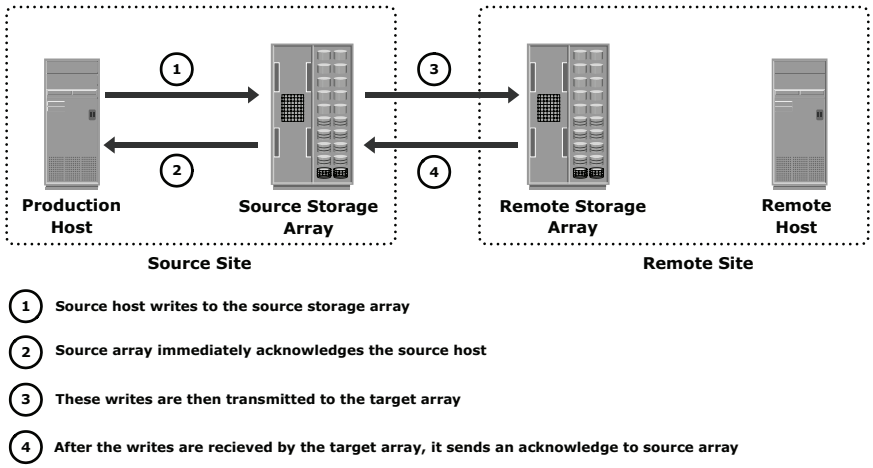


Figure 14-7: Array-based asynchronous remote replication

Some implementations of asynchronous remote replication maintain write ordering. A time stamp and sequence number are attached to each write when it is received by the source. Writes are then transmitted to the remote array, where they are committed to the remote replica in the exact order in which they were buffered at the source. This implicitly guarantees consistency of data on the remote replicas. Other implementations ensure consistency by leveraging the dependent write principle inherent to most DBMSs. The writes are buffered for a predefined period of time. At the end of this duration, the buffer is closed, and a new buffer is opened for subsequent writes. All writes in the closed buffer are transmitted together and committed to the remote replica.

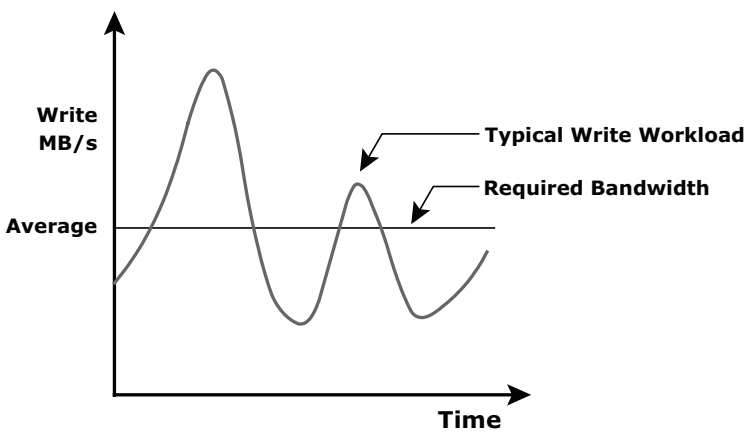


Figure 14-8: Network bandwidth requirement for asynchronous replication

Asynchronous remote replication provides network bandwidth cost savings, as only bandwidth equal to or greater than the average write workload is needed, as represented by the "Average" line in Figure 14-8. During times when the write workload exceeds the average bandwidth, sufficient buffer space has to be configured on the source storage array to hold these writes.

Disk-Buffered Replication Mode

Disk-buffered replication is a combination of local and remote replication technologies. A consistent PIT local replica of the source device is first created. This is then replicated to a remote replica on the target array.

The sequence of operations in a disk-buffered remote replication is shown in Figure 14-9. At the beginning of the cycle, the network links between the two arrays are suspended and there is no transmission of data. While production application is running on the source device, a consistent PIT local replica of the source device is created. The network links are enabled, and data on the local replica in the source array is transmitted to its remote replica in the target array. After synchronization of this pair, the network link is suspended and the next local replica of the source is created. Optionally, a local PIT replica of the remote device on the target array can be created. The frequency of this cycle of operations depends on available link bandwidth and the data change rate on the source device.

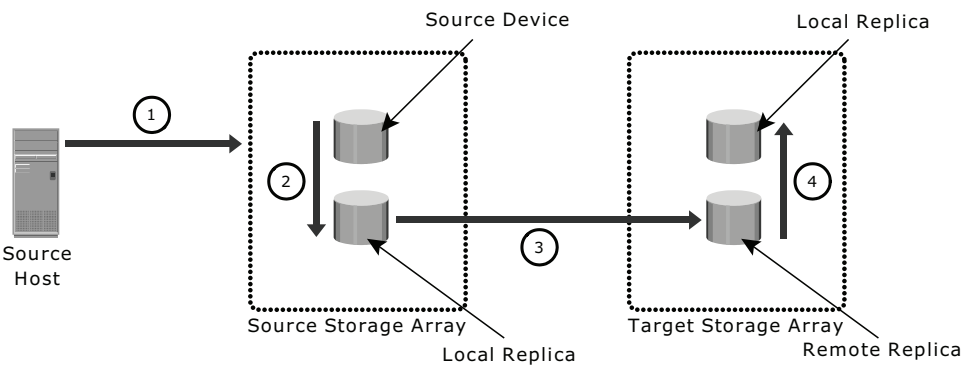


Figure 14-9: Disk-buffered remote replication

Array-based replication technologies can track changes made to the source and target devices. Hence, all resynchronization operations can be done incrementally.

For example, a local replica of the source device is created at 10:00 AM and this data is transmitted to the remote replica, which takes one hour to complete. Changes made to the source device after 10:00 AM are tracked. Another replica of the source device is created at 11:00 AM by applying track changes between the source and local replica (10:00 AM copy). During the next cycle of transmission (11:00 AM data), the source data has moved to 12:00 PM The local replica in the remote array has the 10:00 AM data until the 11:00 AM data is successfully transmitted to the remote replica. If there is a failure at the source site prior to the completion of transmission, then the worst-case RPO at the remote site would be two hours (as the remote site has 10:00 AM data).

Three-Site Replication

In synchronous and asynchronous replication, under normal conditions the workload is running at the source site. Operations at the source site will not be disrupted by any failure to the target site or to the network used for replication. The replication process resumes as soon as the link or target site issues are resolved. The source site continues to operate without any remote protection. If failure occurs at the source site during this time, RPO will be extended.

In synchronous replication, source and target sites are usually within 200 KM (125 miles) of each other. Hence, in the event of a regional disaster, both the source and the target sites could become unavailable. This will lead to extended RPO and RTO because the last known good copy of data would have to come from another source, such as offsite tape library.

A regional disaster will not affect the target site in asynchronous replication, as the sites are typically several hundred or several thousand kilometers apart. If the source site fails, production can be shifted to the target site, but there will be no remote protection until the failure is resolved.

Three-site replication is used to mitigate the risks identified in two-site replication. In a three-site replication, data from the source site is replicated to two remote data centers. Replication can be synchronous to one of the two data centers, providing a zero-RPO solution. It can be asynchronous or disk buffered to the other remote data center, providing a finite RPO. Three-site remote replication can be implemented as a cascade/multi-hop or a triangle/multi-target solution.

Three-Site Replication—Cascade/Multi-hop

In the *cascade/multi-hop* form of replication, data flows from the source to the intermediate storage array, known as a *bunker,* in the first hop and then from a bunker to a storage array at a remote site in the second hop. Replication between the source and the bunker occurs synchronously, but replication between the bunker and the remote site can be achieved in two ways: disk-buffered mode or asynchronous mode.

Synchronous + Asynchronous

This method employs a combination of synchronous and asynchronous remote replication technologies. Synchronous replication occurs between the source and the bunker. Asynchronous replication occurs between the bunker and the remote site. The remote replica in the bunker acts as the source for the asynchronous replication to create a remote replica at the remote site. Figure 14-10(a) illustrates the synchronous + asynchronous method.

RPO at the remote site is usually on the order of minutes in this implementation. In this method, a minimum of three storage devices are required (including the source) to replicate one storage device. The devices containing a synchronous remote replica at the bunker and the asynchronous replica at the remote are the other two devices.

If there is a disaster at the source, operations are failed over to the bunker site with zero or near-zero data loss. But unlike the synchronous two-site situation, there is still remote protection at the third site. The RPO between the bunker and third site could be on the order of minutes.

If there is a disaster at the bunker site or if there is a network link failure between the source and bunker sites, the source site will continue to operate as normal but without any remote replication. This situation is very similar to two-site replication when a failure/disaster occurs at the target site. The updates to the remote site cannot occur due to the failure in the bunker site. Hence, the data at the remote site keeps falling behind, but the advantage here is that if the source fails during this time, operations can be resumed at the remote site. RPO at the remote site depends on the time difference between the bunker site failure and source site failure.

A *regional disaster* in three-site cascade/multihop replication is very similar to a source site failure in two-site asynchronous replication. Operations will failover to the remote site with an RPO on the order of minutes. There is no remote protection until the regional disaster is resolved. Local replication technologies could be used at the remote site during this time.

If a disaster occurs at the remote site, or if the network links between the bunker and the remote site fail, the source site continues to work as normal with disaster recovery protection provided at the bunker site.

Synchronous + Disk Buffered

This method employs a combination of local and remote replication technologies. Synchronous replication occurs between the source and the bunker: A consistent PIT local replica is created at the bunker. Data is transmitted from the local replica at the bunker to the remote replica at the remote site. Optionally, a local replica can be created at the remote site after data is received from the bunker. Figure 14-10(b) illustrates the synchronous + disk buffered method.

In this method, a minimum of four storage devices are required (including the source) to replicate one storage device. The other three devices are the

synchronous remote replica at the bunker, a consistent PIT local replica at the bunker, and the replica at the remote site. RPO at the remote site is usually in the order of hours in this implementation. For example, if a local replica is created at 10:00 AM at the bunker and it takes an hour to transmit this data to the remote site, changes made to the remote replica at the bunker since 10:00 AM are tracked. Hence only one hour's worth of data has to be resynchronized between the bunker and the remote site during the next cycle. RPO in this case will also be two hours, similar to disk-buffered replication.

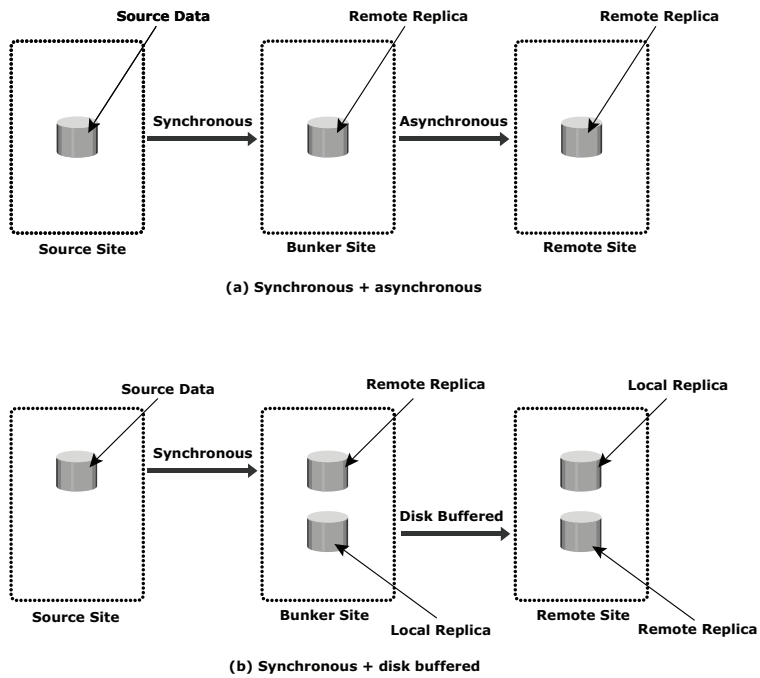


Figure 14-10: Three-site replication

The process of creating the consistent PIT copy at the bunker and incrementally updating the remote replica and the local replica at the remote site occurs continuously in a cycle. This process can be automated and controlled from the source.

Three-Site Replication—Triangle/Multi-target

In the *three-site triangle/multi-target replication,* data at the source storage array is concurrently replicated to two different arrays. The source-to-bunker site (target 1) replication is synchronous, with a near-zero RPO. The source-to-remote site (target 2) replication is asynchronous, with an RPO of minutes. The distance between the source and the remote site could be thousands of

miles. This configuration does not depend on the bunker site for updating data on the remote site, because data is asynchronously copied to the remote site directly from the source.

The key benefit of three-site triangle/multi-target replication is the ability to failover to either of the two remote sites in the case of source site failure, with disaster recovery (asynchronous) protection between them. Resynchronization between the two surviving target sites is incremental. Disaster recovery protection is always available in the event of any one site failure.

During normal operations all three sites are available and the workload is at the source site. At any given instant, the data at the bunker and the source is identical. The data at the remote site is behind the data at the source and the bunker. The replication network links between the bunker and remote sites will be in place but not in use. Thus, during normal operations there is no data movement between the bunker and remote arrays. The difference in the data between the bunker and remote sites is tracked, so that in the event of a source site disaster, operations can be resumed at the bunker or the remote sites with incremental resynchronization between the sites.

14.2.3 SAN-Based Remote Replication

SAN-based remote replication enables the replication of data between heterogeneous storage arrays. Data is moved from one array to the other over the SAN/WAN. This technology is application and server operating system independent, because the replication operations are performed by one of the storage arrays (the control array). There is no impact on production servers (because replication is done by the array) or the LAN (because data is moved over the SAN).

SAN-based remote replication is a point-in-time replication technology. Uses of SAN-based remote replication include data mobility, remote vaulting, and data migration. Data mobility enables incrementally copying multiple volumes over extended distances, as well as implementing a tiered storage strategy. Data vaulting is the practice of storing a set of point-in-time copies on heterogeneous remote arrays to guard against a failure of the source site. Data migration refers to moving data to new storage arrays and consolidating data from multiple heterogeneous storage arrays onto a single storage array.

The array performing the replication operations is called the *control array*. Data can be moved to/from devices in the control array to/from a remote array. The devices in the control array that are part of the replication session are called *control devices*. For every control device there is a counterpart, a *remote device*, on the *remote* array.

The terms "control" or "remote" do not indicate the direction of data flow, they only indicate the array that is performing the replication operation. Data movement could be from the control array to the remote array or vice versa. The direction of data movement is determined by the replication operation.

The front-end ports of the control array must be zoned to the front-end ports of the remote array. LUN masking should be performed on the remote array to allow access to the remote devices to the front-end port of the control array. In effect, the front-end ports of the control array act as an HBA, initiating data transfer to/from the remote array.

SAN-based replication uses two types of operations: *push* and *pull*. These terms are defined from the perspective of the control array. In the *push* operation, data is transmitted from the control storage array to the remote storage array. The control device, therefore, acts like the source, while the remote device is the target. The data that needs to be replicated would be on devices in the control array.

In the *pull* operation, data is transmitted from the remote storage array to the control storage array. The remote device is the source and the control device is the target. The data that needs to be replicated would be on devices in the remote array.

When a push or pull operation is initiated, the control array creates a protection bitmap to track the replication process. Each bit in the protection bitmap represents a data chunk on the control device. Chunk size may vary with technology implementations. When the replication operation is initiated, all the bits are set to one, indicating that all the contents of the source device need to be copied to the target device. As the replication process copies data, the bits are changed to zero, indicating that a particular chunk has been copied. At the end of the replication process, all the bits become zero.

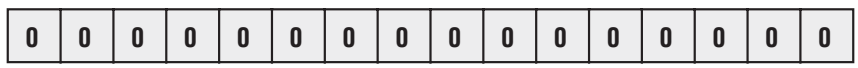
During the push and pull operations, host access to the remote device is not allowed because the control storage array has no control over the remote storage array and cannot track any change on the remote device. Data integrity cannot be guaranteed if changes are made to the remote device during the push and pull operations. Therefore, for all SAN-based remote replications, the remote devices should not be in use during the replication process in order to ensure data integrity and consistency.

The push/pull operations can be either *hot* or *cold*. These terms apply to the control devices only. In a cold operation, the control device is inaccessible to the host during replication. Cold operations guarantee data consistency because both the control and the remote devices are offline to every host operation. In a hot operation, the control device is online for host operations. With hot operations, changes can be made to the control device during push/pull because the control array can keep track of all changes, and thus ensures data integrity.

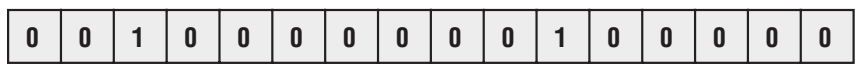
When the hot push operation is initiated, applications can be up and running on the control devices. I/O to the control devices is held while the protection bitmap is created. This ensures a consistent PIT image of the data. The protection bitmap is referred prior to any write to the control devices. If the bit is zero, the write is allowed. If the bit is one, the replication process holds the write, copies the required chunk to the remote device, and then allows the write to complete.

In the hot pull operation, the hosts can access control devices after starting the pull operation. The protection bitmap is referenced for every read or write operation. If the bit is zero, a read or write occurs. If the bit is one, the read or write is held, and the replication process copies the required chunk from the remote device. When the chunk is copied, the read or write is completed. The control devices can be used after the pull operation is initiated and as soon as the protection bitmap is created.

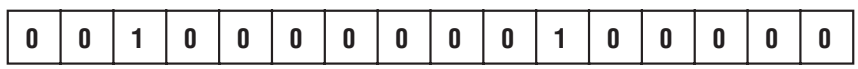
In SAN-based replication, the control array can keep track of changes made to the control devices after the replication session is activated. This is allowed in the incremental push operation only. A second bitmap, called a *resynchronization bitmap,* is created. All the bits in the resynchronization bitmap are set to zero when a push is initiated, as shown in Figure 14-11 (a). As changes are made to the control device, the bits are flipped from zero to one, indicating that changes have occurred, as shown in Figure 14-11 (b). When resynchronization is required, the push is reinitiated and the resynchronization bitmap becomes the new protection bitmap, as shown in Figure 14-11 (c), and only the modified chunks are transmitted to the remote devices. If changes are made to the remote device, the SAN-based replication operation is unaware of these changes, therefore, data integrity cannot be ensured if an incremental push is performed.



(a) Resynchronization bitmap when push is initiated



(b) Resynchronization bitmap when data chunks are updated



(c) Resynchronization bitmap becomes the protection bitmap

Figure 14-11: Bitmap status in SAN-based replication

14.3 Network Infrastructure

For remote replication over extended distances, optical network technologies such as dense wavelength division multiplexing (DWDM), coarse wavelength division multiplexing (CWDM), and synchronous optical network (SONET) are deployed.

14.3.1 DWDM

DWDM is an optical technology by which data from different channels are carried at different wavelengths over a fiber-optic link. It is a fiber-optic transmission technique that uses light waves to transmit data parallel by bit or serial by character. It integrates multiple light waves with different wavelengths in a group and directs them through a single optical fiber.

The multiplexing of data from several channels into a multicolored light stream transmitted on a single optical fiber has opened up the conventional optical fiber bandwidth by breaking it into many channels, each at a different optical wavelength. Each wavelength can carry a signal at a bit rate less than the upper limit defined by the electronics; typically up to several gigabits per second. Using DWDM, different data formats at different data rates can be transmitted together. Specifically, IP ESCON, FC, SONET and ATM data can all travel at the same time within the optical fiber (see Figure 14-12).

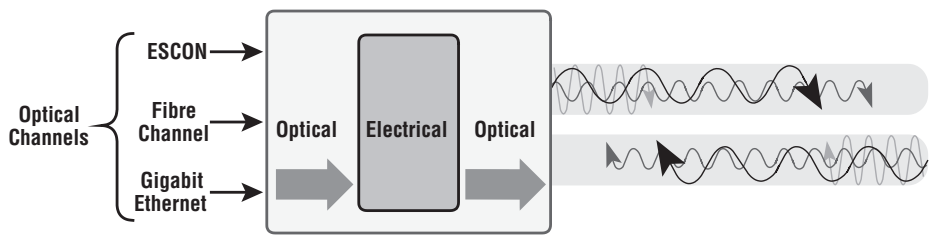


Figure 14-12: Dense wavelength division multiplexing (DWDM)

CWDM, like DWDM, uses multiplexing and demultiplexing on different channels by assigning varied wavelengths to each channel. Compared to DWDM, CWDM is used to consolidate environments containing a low number of channels at a reduced cost.

14.3.2 SONET

SONET (synchronous optical network) is a network technology that involves transferring a large payload through an optical fiber over long distances. SONET multiplexes data streams of different speeds into a frame and sends them across the network. The European variation of SONET is called synchronous digital hierarchy (SDH). Figure 14-13 shows the multiplexing of data streams of different speeds in SONET and SDH technologies.

SONET/SDH uses generic framing procedure (GFP) and supports the transport of both packet-oriented (Ethernet, IP) and character-oriented (FC) data.

Figure 14-13: Data stream multiplexing in SONET and SDH

SONET transfers data at a very high speed (for example, OC-768 provides line rate up to 40 Gbps). The basic SONET/SDH signal operates at 51.84 Mbps and is designated synchronous transport signal level one (STS-1). The STS-1 frame is the basic unit of transmission in SONET/SDH. Multiple STS-1 circuits can be aggregated to form higher-speed links. STS-3 (155.52 Mb/s) is equivalent to SONET level OC-3 and SDH level STM-1.

14.4 Concepts in Practice: EMC SRDF, EMC SAN Copy, and EMC MirrorView

This section discusses three EMC products that use remote replication technology. EMC Symmetrix Remote Data Facility (SRDF) and EMC MirrorView are storage array–based remote application softwares supported by EMC Symmetrix and CLARiiON respectively. EMC SAN Copy is SAN-based remote replication software deployed in an EMC CLARiiON storage array. For the latest information, visit `http://education.EMC.com/ismbook`.

14.4.1 SRDF Family

SRDF offers a family of technology solutions to implement storage array-based remote replication technologies. The three Symmetrix solutions are:

- **SRDF/Synchronous (SRDF/S):** A remote replication solution that creates a synchronous replica at one or more Symmetrix targets
- **SRDF/Asynchronous (SRDF/A):** A remote replication solution that enables the source to asynchronously replicate data, incorporating delta set technology and dependent write consistency. A delta set enables write ordering by employing a buffering mechanism.
- **SRDF/Automated Replication (SRDF/AR):** A remote replication solution that uses both SRDF and TimeFinder/Mirror to implement disk-buffered replication technology

14.4.2 Disaster Recovery with SRDF

The source arrays have SRDF R1 devices (source devices), and the target arrays have SRDF R2 devices (replica devices). Data written to R1 devices is replicated to R2 devices, either synchronously or asynchronously. SRDF R1 and R2 devices can have any local RAID protection, such as RAID 1 or RAID 5. SRDF R2 devices are in a read-only (R/O) state when remote replication is in effect. Hence, under normal operating conditions, changes cannot be made directly to the R2 devices. The R2 devices can only receive data from their corresponding R1 devices on the source storage array.

SRDF uses dedicated adapters (controllers) to send data from the source to the target storage array. The supported adapters for remote replication are ESCON, FC, and GigE.

Shifting production work from the source site to the target site is done by the SRDF *failover* operation, whereas shifting the production work from the target site back to the source site is done by the SRDF *failback* operation:

- **Failover:** The failover operation is initiated if the SRDF R1 devices are unavailable and if BC operations need to restart on the R2 devices. The failover operation can also be performed for testing the disaster-recovery processes and for any maintenance tasks at the source site. Figure 14-14 shows the I/O status before and after failover. Before failover, the source allows the R/W operation, whereas a host accesses the target only in the R/O state. The failover process enables a host to perform R/W operations with the targets.

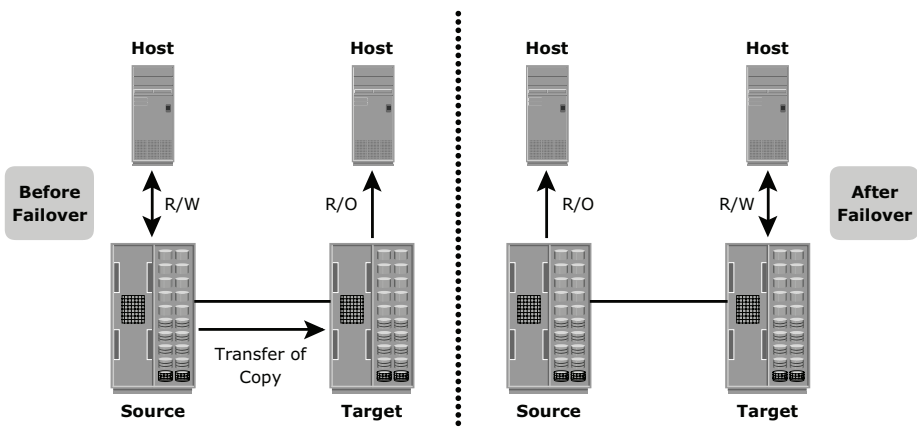


Figure 14-14: EMC SRDF - before and after failover

- **Failback:** The failback operation allows normal business operations to resume at the source site on the R1 device. When failback is invoked, the target becomes R/O, the source becomes R/W, and any changes that were

made at the target while in the failover state are incrementally propagated back to the source, as illustrated in Figure 14-15.

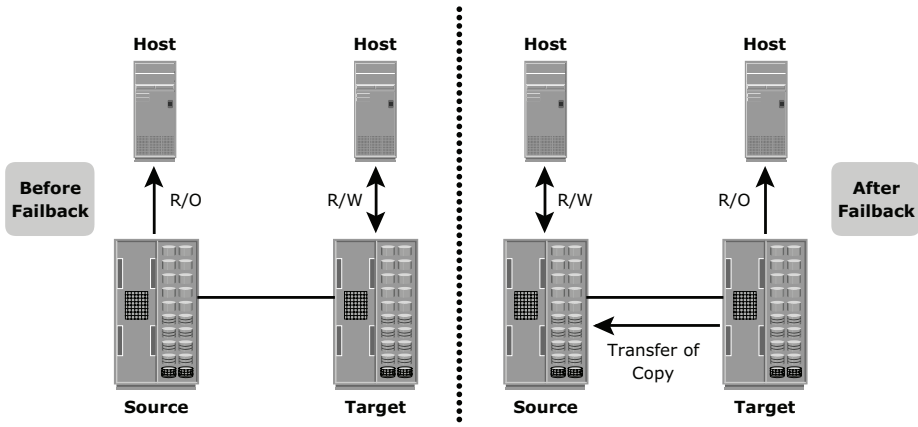


Figure 14-15: EMC SRDF - before and after failback

14.4.3 SRDF Operations for Concurrent Access

SRDF provides split operations to enable concurrent access to both source and target devices. The establish and restore operations are used to return the source-target pairs to the normal SRDF state.

In *split* operation, when R2 is split from R1, BC operations can be performed on R2. The split operation enables concurrent access to both the source and the target devices. In this operation, target devices are made R/W, and the SRDF replication between the source and the target is suspended, as shown in Figure 14-16. For the duration of the split, there is no remote data protection.

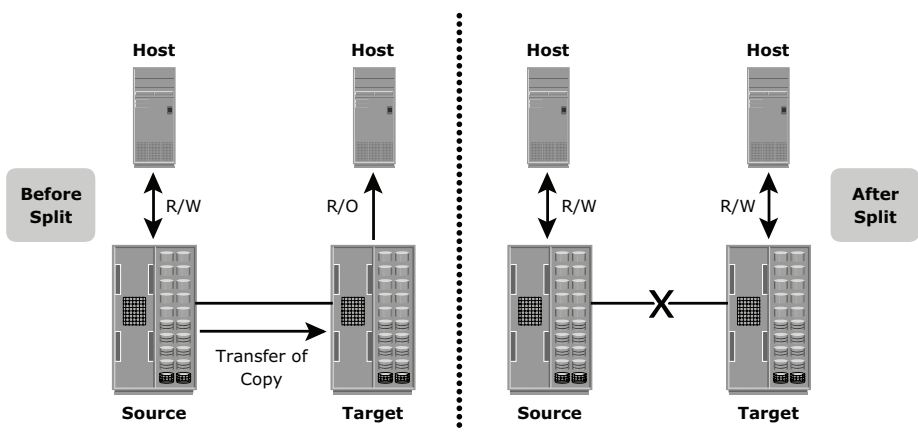


Figure 14-16: Concurrent access with EMC SRDF

During concurrent operations while in a SRDF split state, changes could occur on both the source and the target devices. Normal SRDF replication can be resumed by performing an *establish* or a *restore* operation. With either establish or restore, the status of the target device becomes R/O (see Figure 14-17).

The establish operation is used when changes to the target device should be discarded, while preserving changes that were made to the source device. The restore operation is used when changes to the source device should be discarded, while preserving changes that were made to the target device. Prior to a restore operation, all access to the source and target devices must be stopped. The target device switches to R/O status, while the data on the source device is overwritten with the data from the target device.

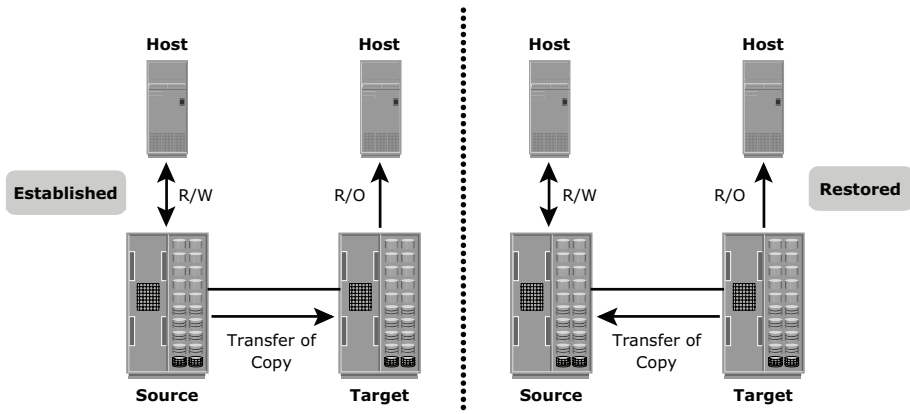


Figure 14-17: Restarting SRDF replication after concurrent access

14.4.4 EMC SAN Copy

SAN Copy is CLARiiON software that performs SAN-based remote replication between CLARiiON and Symmetrix or other vendor storage arrays. It enables simultaneous creation of one or more copies of source devices to target devices through a SAN. Source and target devices could either be on a single array or on multiple arrays. SAN Copy software on the CLARiiON (designated as the control storage array) controls the entire replication process. The source ports utilized during SAN Copy data transfer must be zoned to the target port, and LUN masking must be implemented on the target to perform remote replication. Additional features offered by SAN Copy include the following:

- **Automatic check pointing in the event of a link failure:** Checkpoints are written on the disk drive of the source CLARiiON at the administrator-defined time intervals. This feature enables SAN Copy to resume an interrupted replication session from the last checkpoint.

- **Transfer rate throttle:** The SAN Copy transfer rate can be controlled by throttling network bandwidth. The throttle value ranges from 1 (low) to 10 (high). The transfer rates of each concurrent replication session are adjusted to different throttle values depending on bandwidth requirements.
- **Incremental SAN Copy:** The target logical device is synchronized with the source logical device after an initial full copy replication session is over. After a full copy, SAN Copy establishes an incremental session to transfer only updated data to the target, which lowers the network bandwidth requirement. Changes at the source are tracked by a protection bitmap. The source device is R/O during a full-copy replication session, whereas the incremental session allows R/W at the source.

14.4.5 EMC MirrorView

EMC MirrorView is a CLARiiON-based software that enables storage array–based remote replication over FC SAN, IP extended SAN, and TCP/IP networks. MirrorView family consists of MirrorView/Synchronous (MirrorView/S), and MirrorView/Asynchronous (MirrorView/A). MirrowView software must be installed at both source and target CLARiiON in order to perform remote replication. Any CLARiiON running MirrorView can simultaneously house primary (source) LUNs for some applications and secondary (target) LUNs for others. MirrorView supports both synchronous and asynchronous replication of data on the same CLARiiON. It also supports consistency groups for maintaining data consistency across write-order dependent LUNs.

MirrorView Operations

Initial Synchronization is a replication process that is used for new mirrors (target) to create an initial copy of the primary/primary image (LUN on source CLARiiON containing production data). During the initial synchronization process, the primary images remain online whereas the secondary/secondary image (LUN that contains a mirror of the primary image) is inaccessible. Until the initial synchronization process is complete, secondary images are in the *synchronizing* state. If the synchronization were interrupted, the secondary image would be in the *out-of-sync* state indicating the secondary data is not in a consistent state. When synchronization completes, the mirror data state will be *in-sync*.

A *fracture* operation stops MirrorView replication. An administrator can initiate fracture to suspend the replication. MirrorView software can automatically fracture when it senses a connectivity failure between the primary and secondary LUNs. Replication can resume when the user executes the *synchronize* command.

In the event of a link failure, MirrorView recovery policy provides two options. An *auto-recovery* option starts synchronization as soon as the secondary image is reachable, or a *manual recovery,* where MirrorView waits for a synchronization request from the user.

MirrorView/S invokes a *fracture log* when the secondary image is fractured. The fracture log is a bitmap held in the memory of the storage processor that owns the primary LUN. When the secondary is reachable, using the fracture log, the primary and secondary LUNs can be synchronized by transmitting only the updated information to the secondary. MirrorView/S also uses a *write intent log*, but unlike the fracture log, which is enabled when the mirror is fractured, the write intent log is always active. The write intent log is stored persistently on the disk in the source CLARiiON. Before the primary and secondary LUNs are updated, an entry takes place at the write intent log to indicate locations at the primary LUNs where data changes will occur. In the event of a storage processor failure, the write intent log will be used to determine which locations must be synchronized from the primary LUNs to the secondary LUNs.

MirrorView/A does not use fracture and write intent logs, but it tracks locations (using SnapView technology) at the primary LUNs where updates occur. MirrorView/A utilizes the delta set mechanism to periodically transfer data to the secondary LUNs. MirrorView uses two bitmaps on the primary LUNs. For each update, one bitmap (the *tracking map*) tracks changes between updates, and the other bitmap (the *transfer map*) tracks the progress of the update when transferring to the secondary. The tracking and transfer maps are persistently stored at the reserved LUN pool before mirror operations are initiated.

A secondary image is *promoted* to the role of primary, when it is necessary to run production applications at the disaster recovery site. This may be in response to an actual disaster at the source site, part of a migration strategy, or simply for testing purposes.

Summary

This chapter detailed remote replication. As a primary utility, remote replication provides disaster recovery and disaster restart solutions. It enables business operations to be rapidly restarted at a remote site following an outage, with acceptable data loss.

Remote replication enables BC operations from a target site. The replica of source data at the target can be used for backup and testing. This replica can also be used for data repurposing, such as report generation, data warehousing, and decision support. The segregation of business operations between the source and target protects the source from becoming a performance bottleneck, ensuring improved production performance at the source.

Remote replication may also be used for data center migrations, providing the least disturbance to production operations because the applications accessing the source data are not affected.

This chapter also described different types of remote replication solutions. The distance between the primary site and the remote site is a prime consideration when deciding which remote replication technology solution to deploy. Asynchronous replication may adequately meet the RPO and RTO needs, while permitting greater distances between the sites.

Storage management solutions provide the capability to not only automate business continuity solutions, but also enable centralized management of the overall storage infrastructure. Organizations must ensure security of the information assets. The next chapter details storage security and management.

EXERCISES

1. **An organization is planning a data center migration. They can only afford a maximum of two hours downtime to complete the migration. Explain how remote replication technology can be used to meet the downtime requirements. Why will the other methods not meet this requirement?**
2. **Explain the RPO that can be achieved with synchronous, asynchronous, and disk-buffered remote replication.**
3. **Discuss the effects of a bunker failure in a three-site replication for the following implementation:**
   - **Multihop—synchronous + disk buffered**
   - **Multihop—synchronous + asynchronous**
   - **Multi-target**
4. **Discuss the effects of a source failure in a three-site replication for the following implementation, and the available recovery options:**
   - **Multihop—synchronous + disk buffered**
   - **Multihop—synchronous + asynchronous**
   - **Multi-target**
5. **A host generates 8,000 I/Os at peak utilization with an average I/O size of 32 KB. The response time is currently measured at an average of 12 ms during peak utilizations. When synchronous replication is implemented with a Fibre Channel link to a remote site, what is the response time experienced by the host if the network latency is 6 ms per I/O?**
6. **Research the remote replication options in a NAS environment. Which type of replication is best suited in integrated and gateway NAS solutions?**