the Availability Digest

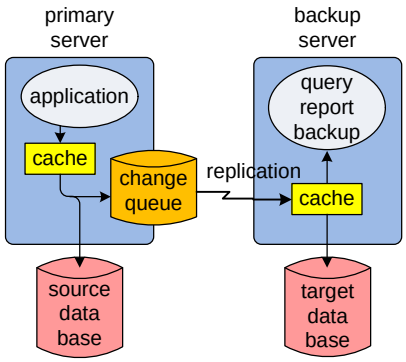
www.availabilitydigest.com

EMC's SRDF Data-Replication Engine

April 2011

Logical Replication versus Hardware Replication

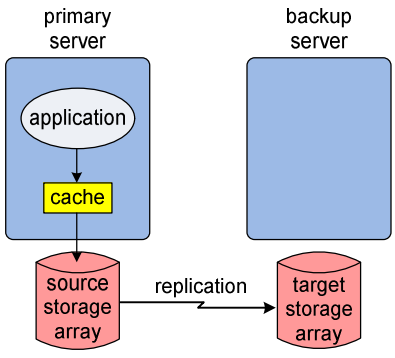
In most of our reviews of data-replication engines, we have focused on server-based logical replicators.[1] These replicators provide a consistent copy of a source database on a remote target database. Logical replicators depend upon a queue of changes maintained by the database manager or application. They follow the changes and replicate them to the target database.



Server-Based Logical Replication

The target database can therefore be used for query and reporting purposes, and it can participate in active/active application networks if bidirectional replication is provided. The target database is also available to be brought online rapidly to provide high availability in the event of a source-system failure. Failover in an active/backup architecture can be achieved in minutes.

Server-based replication requires the use of host-system resources. Though this usage is often minimal, it is eliminated if a storage array is used and if data replication is performed by the storage array itself.[2] Many storage-array replicators replicate disk blocks as they are physically written to the source storage array. Therefore, the target storage array always reflects the contents of the source array.

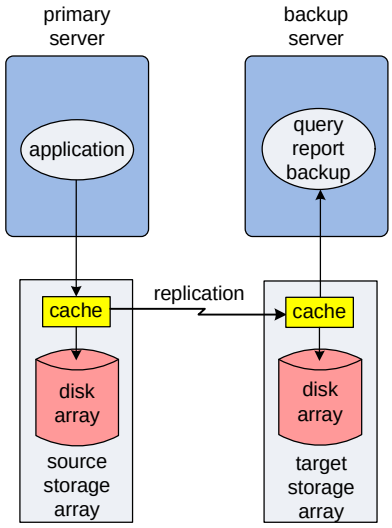


Storage-Array Physical Replication

However, the contents of the storage array do not represent a consistent view of the database since much of the current content of the database is maintained in cache and is not written to the storage array until cache space is needed. In effect, the contents of the target storage array represent a corrupted view of the source database. If the source system fails, there is a great deal of work required to bring the target database into a consistent state, often over an extended period of time typically measured in hours. Therefore, these storage array replicators provide disaster recovery; but they do not provide high availability.

[1] Asynchronous Replication Engines, *Availability Digest*; November 2006.
Synchronous Replication, *Availability Digest*; December 2006.
[2] Hardware Replication, *Availability Digest*; January 2007.

An exception to this is provided by storage arrays that provide cache storage as well as disk storage. These storage arrays replicate disk blocks from cache as they are changed. Therefore, the target storage array always represents a consistent copy of the source database. The backup system can be used for functions such as queries, reports, and backups. Furthermore, if the source system fails, recovery is rapid, often measured in minutes. High availability is achieved.



Storage-Array Logical Replication

Such a replication engine is EMC's Symmetrix Remote Data Facility (SRDF), which we review in this article (www.emc.com).

Symmetrix Remote Data Facility (SRDF)

Symmetrix Storage Arrays

SRDF replicates data stored in Symmetrix DMX-4 storage arrays. They scale to massive amounts of storage and are architected to provide high performance under large workloads.

A DMX-4 array can contain about 2,000 disks organized as RAID 0 (no redundancy), RAID 1 (mirrored), RAID 10, RAID 5, or RAID 6. Disk capacities range from 146 gigabytes to one terabyte. A fully configured array can provide over 500 terabytes of mirrored or RAID storage.

Flash drives are supported by DMX-4. They increase performance by a factor of ten and reduce power consumption by 98%.

A DMX-4 can be configured with up to 256 gigabytes of mirrored global memory, which is used in large part for cache memory. To achieve maximum performance, DMX-4's Direct Matrix Architecture provides up to 128 one gigabit/second directly connected data paths between global memory and its I/O directors.

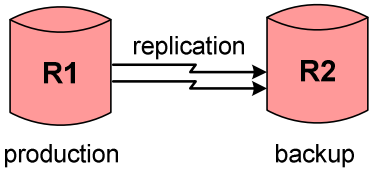
For host connectivity and remote replication communication, a Symmetrix DMX-4 can support up to 64 host ports and eight remote 4-Gb/s Fibre Channel or ESCON ports. Alternatively, it can support up to 48 host 1-Gb/s iSCSI or 4-Gb/s FICON ports and/or up to eight 1-Gb/s GigE remote ports.

The Symmetrix storage array is designed to provide totally nondisruptive operations. Hardware and software can be maintained and upgraded, and storage configurations can be changed with no application downtime.

Symmetrix DMX-4 arrays support mainframes and Windows, Linux, UNIX, and AIX platforms. Over 70,000 installations have been made around the world.

Data Replication Options

Introduced by EMC in 1994, SRDF provides several data replication options for DMX-4 storage arrays - both asynchronous (SRDF/A) and synchronous (SRDF/S). In SRDF terms, the production storage array is known as *R1*; and the backup storage array is known as *R2*.



SRDF must always be configured with redundant channels between R1 and R2 so that replication is not lost on a link failure.

SRDF/A

Using SRDF's *asynchronous-replication* mode (SRDF/A), data is replicated with no impact on the applications. As applications make changes to their databases, the changes complete immediately on R1 and are queued for replication to the target storage array, R2.

The replication queue in the source array, R1, is a set of pointers to data blocks that have changed (SRDF has no knowledge of the structure of the data beyond data blocks, such as table rows or fields). Replication is scheduled periodically according to configuration parameters. The set of changes that accumulate in a replication interval is known as a *delta set*. Replication intervals can be as short as seconds but are typically measured in minutes.

When it comes time to replicate a data set, SRDF uses the pointers in the replication queue to access the changed disk blocks and to transmit them to R2's cache. If a disk block is no longer in the cache of the source storage array, it is read from disk.

Note that a queue of changes is not maintained in the delta set in the R1 queue. Rather, the pointers point only to the latest value of a disk block in R1. If multiple changes are made to a disk block in a replication interval, only the latest change is sent; and SRDF sends that disk block only once even though it may appear several times in the queue. This means that during the replication process, R2 may not hold a consistent copy of the database. However, once the delta set has been replicated, R2's database will be consistent.

If a fault occurs during the transmission of a delta set, R2 will discard the entire delta set, thus maintaining the consistency of the target database.

If replication falls behind, SRDF/A has the capability to throttle host I/O so that replication can catch up.

SRDF/A can replicate over any distance, up to halfway around the world, typically over IP channels.

SRDF/S

SRDF/S, SRDF's *synchronous-replication* mode, ensures that each change has been stored in the target storage array's cache before the change is allowed to complete on the source storage array. This means that the application is delayed as SRDF awaits a completion acknowledgement from R2. However, it also means that R2's database is always a consistent copy in exact time synchronization with R1's database.

Because of the impact on application performance caused by synchronous-replication delay, SRDF/S can only be used over limited campus or metropolitan distances using optical fiber channels. The source and target systems can be three kilometers apart using a direct fiber connection. They can be separated by 66 kilometers by using repeaters and converters in 20 km segments.

SRDF/S supports an optional *semisynchronous-replication* mode. In this mode, a change completes immediately on R1 and does not wait for replication to complete. However, the next change is not accepted by R1 until the previous change is acknowledged by R2. This allows an application to proceed with its processing chores, including the reading of data from disk. The application is delayed only if it attempts to make a change before the previous change has been

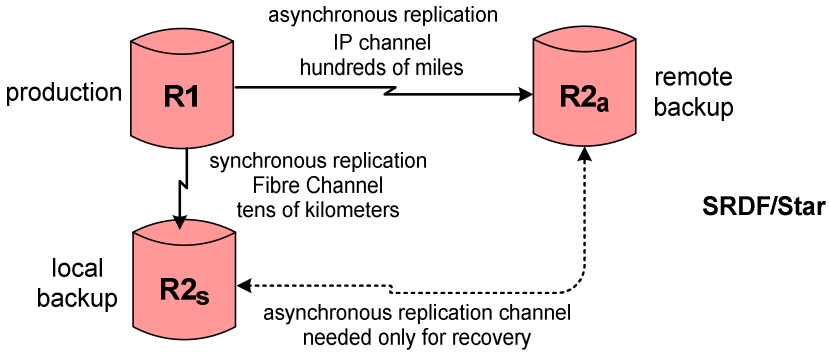
successfully replicated. Semisynchronous replication eases the application performance penalty imposed by pure synchronous replication, though it does not eliminate it.

SRDF/Star

Both SRDF/A and SRDF/S have their limitations. With SRDF/A, there is a delay from when a change is made to the source database and when it appears in the target database. This delay is known as *replication latency*. Should the source node fail, any changes that have not been replicated are lost. Thus, a Recovery Point Objective (RPO) of zero cannot be achieved. In fact, the RPO will be no better than the configured replication interval for replication delta sets. Since this interval is typically measured in minutes, RPOs will be in the order of minutes.

SRDF/S solves this problem. Since the replication of changes is synchronous, no data is lost if the source system fails. However, the distance limitation between the source and target nodes may not allow these nodes to be separated far enough to satisfy disaster-tolerance requirements. A 50-kilometer distance may protect against a building fire, but it may not protect against a major earthquake or flood. A separation of hundreds of miles or more may be required.

SRDF/Star is a configuration that overcomes these problems to a great extent. With SRDF/Star, there are three nodes - a production node, R1, and two target nodes. The production node replicates its database to both target nodes. One target node. $R2_S$, is nearby and is kept synchronized with SRDF/S synchronous replication. Its database is therefore an exact copy of the production database. The other target node, $R2_a$, can be hundreds or thousands of miles away and is synchronized by SRDF/A asynchronous replication. It therefore will survive any disaster that affects the production system.[3] The synchronous local backup system, $R2_S$, is often referred to as a *data bunker* because it safe-stores all data on behalf of the asynchronous remote backup, $R2_a$, which will lose data following a production-system failure.



There are several disaster recovery scenarios with SRDF/Star. If the production node, R1, goes down, operations can be moved to the nearby synchronously-replicated $R2_S$ node. Asynchronous replication is initiated from this node to the remote node, $R2_a$; and full production resumes with a completely current database and a disaster-recovery site. Alternatively, the database at $R2_a$ can be brought up-to-date by replicating missing changes to it from $R2_S$. Asynchronous replication can then be established from $R2_a$ to $R2_S$, and operations can be resumed at $R2_a$ with $R2_S$ acting as its disaster-recovery site. In either case, full operation is resumed with no data loss and with a functioning disaster-recovery site.

[3] For case studies of SRDF/Star in use, see Banks Use Synchronous Replication for Zero RPO, *Availability Digest*; February 2010.

If $R2_S$ should fail, disaster recovery is still provided by $R2_a$, though some data may be lost if the production node should subsequently fail. If $R2_a$ should fail, the system is subject to the possibility of a common disaster taking out both R1 and $R2_S$. Of course, the probability of a dual-node failure or a common disaster following a remote-node failure is extremely small.

Recovery

SRDF provides automatic recovery from a node failure. If the production node, R1, fails, R2 is automatically promoted to R1 and production continues on backup servers located at the backup site. Alternatively, if synchronous replication over Fibre Channel is being used, the servers at the production site, if still operable, can continue processing with the new R1 storage array at the backup site. In any event, starting processing at the R2 site is no different than restarting R1 following a power failure.

Fallback to the production site once the failed node is restored is also automatic under manual command. The backup site is quiesced to let all in-progress database activity complete. The production site's storage array is then synchronized with the backup site's storage array. The production site is brought up as R1, and the backup site's storage array is reverted to R2. The replication channel is activated, and production can now continue at the production site.

It is not necessary to have full synchronization before starting the applications on the system being synchronized. This is because SRDF knows where valid data resides. As the application proceeds, SRDF will preemptively move data from the current R1 system to the system being synchronized if the application requests that data before it has been moved by the resynchronization process.

System Splitting

The redundant storage-array configuration can be split so that the backup site can be used for other purposes. To do this, processing at R1 is quiesced; and in-progress activity is allowed to complete. Following the next replication cycle, R2 is now completely synchronized with R1. At this point, replication is terminated; and R2 can be used for other purposes such as fielding large queries, generating reports, or backing up the database. Since the R2 database is consistent, it can support write activity as well. However, any changes made to R2's database will be lost when R2 is resynchronized with R1.

When R2 processing has been completed, it is returned to service as the backup for R1. R1 processing is quiesced, R2 is synchronized with R1, replication is initiated, and normal production is restored.

Eliminating Planned Downtime

System splitting can also be used to eliminate planned downtime for system maintenance such as hardware or software upgrades and configuration changes. Either system can be upgraded first. For instance, R2 can be taken offline and upgraded. After it has been thoroughly tested, it can be synchronized with R1.

A failover to R2 can now be made, promoting it to R1 to continue processing. The production system can be upgraded and then resynchronized with the backup system. At this point, production processing can be transferred back to the production system.

In an SRDF/Star configuration, when the local synchronous node, $R2_S$, is promoted to R1 so that the production node can be upgraded, $R2_S$ must establish a temporary R1-R2 relationship with the remote asynchronous node, $R2_a$, to continue disaster protection.

Other Configurations

SRDF supports many other configurations:

- Volumes in the storage array can be individually configured to use either SRDF/A or SRDF/S.

- A source system can feed many target systems, thus making data available to other applications such as data warehouses. Database volumes can be selectively replicated over specific communication channels to targeted remote systems.

- Many source systems can feed a single target. In this way, a single target system can back up multiple source systems.

- Bidirectional replication is supported. A node may be an R2 node to provide backup for a set of remote applications, and it can be an R1 node serving its own applications and replicating to a remote node for backup.

- SRDF Consistency Groups (SRDF/CG) ensure consistency of data spread across multiple storage arrays. When SRDF/CG detects any write to a volume that cannot communicate with its remote mirror, it suspends the remote mirroring for all volumes defined in the consistency group.

- Cascaded SRDF allows a storage array to be both a synchronous R2 and an asynchronous R1 replicating to a remote R2 site.

- Dynamic SRDF allows a storage array to be configured either as an R1 node or an R2 node. This capability is required for SRDF/Star.

- SRDF/Data Mobility (SRDF/DM) provides for the resynchronization of a storage array before returning it to service. It also provides periodic transfer of data for data warehousing or information sharing for decision support.

- SRDF FarPoint allows I/O from multiple logical volumes to be serially transmitted on a single SRDF replication link. This enables the SRDF link to be more fully utilized.

- Source/target switching allows the R1 and R2 roles to be reversed, facilitating simple disaster-recovery testing.

System Management

The Symmetrix Management Console is a browser-based, intuitive user interface for the configuration and management of Symmetrix systems. It is used for the operation and monitoring of SRDF remote mirroring operations. It supports simple storage allocation and administers both open system and mainframe-attached Symmetrix systems.

The Management Console provides health indicators, cycle time, and throughput of the SRDF facility at user-specified polling intervals. It features multiple levels of threshold alerting.

The Symmetrix Performance Analyzer provides enhanced monitoring of Symmetrix operations using real-time dashboards and heat maps.

Summary

EMC's SRDF storage array replication facility provides consistent target database copies kept current with either asynchronous or synchronous replication. SRDF is used with EMC's DMX-4 massively scalable storage arrays and supports mainframes, AIX systems, and Windows, Linux, and UNIX open systems.

A variety of configurations are supported. Of particular use is SRDF/Star, which provides synchronous replication to a nearby data bunker and asynchronous replication to a remote disaster- recovery site.