

Dell PowerVault MD3600f/MD3620f Remote Replication Functional Guide



THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

© 2011 Dell Inc. All rights reserved. Reproduction of this material in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell.

Dell, the *DELL* logo, and the *DELL* badge, *PowerConnect*, and *PowerVault* are trademarks of Dell Inc. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell Inc. disclaims any proprietary interest in trademarks and trade names other than its own.

June 2011

Abstract

Today's businesses face an ever-increasing amount of data, and protecting this data is no longer a matter of simply copying yesterday's changed files to tape. Critical data changes occur throughout the day, and this data needs to be protected against damage, loss, or lack of availability. Dell MD3600f/MD3620f (MD36X0f) series of storage arrays provide the critical and necessary protection of your data assets.

Dell PowerVault MD3600f/MD3620f arrays (MD36X0f) support the following features for data protection:

- Built-in snapshot feature enables quick recovery of files
- Virtual Disk Copy which clones files or virtual disks for recovery
- The Remote Replication (RR) feature allows the implementation of disaster recovery initiatives and centralization, consolidation and migration of data

This whitepaper explains the functional aspects of Dell PowerVault Remote Replication (RR), its features and functionality.

Table of Contents

Abstract	iii
Dell MD36X0f Remote Replication (RR) Overview	2
Key Benefits of PowerVault MD Remote Replication (RR)	2
Remote Replication Functional Overview.....	2
Remote Replication Concepts	3
Primary and Secondary Virtual Disks.....	3
Primary (Local) Virtual Disk	3
Secondary (Remote) Virtual Disk	3
Fail Over/Role Reversal.....	3
Replication Repository.....	4
Replication Relationships	4
Data Replication Modes	5
Synchronous Replication	5
Asynchronous Replication	7
Asynchronous Replication with Write Order Consistency.....	8
Switching Replication Modes	9
RR Suspend and Resume with Delta Log Resynchronization	10
Database Consistency and Database Hot Backups.....	11
Summary	12

Dell MD36X0f Remote Replication (RR) Overview

Remote Replication (RR) is the MD36X0f array feature that provides the ability to remotely replicate data from one MD36X0f storage array to another. RR is an optional premium feature that requires activation, configuration, and in normal replication operations, administration and control; the latter two tasks do not require lots of time and manpower resources, and they can be automated through scripting.

Key Benefits of PowerVault MD Remote Replication (RR)

- **SAN Based Replication** - Data replication between the primary virtual disk and the secondary virtual disk is managed by the MD36X0f storage arrays and is transparent to host machines and applications.
- **Replication modes** - RR supports *synchronous*, *asynchronous*, and *asynchronous with write order consistency* replication modes. These modes help administrators choose the replication method that best meets protection, distance or performance requirements.
- **Dynamic replication mode switching without suspending the replication** - Users can switch between replication modes at any time. This enables administrators to accommodate changing application and bandwidth requirements without sacrificing protection.
- **Ease of use** - IT administrators can enable replication through the Modular Disk Management System (MDSM) GUI or MDSM CLI. Once enabled, RR functions can be managed using the MDSM GUI, the CLI, or MDSM vCenter plug-in. These tools are available free of charge.
- **Multiple replication relationships** - A replication connection is not limited to a single primary RR system and a single remote RR system. In general, each MD36X0f with the optional RR premium feature installed and activated may be either a primary system or a secondary system, or both. Up to 16 replication relationships are supported.
- **Suspend and resume** - RR provides the ability to suspend replication by explicitly using the Suspend command, and under certain circumstances to automatically suspend replication (consistency group suspension, asynchronous replication when link bandwidth exceeded).
- **Role reversal/fail over** - RR allows reversing the roles of primary and secondary virtual disks to recover from a disaster.
- **Read-only replica access (includes Snapshot creation)** - RR enables the remote data to be utilized prior to a disaster without sacrificing protection of the primary site data.

Remote Replication Functional Overview

Once activated, RR requires that you specify the virtual disks that you wish to replicate, configure the replication by selecting the remote MD36X0f you wish to have the secondary replicated copies on, and then to select and create the secondary virtual disk storage on the remote MD36X0f. Initially, RR will begin operations by copying the entire contents of the virtual disks that you wish to replicate from the primary site's MD36X0f to the remote site's MD36X0f. Once this initial copy operation is complete, active replication begins. For every subsequent write to the primary MD36X0f storage array, there is a corresponding write sent to the remote site MD36X0f and then written to the corresponding remote virtual disk.

RR can be activated and configured at any time; additional virtual disk replication can be configured when necessary.

There are some key, critical design issues regarding remote replication that need to be addressed before designing a DR solution using remote replication.

- Distances involved (primarily the distance between the primary and secondary sites)
- The amount of data required to be replicated
- The recovery objectives should a disaster occur
- The requirements for database applications

For more details about designing a RR solution, please refer to the *Dell PowerVault MD3600f/MD3620f Remote Replication Design Guide*.

Remote Replication Concepts

This section introduces you to primary, secondary, and replication repository, and describes how they interact to replicate data between arrays using the Remote Replication feature.

Primary and Secondary Virtual Disks

When you start remote replication, a replicated virtual disk pair is created and consists of a primary (local) virtual disk on the primary storage array and its replicated pair virtual disk on the secondary storage array.

Primary (Local) Virtual Disk

The primary virtual disk is the virtual disk that accepts host I/O and stores application data. When the replication relationship is first created, data from the primary virtual disk is copied in its entirety to the secondary virtual disk. This process is known as a full synchronization and is controlled by the local MD36X0f storage array and owner of the primary virtual disk. During a full synchronization, the primary virtual disk remains fully accessible for all normal read and write I/O operations. The local MD36X0f storage array is responsible for initiating remote writes to the secondary virtual disk to keep the data on the two virtual disks synchronized.

Secondary (Remote) Virtual Disk

The remote or secondary MD36X0f storage array receives remote writes from the primary MD36X0f storage array and applies these writes to the remote virtual disks. While these secondary virtual disks (replicas) are also regular virtual disks they are not directly accessible for I/O by any hosts: the remote storage array will not accept direct host write requests for those virtual disks acting as remote pairs. This is to prevent unintentional corruption of the data being replicated to these secondary virtual disks. Virtual disks in secondary mode can be mapped to hosts, and are seen as read only disks until later promoted to a primary mode (a role reversal). MD36X0f storage also provides a feature that allows secondary virtual disks to be read through the use of the Snapshot function.

NOTE: Only standard virtual disks may be used in any replicated virtual disk pair (no Snapshot virtual disks). There can be up to 16 defined virtual disk replication pairs per MD3600f storage array.

Fail Over/Role Reversal

During normal RR operations, the remote virtual disks are in target mode and not in primary virtual disk mode. This prevents host access of these virtual disks for normal I/O. In the event of a disaster or a catastrophic failure of the primary site, a role reversal can be performed to promote the secondary

virtual disk to a primary role. Hosts will then be able to read and write to the newly promoted virtual disk and business operations can continue.

Replication Repository

A replication repository is a special virtual disk in the storage array created as a resource for the primary MD36X0f storage array whenever Remote Replication is used. The MD36X0f stores replication information in this repository, including information about remote writes that are not yet complete. The MD36X0f uses this information to recover from controller resets, temporary network outages, and accidental powering-down of arrays. When you activate the Remote Replication premium feature on the array, two replication repository virtual disks are created, one for each controller in the MD36X0f. An individual replication repository virtual disk is not needed for each replication pair. When you create the replication repository virtual disks, you specify the location of the virtual disks. Either you can use existing free capacity, or you can create a disk group for the virtual disks from unconfigured capacity and then specify the RAID level.

Because of the critical nature of the data being stored, the RAID level of replication repository virtual disks cannot be RAID 0 (for data striping). The required size of each virtual disk is 128 MB, or 256 MB total for both replication repository virtual disks of a dual controller storage array.

Replication Relationships

Prior to creating a replication relationship, the Remote Replication premium feature must be enabled and activated on both the primary and secondary storage arrays. The Remote Replication premium feature enables the creation of up to 16 virtual disks in a replication relationship with virtual disks in other Remote Replication enabled MD36X0f arrays. Each virtual disk within the array is replicated in a 1:1 relationship with its corresponding virtual disk and can be either in a primary or secondary role.

The following steps outline the creation of a remote relationship.

1. For use as secondary virtual disks, virtual disks on the secondary storage array must be created on the secondary site if they do not already exist and secondary virtual disk must be a standard virtual disk.
2. The secondary virtual disk candidates must be configured with equal or greater capacity than the associated primary virtual disk.
3. Secondary virtual disk candidates may be of different RAID configurations – it is not required to have the same RAID configuration for both virtual disks in a replicated pair relationship. But, the available capacity must meet the step 2 requirements.
4. When secondary virtual disk candidates are available, a replication relationship can be established in the storage management software (MDSM) by identifying the array containing the primary virtual disk and the array containing the secondary virtual disk, then choosing the specific virtual disk as the secondary virtual disk.
5. The replication relationships can be any of the replication relationships defined in [Figure 1](#).

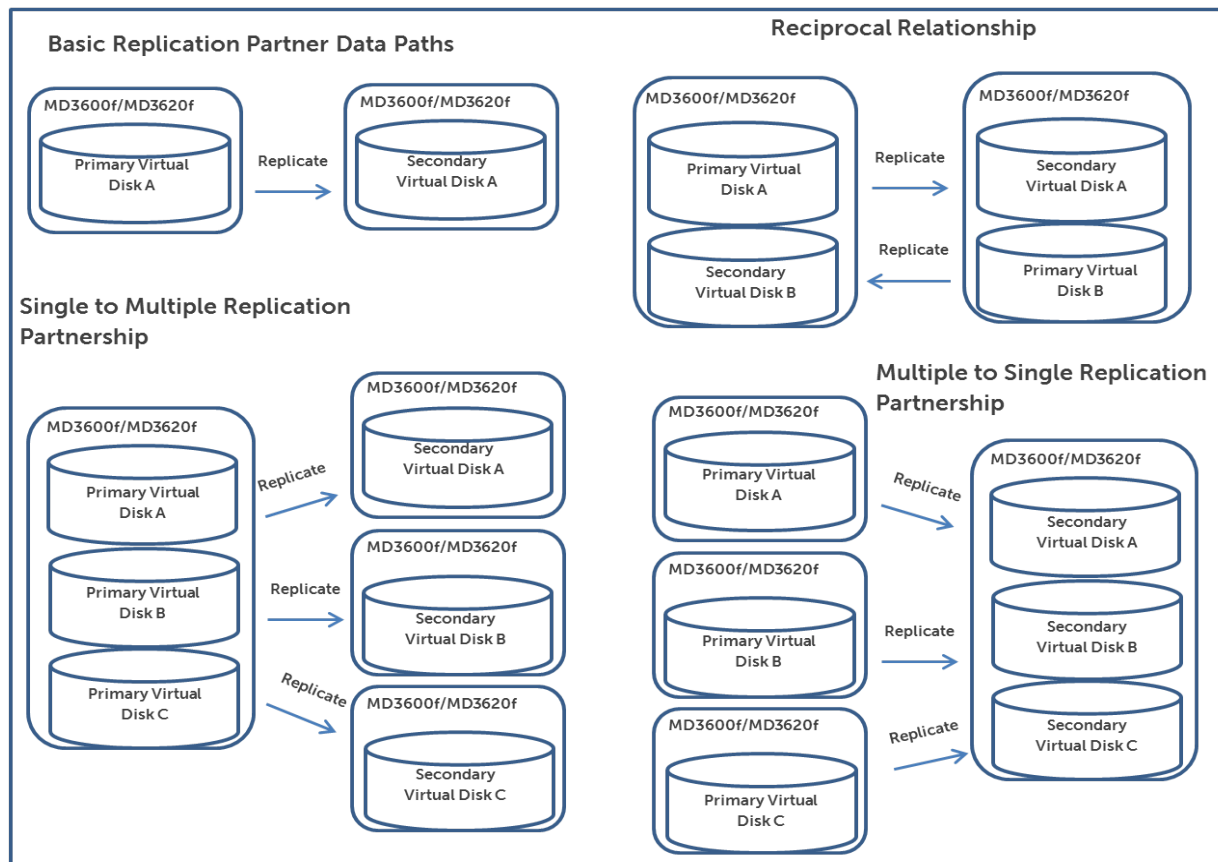


Figure 1: Examples of Replication Relationships

Data Replication Modes

Data replication between the primary virtual disk and the secondary virtual disk is managed by the MD36X0f storage arrays and is transparent to host machines and applications.

This section describes how data is replicated between arrays participating in Remote Replication and the actions taken by the MD36X0f controller of the primary virtual disk if a link interruption occurs between arrays.

The MD36X0f offers three replication modes: Synchronous Replication, Asynchronous Replication, and Asynchronous with Write Order Consistency Replication. Each of these replication modes is designed around a specific set of operational requirements and offers features that may be required of the disaster recovery solution. These replication modes may be dynamically changed from one mode to another.

Synchronous Replication

Synchronous replication is defined as the completion of both the local primary write and the remote write before acknowledgement is made to the host server issuing the I/O. This guarantees that the remote virtual disks are identical to the primary side virtual disks with each and every write and provides the highest level of data protection. Synchronous Replication Mode flow is shown in [Figure 2](#).

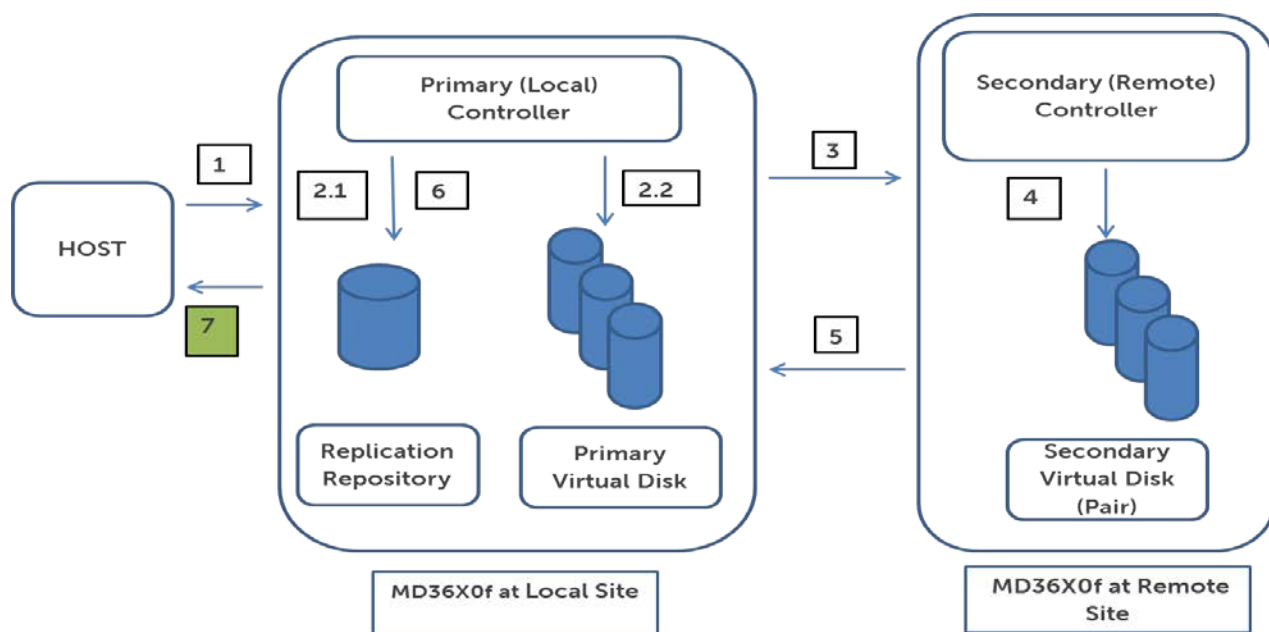


Figure 2: Synchronous Replication Mode

1	Host sends Write Request
2.1	Primary (local) controller at the local site takes in the request, and stores the request in the Replication Repository
2.2	Primary (local) controller writes the data into the primary virtual disk
3	Primary (local) Storage controller then sends the write request to the Remote Storage controller at the Remote Site
4	The secondary controller then processes the write request
5	Once the write request is processed, Secondary Controller sends acknowledgement to Primary Controller
6	Primary controller removes the entry from the Replication Repository
7	Primary controller sends the completion status to Host.

Synchronous replication is usually limited to remote MD36X0f storage arrays located less than 10KM away. This happens to be the fibre channel distances that most equipment supports. The reason for this limitation lies in the fact that the remote write has to complete with each associated local primary write. Extending the distances increases the latency or the time it takes to send the write to the remote MD36X0f storage array. In turn, longer communication time reduces the number of write I/Os that can complete in a given period of time. The benefit is that high IOPs applications with remote replication requirements will not be constrained by the latencies involved in long distance replication.

Asynchronous Replication

Asynchronous replication removes the restriction that the remote write has to complete along with the local primary write. The MD36X0f will separately queue each remote write to allow the local primary write completion to signal Host I/O completion (thus no waiting for the remote write to complete). Instead, the MD36X0f will manage the queue of remote writes and ensure they all complete eventually.

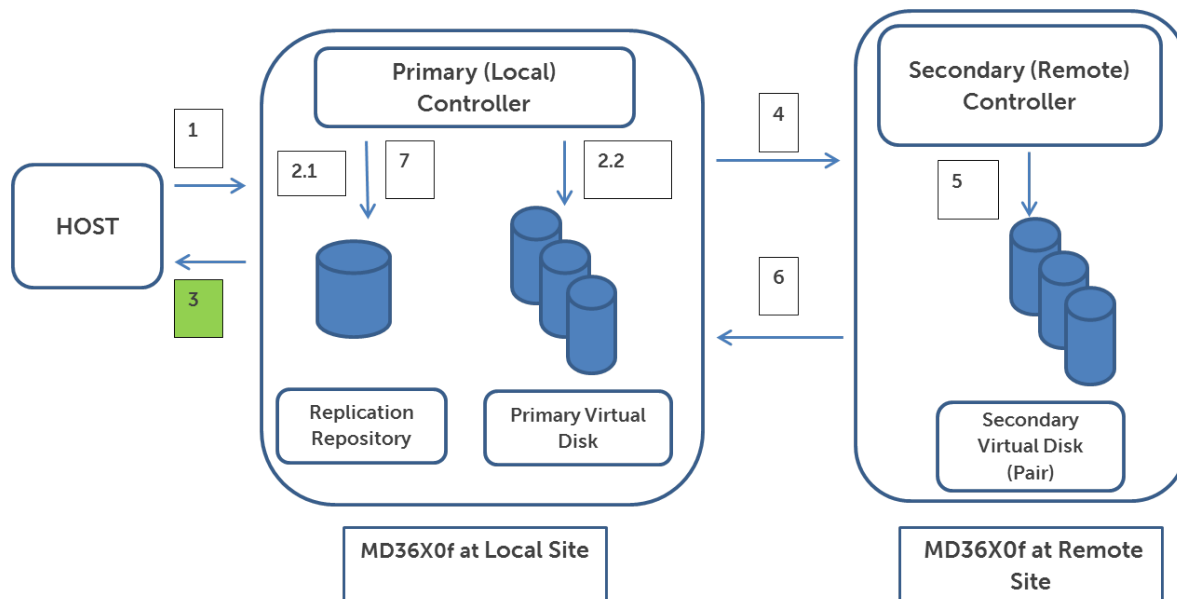


Figure 3: Asynchronous Replication Mode

1	Host sends Write Request
2.1	Primary(local) controller at the local site takes in the request and stores the request in the Replication Repository
2.2	Primary (local) controller writes the data into the primary virtual disk
3	Primary controller sends the completion status to Host
4	Primary controller then sends the write request to the Secondary controller at the Remote Site
5	The secondary controller then processes the write request
6	Once the write request is processed, Secondary Controller sends acknowledgement to Primary Controller
7	Primary controller removes the entry from the Remote Repository

NOTE: The write completion signal is sent in Step 3 as opposed to Step 7 in synchronous mode as shown in [Figure 3](#).

Asynchronous replication is usually required when the remote site distances are greater than 10KM and cannot tolerate slowing the host application I/Os by waiting for remote write completions.

The number of write I/Os supported by asynchronous replication is distance determined. The usual rule of thumb is to allow 1ms for each 100KM in distance. Therefore a 1000KM (621 miles) distance will require 10ms for a round-trip RR I/O and will, therefore, limit the numbers of IOPs to 100. Longer than 1000KM distances will proportionally reduce this IOPs limit. It is imperative that the distance between MD36X0f storage arrays be a critical design factor that is surfaced early in the design and accounted for in the implementation of remote replication.

There are several reasons for highlighting the replication latencies. First, peak period writes (in IOPs) will determine whether RR can keep its replication queue from exceeding the queue limits. If the communications link cannot handle the peak period IOPs, and if the queues are then exceeded, then RR will suspend replication and go into suspend mode. We will take a deeper look into this later.

Asynchronous Replication with Write Order Consistency

For some applications such as database applications, out of order write I/Os will cause database recovery to fail and lead to data inconsistency and data loss. Databases operate under a strict sequencing protocol of writes to tablespaces and to logs. The logs are used to ensure the consistency or correctness of the database transactions as they occur and especially when they complete or commit. The order of the writes to each of the database tablespace objects that reside on virtual disks must also be the same order the remote writes are applied at the remote MD36X0f. This includes the same relative order within the same virtual disk and between the tablespace virtual disks and the log files.

Replication write activity in a synchronous mode guarantees the same write completion order on the remote array, providing the best chance of data recovery from the remote site. With asynchronous write mode, replication write requests are issued in parallel, and therefore not guaranteed to be sent and/or completed in the same order as received by the primary array.

In order to ensure that all writes to the secondary virtual disks are in the same exact order as they were applied to the local primary virtual disks, RR provides an Asynchronous Write Order Consistency Replication (AWOC) mode. The collection of virtual disks configured for write consistency is often referred to as a consistency group. AWOC provides a consistency group in which all associated virtual disks that need remote writes to be completed in the same order as the primary writes occurred are included. RR then ensures that all the writes for this consistency group's virtual disks are sent and written in absolute in-order sequence. Only one write consistency group is allowed per array.

In [Figure 4](#), when host issues writes X, Y and Z to the primary site array, preserved write order consistency ensures write requests are issued to the remote array in the same order as on the local array. So, the remote array receives the write request in the same X, Y and Z order as the primary system did.

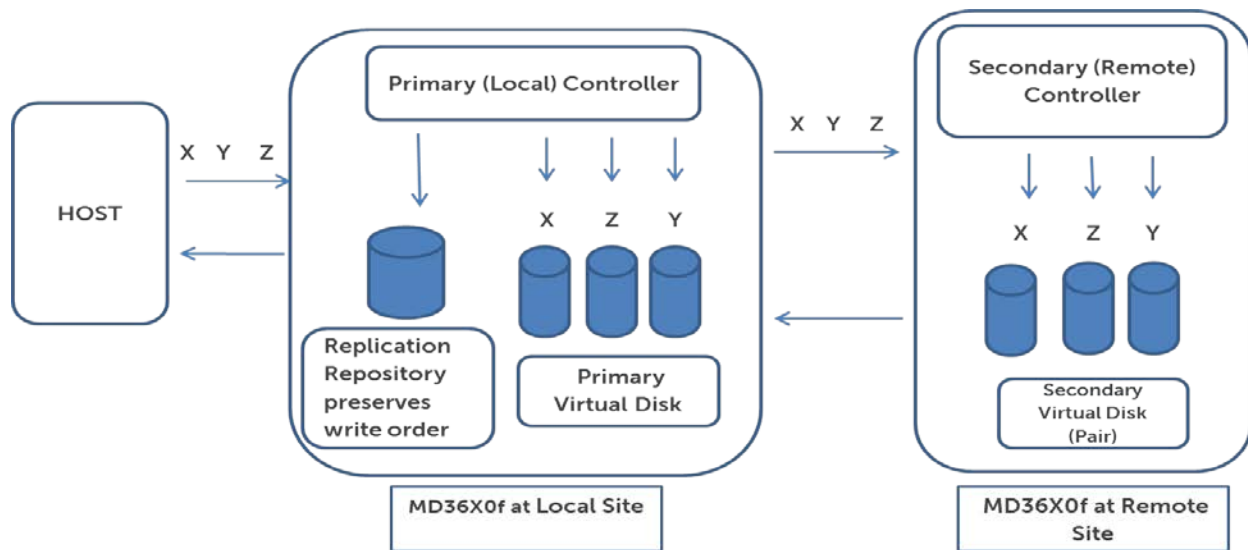


Figure 4: Asynchronous Mode with Write Order Consistency

A second feature of AWOC is that if a single replicated pair goes out of synchronization due to a failure to write to the remote virtual disk, remote write I/Os of the entire consistency group are suspended. This preserves the consistency as of the last remote write. In turn, this guarantees that the virtual disks that constitute the database will be able to be recovered for subsequent use by the remote site applications and servers.

All databases will have this requirement to have inter-virtual disk and intra-virtual disk write I/Os be in absolute primary sequence when remote replication is used.

Switching Replication Modes

RR provides the ability to dynamically switching replication modes. Synchronous replication can be switched to Asynchronous Replication and then to Asynchronous Write Order Consistency replication mode. This order can be reversed and in the specific reverse order. Applications may find this useful to enhance data protection by switching to synchronous mode or to increase IOPs throughput by switching away from synchronous replication mode.

An additional requirement if the remote replication communications is bridged to TCP/IP is that all equipment from one end to the other must have the *in-order packet delivery* option turned on. Due to the nature of TCP/IP networks being routable, it is entirely possible that two sequential I/Os arrive out-of-order. In order to eliminate this possibility, the in-order-packet delivery must be turned on for all bridges, routers, and switches that are configured for RR using AWOC.

With throughput and latency being critical factors in meeting peak period I/O requirements, it is also recommended that the *jumbo packets* option, if available, be turned on for all equipment end-to-end. This increases the communications link utilization by using larger Ethernet packets to contain the fibre channel data payloads.

RR Suspend and Resume with Delta Log Resynchronization

It is important to be able to size the communications link for throughput and for IOPs. Peak period measurement for the application's set of virtual disks is necessary in order to determine the bandwidth required for remote replication. Peak period IOPs measured will determine whether the intended distance of the remote site can support the peak IOPs. Additional factors that impact the peak period requirements include future growth needs, the need to share a single communications link, and whether there is a need to use redundant communications links for safety.

In the event that peak period write traffic exceeds the throughput and/or the IOPs capability of the communications links, it will be necessary to consider using RR's Suspend and Resume features.

RR provides the ability to suspend replication by explicitly using the Suspend command and under certain circumstances, to automatically suspend replication (consistency group suspension, asynchronous replication when link bandwidth exceeded). This allows RR to continue to capture all write I/Os but not use the overloaded communications links until a later time when the peak period traffic eases.

Asynchronous replication will automatically suspend when the peak period traffic exceeds the communications link capability (after exceeding the RR asynchronous queue).

AWOC will not automatically suspend when the peak period traffic exceeds the communications link capability. Instead host I/O will be slowed down to keep the peak period traffic below the link capability. This preserves data integrity of the remote virtual disks without resorting to RR suspension.

Suspending replication will create a delta log for each affected virtual disk at the time of suspension. This delta log tracks the missing write I/O to the remote virtual disk for the duration of the suspension and marks a data range for the write and will use this delta to sequence writes to that remote virtual disk when RR resumes.

Resume will cause RR to resume replication for the suspended virtual disk. The delta log will be read front to back from the primary virtual disk for all the missing I/Os in the secondary virtual disk and RR will send these missing I/Os to the secondary MD36X0f. Once the delta log is completely processed, RR resumes active replication of the primary virtual disk.

It is important to note that the delta log reads will be larger than the original write I/O due to the fact that the delta log is composed of data ranges that are often larger than I/Os. Because the delta log is fixed in size, larger virtual disks will have larger data ranges. This means the resumption time may be longer than it would have taken if RR did not suspend.

If the communications link is sufficiently undersized (for example, due to peak period traffic growth), there is the possibility that the RR resume operation may never complete. The assumption for the resume operation completing is that there is sufficient time and bandwidth to process all missing writes in the delta log before the next peak period occurs. This may not be the case if the peak period traffic continues or if the non-peak traffic prevents emptying the delta log.

It is important to size the communications link to handle current workloads, future growth, and RR resume operations.

Database Consistency and Database Hot Backups

After suspending AWOOC for database virtual disk replication, and, if a disaster occurs while resuming RR then the remote database virtual disks are not recoverable. This is due to the delta log processing of missing I/O in non-original write order. Once RR's resume operations are complete, the virtual disk pairs are once again in synchronization, and the remote database virtual disks are recoverable in disaster recovery processing.

In order to have usable database virtual disks in the event that a disaster occurs in the middle of a resume operation, it is important to snapshot the remote database virtual disks at the very instant that they are suspended and before the replication is resumed and uses the delta log to re-synchronize. This requires the auto-resume feature to be turned off and a snapshot is taken at the time of suspension.

Once the resume operation completes, a second snapshot is taken. This provides the ability to recover from a primary site disaster. If a disaster occurs during resume operations, the first snapshot is used for recovery. Once the resume operation completes, the second snapshot is used for recovery.

In both cases, either snapshot will be database consistent and usable for database recovery in the event of a disaster. We can achieve this using Database Hot Backups.

All databases provide a hot backup operation that allows a storage array to snapshot the database virtual disks to create a point-in-time copy for backup and/or recovery purposes.

Database hot backups perform a number of functions that ensure recoverability using snapshots or copies of the database virtual disks. First, the database internal data (pointers, buffers, tablespace data, and recovery information) are all persisted to the database virtual disks. Enough information is therefore stored to ensure a successful recovery using the database virtual disks including the important logs.

Disaster Recovery using RR will use this same hot backup operation to create a remote snapshot of the database virtual disks. In essence, the remote snapshot is identical to the local snapshot.

During this hot backup window we will use RR's suspend feature to flush the data written on the primary site virtual disks to their replication counter parts on the remote system; we can then take a snapshot of the remote virtual disks to create a recoverable set of database virtual disks. RR will then resume and the database will exit hot backup mode.

NOTE: Periodically performed, this process will create recovery snapshots (also called Recovery Points) which can be used as remote backup images, or to reprovision the database itself (a second copy residing at the remote site), and in the event of a disaster and the current set of virtual disks are not recoverable, the last recovery snapshot will be used instead.

Summary

Dell PowerVault MD36X0f Remote Replication option provides a robust set of features to implement disaster recovery solutions. Features include SAN based replication between multiple MD36X0f storage arrays in any combination of primary or secondary replication roles and dynamic replication mode switching. For critical database applications, Asynchronous Write Order consistency replication provides consistency group support for database applications. With full control over replication including suspend, resume, and role reversal commands, RR solutions handle the contingencies necessary for a Disaster Recovery solution.