

GUIDE TO MULTISITE DISASTER RECOVERY FOR VMWARE VSPHERE ENABLED BY EMC SYMMETRIX VMAX, SRDF, AND VPLEX

A Detailed Review

EMC GLOBAL SOLUTIONS

Abstract

This white paper offers guidelines for helping customers to choose and set up a configuration for a three-site disaster recovery method for VMware® environments by using EMC® Symmetrix® VMAX™, EMC Symmetrix Remote Data Facility (SRDF®), and EMC VPLEX™.

April 2011

vmware[®]
PARTNER

TECHNOLOGY
ALLIANCE

Copyright © 2011 EMC Corporation. All Rights Reserved.

EMC believes the information in this publication is accurate of its publication date. The information is subject to change without notice.

The information in this publication is provided “as is”. EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

VMware, VMware ESX, VMware ESXi, vMotion, VMware vSphere, and VMware vCenter are registered trademarks or trademarks of VMware, Inc. in the United States and/or other jurisdictions. All other trademarks used herein are the property of their respective owners.

Part Number: H8183

Table of contents

Executive summary	5
Business case.....	5
Use case overview.....	5
Key results/ recommendations	5
Introduction	6
Purpose	6
Scope	6
Audience	6
EMC Global Solutions Proof of Concept program	6
Technology overview	7
Overview.....	7
VMware vSphere	7
vSphere PowerCLI.....	8
EMC Symmetrix VMAX.....	8
Symmetrix Solutions Enabler.....	8
EMC SRDF	9
SRDF/S and SRDF/A	9
SRDF/Star.....	9
EMC VPLEX.....	9
VPLEX Metro	10
Establishing a three-site disaster recovery solution for VMware environments	11
Method 1: EMC Symmetrix VMAX and SRDF/Star.....	11
Advantage.....	11
Disadvantage	11
Method 2: EMC VPLEX and SRDF/A.....	11
Advantage.....	11
Disadvantage	11
Using VMware vCenter Site Recovery Manager for disaster recovery.....	12
VMware vCenter Site Recovery Manager.....	12
EMC SRDF/Star environment	13
Star environment	13
Data flow in a star environment.....	14
Storage failover in a star environment.....	14
Storage failback in a star environment.....	15

VPLEX with SRDF/A environment.....	16
VPLEX environment	16
Cluster detach	17
Best practice	17
Supported VMware HA configurations	17
Data flow with VPLEX/SRDF/A environment.....	18
Storage failover in a VPLEX/ SRDF/A environment	19
Storage failback in a VPLEX/ SRDF/A environment	20
Raw device mapping	21
Raw device mapping considerations	21
SRDF example	21
Placing and configuring mapping files to allow seamless failover between sites with SRDF/Star....	23
Placing and configuring mapping files to allow seamless failover between sites with VPLEX and SRDF/A	25
Mounting datastores in alternate locations	27
Adding virtual machines.....	28
Adding virtual machines to inventory on alternate sites	28
Adding virtual machines through PowerCLI cmdlets	28
Allowing for virtual machine power-on sequencing	28
Conclusion	30
Summary	30
Findings.....	30
References.....	31
Product documentation.....	31
Other documentation.....	31

Executive summary

Business case

In traditional two-site recovery, the standard approach is to use some form of synchronous replication between sites. This ensures that any writes made to storage on the production site are replicated to the remote site before the production host receives an acknowledgment that the write is complete.

Should the production site cease to function completely, all input/output (I/O) committed on the primary site is also committed on the secondary site. Anything that was “in flight” on the primary site is understood by the host to be incomplete, so no completed transactions can be lost between failing on the primary site and recovering on the secondary site.

The impact of writing I/O this way is that write latencies are increased due to the need to replicate the I/O to another site and receive acknowledgment before confirming to the host. This can limit the practical distance between the primary and secondary sites (usually to the range of 60 to 100 km based on latencies).

The distance limitation is one of the reasons that some customers will consider an additional (tertiary) site, which is normally replicated asynchronously to increase the distance between sites.

Use case overview

This use case offers customers guidance when choosing a configuration for a three-site disaster recovery method for VMware® environments by using EMC® Symmetrix® VMAX™, EMC Symmetrix Remote Data Facility (SRDF®), and/or EMC VPLEX™.

Key results/ recommendations

This white paper enables you to:

- Understand the configurations required to establish a three-site disaster recovery environment for VMware infrastructure
- Understand the differences between two different approaches and how these affect:
 - Recovery method
 - Raw device configuration
- Understand how to automate the recovery of virtual machines in alternate locations

Introduction

Purpose

This white paper demonstrates two methods of establishing a three-site disaster recovery solution for VMware environments. Each method has its own unique advantages and disadvantages.

Scope

This white paper provides guidelines for two methods of establishing a three-site disaster recovery solution for a VMware environment. This white paper does not provide any testing or validation results, but does provide detailed logical configuration information.

Audience

This white paper is intended for EMC employees, partners, and customers who want to establish a three-site disaster recovery solution. Readers should have at least a basic understanding of private cloud technology and disaster recovery as well as knowledge about the main components being described in this white paper.

EMC Global Solutions Proof of Concept program

This white paper is based on a proof of concept that was executed within the EMC Global Solutions Labs.

The Proof of Concept program within EMC provides the framework to enable customers to define the test and validation criteria that is critical for them in moving forward with an EMC proposal.

A proof of concept provides EMC with the opportunity to:

- Address a customer's specific questions around a proposed solution
- Define the technical scope that will allow EMC to validate the proof points within the context of the IT architecture that is under review

The program enables EMC to establish its credibility and expertise, and to give customers the confidence to move forward with the solution.

Technology overview

Overview

The main components used in the two methods described in this white paper for building a three-site disaster recovery configuration are:

- VMware vSphere™ 4 (“vSphere”)
- EMC Symmetrix VMAX
- EMC SRDF
- EMC VPLEX

VMware vSphere

You can offer the highest levels of availability and responsiveness with vSphere (shown in Figure 1), the industry’s most complete and robust virtualization platform. The most demanding data centers around the world depend on vSphere, virtualizing business-critical applications with dynamic resource pools for unprecedented flexibility and reliability.



Figure 1. VMware vSphere 4

With vSphere, you can:

- Turn your data center into a flexible cloud infrastructure.
- Use existing assets and applications while offering self-service deployment and provisioning through virtualization.
- Create a private cloud and deliver IT infrastructure as an easily accessible service.
- Automate disaster recovery plans with disaster recovery management.
- Reduce your capital and operating costs and increase control over IT infrastructures while preserving the flexibility to choose any OS, application, and hardware with vSphere.

Available in several different editions, vSphere delivers targeted benefits to small business, mid-size, and enterprise business customers.

vSphere PowerCLI

VMware vSphere PowerCLI is a command line utility that can be installed on any Microsoft Windows host and used to script many of the operations required to manage a VMware environment.

Unlike commands issued at the command line of a vSphere host, the vSphere PowerCLI commands can be targeted at a VMware vCenter™ Server instance. This capability allows the commands to be issued in the context of the entire VMware infrastructure.

All of the discover, add, power-on, and answer prompt tasks can be completed by using vSphere PowerCLI cmdlets.

EMC Symmetrix VMAX

The EMC Symmetrix VMAX system features a revolutionary Virtual Matrix Architecture™. This system architecture builds upon the rich heritage of the Symmetrix multi-controller platform and extends the value of the Direct Matrix Architecture® to deliver unprecedented performance, availability, and functionality at a reduced cost. The unique scale-out architecture of the Symmetrix VMAX system provides the foundation to scale to hundreds of petabytes of capacity. This supports thousands of servers and millions of IOPS delivered to that capacity. They are all flexibly deployed throughout a virtual IT environment and able to be controlled through a single screen.

The Symmetrix VMAX system architecture provides a Virtual Matrix™ that can scale beyond the confines of a single system footprint. The core element of the Virtual Matrix is the Symmetrix VMAX Engine, which includes caches, front-end connectivity, and back-end connectivity.

Symmetrix Solutions Enabler

Symmetrix Management Console (SMC) is a powerful and intuitive GUI that configures and manages multiple Symmetrix arrays. SMC presents the functionality of the Symmetrix Solutions Enabler command line interface in a browser interface, simplifies storage administration tasks through the use of built-in wizards, and provides the ability to:

- Manage Symmetrix access controls, user accounts, and permission roles
- Discover Symmetrix arrays
- Perform configuration operations
- Manage devices
- Perform and monitor replication operations
- Monitor alerts
- Monitor an application's performance
- Configure and manage Fast and Symmetrix Optimizer

EMC SRDF

The EMC SRDF family of remote replication software is the most field-proven, widely deployed, array-based disaster restart solution in the world. By using the industry-leading, high-end Symmetrix system, SRDF offers the most choice and flexibility to meet any service-level agreement.

SRDF provides remote data replication, independent of the host and operating system, application, and database. SRDF helps companies manage planned and unplanned outages, enabling 24x7x365 data availability. It allows businesses to focus on maximizing revenue-generation and customer-support opportunities, improve productivity, and control or reduce costs for increased competitive advantage.

SRDF/S and SRDF/A

SRDF consists of several options including advanced three-site capabilities that use SRDF/Synchronous (SRDF/S) and SRDF/Asynchronous (SRDF/A). SRDF offers the most comprehensive portfolio of remote replication solutions in the industry.

- SRDF/S maintains real-time synchronous remote data replication from one Symmetrix production site to one or more Symmetrix systems located within campus, metropolitan, or regional distances and provides a recovery-point objective of zero data loss.
- SRDF/A maintains asynchronous data replication that is usually at extended distances and provides a recovery-point objective that could be as minimal as a few seconds.

SRDF/Star

SRDF/Star provides a three-site, disaster recovery solution. This architecture protects business data against a primary site failure or a regional disaster. SRDF/Star uses concurrent SRDF or cascaded SRDF capability to mirror the same production data synchronously to one remote site and asynchronously to another remote site:

- The workload site is the primary data center where the production workload is running.
- The synchronous target site is a secondary site usually located in the same region as the workload site. The production data is mirrored to this site by using synchronous replication.
- The asynchronous target site is a secondary site in a distant location. The production data is mirrored to this site by using asynchronous replication.

EMC VPLEX

EMC VPLEX is the next-generation solution for information mobility and access within, across, and between data centers. It is the first platform in the world that delivers both local and distributed federation.

- Local federation provides the transport cooperation of physical elements within a site.
- Distributed federation extends access between two locations across distance. EMC VPLEX Metro is a solution for federating both EMC and non-EMC storage.

VPLEX storage federation provides an extensive offering of new features and functionality for the era of private cloud computing:

- Achieve a recovery-point objective as minimal as one second with SRDF/A in the event of a regional disaster.
- Establish continuous remote data mirroring across three sites and maintain high availability in the event any of them goes offline.

As VPLEX removes physical barriers and enables users to access a single copy of the data at different geographical locations, it also enables geographically stretched virtual and physical host clusters. This enables transparent load sharing between multiple sites while providing the flexibility of relocating workloads between sites in anticipation of planned events.

Built on a foundation of scalable and highly-available processor engines, VPLEX is designed to seamlessly scale from small to large configurations.

With a unique scale-up and scale-out architecture, VPLEX advanced data caching and distributed cache coherency provide workload resiliency, automatic sharing, and balancing, and failover of storage domains, and enable both local and remote data access with predictable service levels.

VPLEX Metro

VPLEX Metro with AccessAnywhere lets you seamlessly relocate data between two sites within synchronous distances. The combination of virtual storage with VPLEX Metro and virtual servers lets you transparently move virtual machines across a distance.

Establishing a three-site disaster recovery solution for VMware environments

Method 1: EMC Symmetrix VMAX and SRDF/Star

Method 1 uses EMC Symmetrix VMAX with a combination of SRDF/S and SRDF/A technologies in a star configuration to provide a consistent and protected version of the data across the three sites.

This environment includes three distinct VMware vCenter Server instances, with recovery between sites achieved by manipulating the SRDF relationships and virtual machine recovery through scripts on the alternate sites.

Advantage

Using SRDF and star configuration to protect your data enables you to seamlessly protect your VMware data across three sites without the need for any full resynchronizations during failback to normal operation.

Disadvantage

Each site must operate as a completely separate VMware vCenter instance. Therefore, no automatic recovery is possible between sites.

Note: Recovery can be automated through a script but is not automatically executed.

Method 2: EMC VPLEX and SRDF/A

Method 2 uses EMC VPLEX to protect data between primary and secondary sites by using Distributed Virtual Volumes and Distributed Cache Coherency. The tertiary site is enabled by SRDF replication of data from either the primary or secondary site to the tertiary site.

This environment uses two VMware vCenter Server instances:

- The first instance manages the primary and secondary sites and uses VMware vMotion® or VMware High Availability (HA) to recover between the primary and secondary sites.
- The second instance is located on the tertiary site and acts independently of the first instance. Recovery on the tertiary site is achieved by manipulating the SRDF relationships and virtual machine recovery through scripts.

Advantage

When using a combination of VPLEX and SRDF to protect across three sites, it is possible to automatically recover the VMware infrastructure by using VMware HA between the primary and secondary sites (those operating through the VPLEX).

Disadvantage

Because the tertiary site is operated by SRDF (and not integrated with VPLEX), failbacks from the tertiary site require that one leg of the VPLEX distributed virtual volumes undergoes a full resynchronization from the SRDF primary volume (R1).

Using VMware vCenter Site Recovery Manager for disaster recovery

VMware vCenter Site Recovery Manager

The standard recommendation for VMware disaster recovery is to use VMware vCenter Site Recovery Manager (VMware SRM), along with some form of array-based replication and its associated Storage Replication Adapter (SRA). This SRA integrates with the VMware SRM infrastructure and controls the storage operations required when a failover is required from one site to another.

The VMware SRM framework allows for granular control over operations such as power-on sequences, as well as virtual machine customization, making the calls to the SRAs installed at the appropriate times to control storage operations.

However, the VMware SRM infrastructure is designed to work in the context of two sites only, or more accurately, two VMware vCenter Server instances. In situations where users have three sites and want to protect themselves against rolling site failures, VMware SRM is not sufficient. In these cases, alternative approaches are required.

EMC SRDF/Star environment

Star environment Figure 2 shows the configuration required to achieve multisite failover for vSphere when using SRDF/Star. Each site has its own independent vCenter Server instance and set of vSphere hosts. Virtual machines on any given datastore can be running on one site at any moment in time.

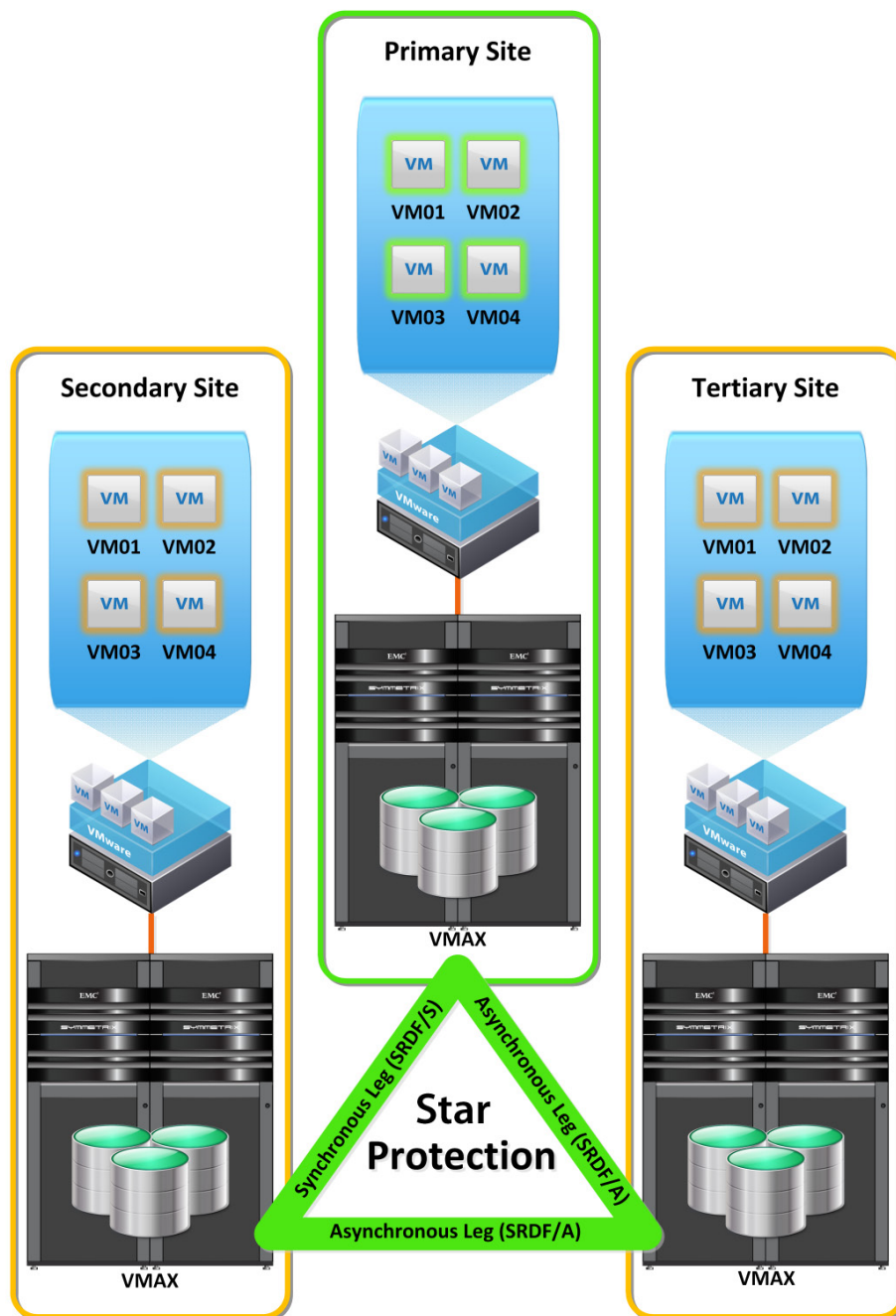


Figure 2. Star environment

Data flow in a star environment

SRDF/Star controls the flow of data from the primary site to the secondary and tertiary sites. In the event of a site failure, SRDF/Star retains protection by maintaining data consistency between the remaining two sites.

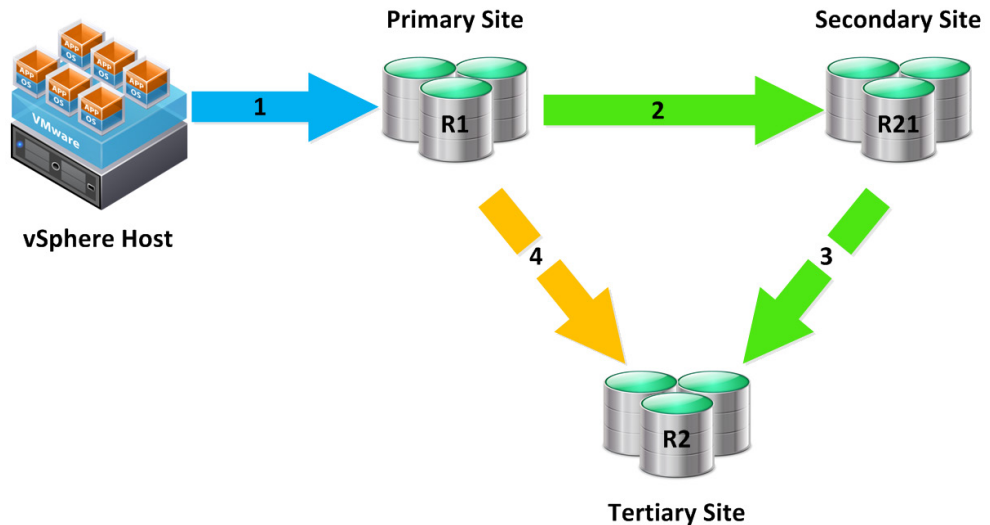


Figure 3. Data flow in a star configuration

Figure 3 shows normal production data flow in a star configuration. The data is available and writable from one site only (primary). Data flows as described in the following steps:

1. Host write is made to primary storage (R1).
2. Write is replicated synchronously to secondary site (R21) before host receives acknowledgment that write is complete.
The R21 annotation reflects the fact that the volume on the secondary site has two personalities. It is an R2 in relation to the R1 on the primary site, and an R1 in relation to the R2 on the tertiary site.
3. Write is asynchronously replicated from secondary site to tertiary site (R2).
4. Asynchronous leg between primary and tertiary site is passive, but maintaining track tables so that this leg can be activated in the event that normal replication to secondary or tertiary site is interrupted.

Storage failover in a star environment

If all legs/arrays of the replication are still available, then it is possible to fail over to an alternate site and reconfigure the star replication to maintain full protection.

Should one of the legs/arrays be unavailable, star replication can be reconfigured to maintain maximum possible site resilience. All synchronizations during a reconfigure will be incremental as the star is maintaining data consistency across all three sites.

EMC Solutions Enabler is a command-line tool that can be used to manipulate storage operations on Symmetrix VMAX units. SRDF and star operations can be executed through the Solutions Enabler by using the `smrdf` and `symstar` commands.

The exact sequence of Solutions Enabler `symstar` commands required to fail over and reverse or reconfigure the replication depends on:

- The nature of the failure
- The copy of the data to be used for continued operations
- The direction and nature of the replications required to the remaining available sites

Once the storage in the preferred location is write-enabled, you can recover the virtual machines on that site.

For exact detail on the commands required to fail over or reconfigure the replication, see the *Symmetrix Customer Procedure Generator* utility available on Powerlink. If you do not have access to this content, contact your EMC representative.

Storage failback in a star environment

Failbacks after a site move or loss of site availability are incremental in terms of data needing to be resynchronized to the other locations. The only exception to this is a destructive loss of site, which requires sending a full copy to the replacement storage on the rebuilt site.

As before, the exact sequence of Solutions Enabler commands required to fail back and reverse or reconfigure the replication depends on the nature of the failure and the method or site on which the customer chooses to continue operations.

Once the storage in the desired location is write-enabled, you can recover the virtual machines on that site.

For exact details on the commands required to fail back or reconfigure the replication, see the *Symmetrix Customer Procedure Generator* utility available on Powerlink. If you do not have access to this content, contact your EMC representative.

VPLEX with SRDF/A environment

VPLEX environment Figure 4 shows the configuration required to achieve multisite failover for VMware vSphere when using a combination of VPLEX and SRDF/A. The primary and secondary sites share a vCenter instance. The vSphere hosts on the two sites can be part of the same vSphere cluster.

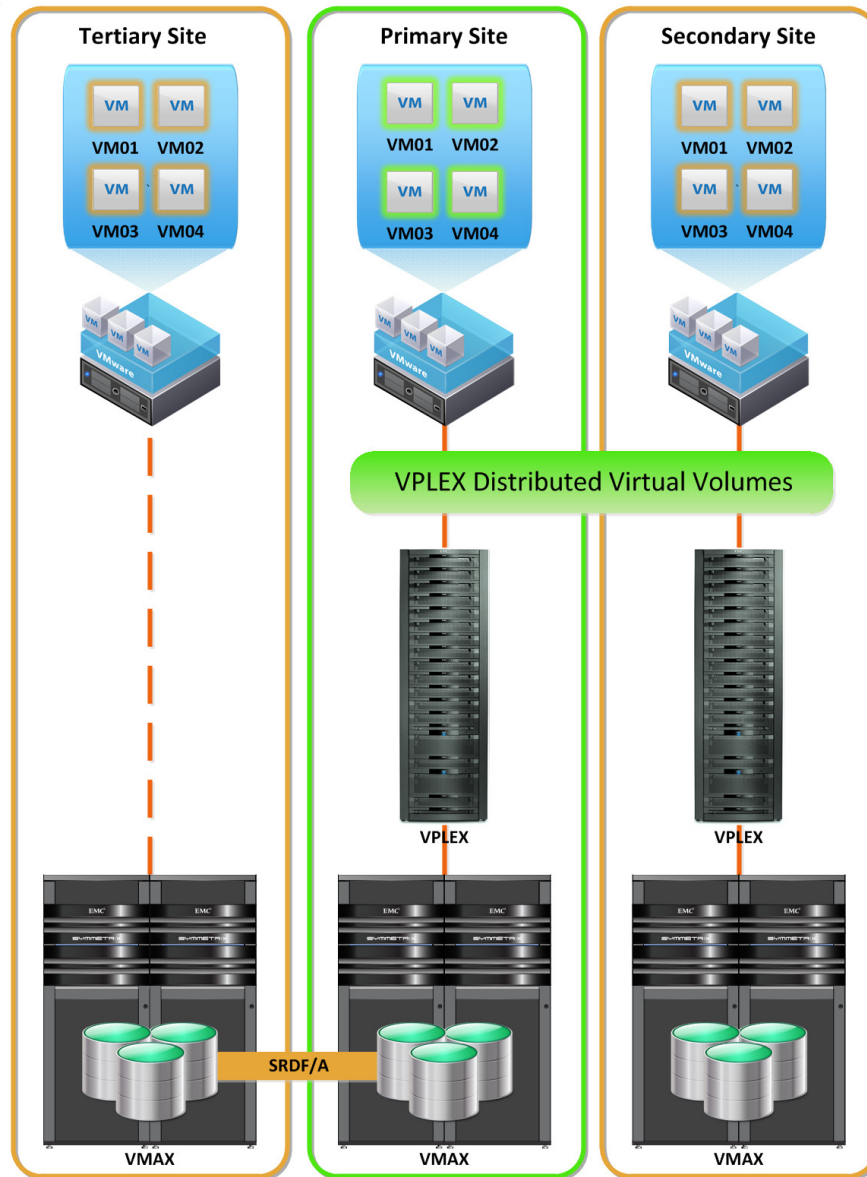


Figure 4. EMC VPLEX environment for multisite failover

In normal operation, virtual machines on any given datastore can run on either the primary or secondary site at any moment in time. In this instance, VMware HA was used to control the failover between the primary and secondary sites. Refer to [Supported VMware HA configurations](#) to fully understand the configuration guidelines, supported configurations, and implications of using VMware HA in this context.

Cluster detach

There is a setting within VPLEX known as “cluster detach”. The cluster detach rules are configured per distributed virtual volume and determine what happens in the event that the two VPLEX units lose connectivity. In that scenario, one site or the other needs to maintain ownership of the volumes. If the cluster detach setting is configured to be the primary site, the primary site will be assigned ownership and continue to run the volumes. To avoid a “split brain” scenario, the secondary site suspends I/O to the leg of the distributed virtual volume on that site.

In this configuration, configure the VPLEX cluster detach setting to be the same site as the primary SRDF (R1) volume. This way, if connectivity to the secondary site is lost, data is still replicated from the primary to the tertiary site through SRDF.

Best practice

VPLEX best practice is to run your virtual machines on the site that assumes ownership of the volumes. This ensures that the virtual machines continue to operate without interruption, should a cluster detach scenario develop.

For more information, see the VMware Knowledge Base article [Using VPLEX Metro with VMware HA](#).

Supported VMware HA configurations

When using VMware HA and VPLEX together to form a distributed VMware ESX™ Cluster, there are a strict set of configuration guidelines to which you must adhere:

- The maximum round-trip latency on both the IP network and the Fibre Channel network between the two sites must not exceed 5 milliseconds (ms). The IP network supports the VMware ESX/ESXi hosts and the VPLEX Management Console, and the Fibre Channel network is required by inter-cluster links connecting the two VPLEX clusters within VPLEX Metro.
- The ESX servers in both data centers must have a private network on the same IP subnet and broadcast domain.
- Any IP subnet used by the virtual machine that resides on it must be accessible from ESX servers in both data centers. This requirement is important so that clients accessing virtual machines that are running on ESX servers on both sides are able to function smoothly upon any VMware HA-triggered virtual machine restart events .
- The data storage locations that include the boot device used by the virtual machines must be active and accessible from ESX servers in both data centers.
- VMware vCenter Server must be able to connect to ESX servers in both data centers.
- The VMware datastore for the virtual machines that are running in the ESX cluster are provisioned on distributed virtual volumes.
- The maximum number of hosts in the HA cluster must not exceed eight hosts, with four hosts on each site.

Based on the VPLEX cluster detach rules discussed earlier, it is important to understand the impact of choosing to run a virtual machine on the non-preferred site (that is, the site that will lose that arbitration in the event of a cluster detach occurring). If a failure occurs which disrupts communication between the VPLEX units,

then virtual machines running on the non-preferred site for a Distributed Virtual Volume will no longer be able to complete I/O successfully. This may cause the virtual machine to stop responding or fail.

For full details, see the VMware Knowledge Base article [Using VPLEX Metro with VMware HA](#).

Data flow with VPLEX/SRDF/A environment

Figure 5 shows the normal production data flow when VPLEX is being used in combination with SRDF/A to maintain three-site protection for VMware.

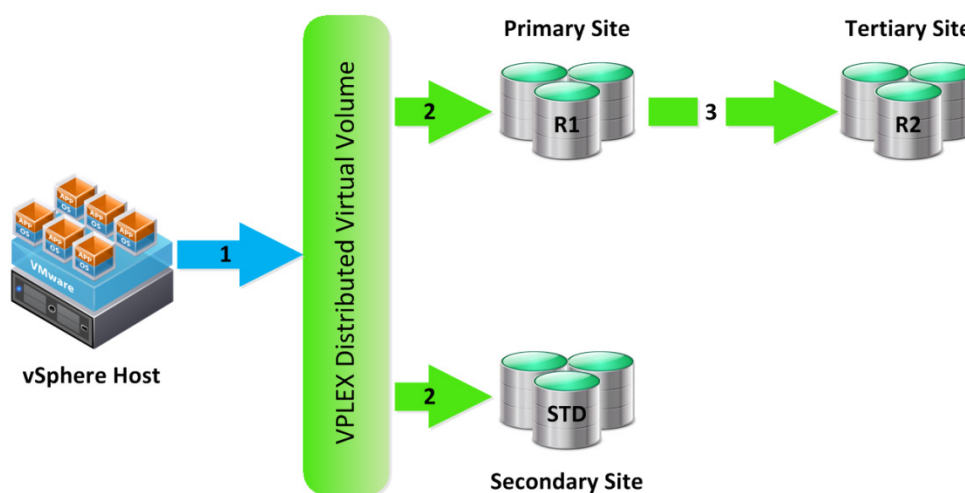


Figure 5. Data flow in a VPLEX/SRDF/A configuration

VPLEX controls the synchronization between the primary and secondary sites. SRDF/A controls the data flow of data to the tertiary site.

One leg of the VPLEX distributed virtual volume is configured as an SRDF primary volume (R1). This volume is responsible for maintaining data consistency between the primary site and the tertiary site. In Figure 5, there is no direct consistency between the secondary site and the tertiary site.

Data flows as described in the following steps:

1. Host write is made to the VPLEX distributed virtual volume.
2. This write is distributed and written to the individual legs of the virtual volume.
3. On the primary site, the leg of the virtual volume is also an SRDF R1, so this write is propagated to the R2 on the tertiary site.

If the primary site fails, the tertiary site becomes stale as it no longer receives any updates made while the systems continue to operate on the secondary site. This can be overcome by also making the secondary site an R1 and replicating to a second distinct R2 on the tertiary site as shown in Figure 6.

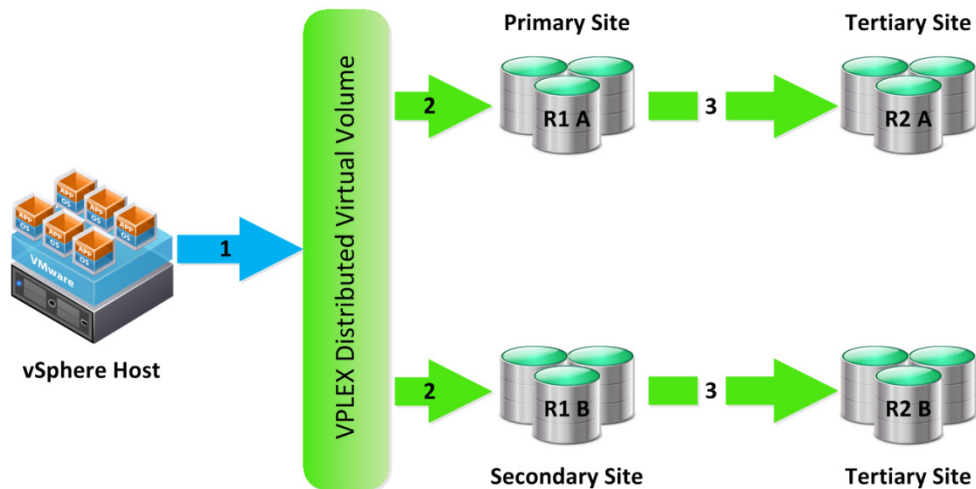


Figure 6. Data flow in a VPLEX/SRDF configuration with primary and secondary R1 sites

While this overcomes the issue of protecting both legs of the distributed virtual volume, it does add some further considerations:

- Four copies of the data are required instead of three copies.
- There are implications with respect to data resynchronizations should production operations be moved to one of the copies on the tertiary site:
 - If production is moved to R2 A, then to resume normal operations, an incremental resynch needs to be completed from R2 A to R1 A by using SRDF.
 - Before the environment is fully consistent again:
 - A full resynch is required between R1 A and R1 B through VPLEX.
 - A full resynch is required from R1 B to R2 B through SRDF.

Storage failover in a VPLEX/SRDF/A environment

VPLEX distributed virtual volumes enable VMware HA functionality across the primary and secondary sites. All hosts in the cluster are accessing the same distributed virtual volumes. VMware HA controls site failover for VMware between primary and secondary locations. No calls to the storage (VPLEX) are required to make this happen.

Failover to the tertiary site involves write-enabling the storage in that location, and adding the required virtual machines to that site's vCenter Server Instance. If a datastore is moved to the tertiary site, all virtual machines on that datastore must be run from the tertiary site (irrespective of whether they originally ran on primary or secondary sites).

The sequence of SRDF Solutions Enabler commands required to fail over to the tertiary site depends on whether source R1 volume is still available. If the R1 is available, then a `failover` command is issued. If the R1 is not available, then a `rw_enable` command is required to resume operations on the tertiary site.

For exact detail on the commands required, see the *Symmetrix Customer Procedure Generator utility* available on Powerlink. If you do not have access to this content, contact your EMC representative.

Storage failback in a VPLEX/SRDF/A environment

Failing back from the tertiary site requires that the primary SRDF volume (R1) be resynchronized from the tertiary site (R2). The resynchronization is incremental, unless the R1 is part of a replacement array installed after a destructive site loss.

However, because the incremental updates from the R2 to the R1 do not pass through the VPLEX infrastructure, you must do the following before failback:

1. Remove the distributed virtual volume from the VPLEX storage views.
2. Disable logging onto the distributed virtual volume.
3. Detach the leg of the distributed virtual volume that is not SRDF.
4. Perform the R2-to-R1 failback or incremental reestablish from R2-to-R1 with SRDF.
5. Add the distributed virtual volume from the VPLEX storage views. At this point, services can be brought online at the site of the R1.
6. Reattach the leg of the distributed virtual volume that is not SRDF. This results in full synchronization between the R1 leg and the standard legs of the distributed virtual volume.
7. Enable logging onto the distributed virtual volumes.

For exact detail on the commands required, see the following resources that are available on Powerlink:

- *VPLEX Procedure Generator utility*
- *VPLEX V4.0 CLI Guide*

If you do not have access to this content, contact your EMC representative.

Raw device mapping

Raw device mapping considerations

Regardless of the method you choose to replicate data from site to site, consider how virtual machines with raw devices will operate.

If no raw device mappings (RDMs) are being used in the environment (that is, only VMFS datastores are used), then recovery of the virtual machines is relatively straightforward from a vSphere perspective. Once the datastores are mounted on the alternate site, the virtual machines (once discovered), should power on with no additional effort. The virtual machine that accesses its storage is dependent on the content of the datastore, not on the identity of the LUN on which the datastore resides.

However, raw devices present a different challenge. Raw devices are, by definition, built upon the unique identifier of the SAN LUN being presented in any given location. Every LUN has a unique ID. Even when data replicated between two LUNs is identical, they both still have unique IDs.

SRDF example

With SRDF, the primary R1 copy of the data will have a unique identifier associated with that LUN. When creating a mapping file for the mapped device, the default is to create it in the same location as the virtual machine. Since that location is replicated to the remote site in order to boot the virtual machine, this means that the mapping file will also be replicated. When you attempt to power on the virtual machines at those locations, it will fail, even if the R2 copy of the mapped devices LUN is available. The mapping file cannot locate a LUN with the identifier it expects.

You can query a mapping file to see the LUN it maps to by using the following command:

```
vmkfstools -q ClusterNode1.vmdk
```

The output of this command will look something like this:

```
Disk ClusterNode1.vmdk is a Passthrough Raw Device Mapping  
Maps to:  
vml.02000900006000097000029260222953303032363453594d4d4554
```

The returned value contains the LUN's unique World Wide Name (WWN - highlighted in red).

This can be confirmed through a Solutions Enabler enquiry against the Symmetrix VMAX LUN as follows:

```
Symdev -sid 229 show 264
```

where 229 is the VMAX ID, and 264 is the LUN ID.

The output looks something like that shown in Table 1:

Table 1. WWN output

Item	Output
Vendor ID:	EMC
Product ID:	SYMMETRIX
Product Revision:	5875
Device WWN:	60000970000292602229533030323634
Device Emulation Type:	FBA
Device Defined Label Type:	N/A
Device Defined Label:	N/A
Device Sub-System ID:	0x0003
Cache Partition Name:	DEFAULT_PARTITION

However, if the mapping file refers to a LUN that is not present, the query fails with an “unable to find relevant LUN” error message.

Alternatively, the EMC Virtual Storage Integrator (VSI) for VMware vSphere plug-in collates this information into a single window.

To avoid failing over and back between sites without continuously recreating the mapping files, create independent mapping files as required on each site on a nonreplicated datastore during initial configuration of the environment.

Every time a new LUN intended for use as a raw device is added to a new or existing virtual machine, this process must be followed.

However, this only needs to be done at the time the new raw device is provisioned. This allows the virtual machine that is accessing that raw device to move around between the three sites without further reconfiguration.

Placing and configuring mapping files to allow seamless failover between sites with SRDF/Star

Figure 7 shows the mapping-file placement in the SRDF/Star configuration.

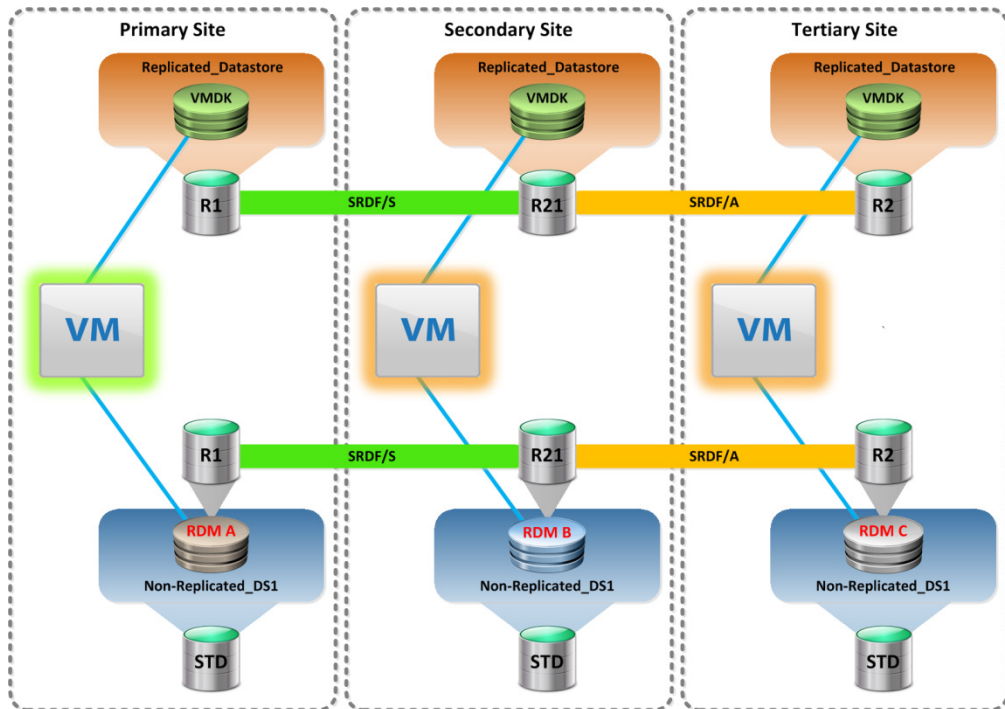


Figure 7. Mapping-file placement in the SRDF/Star configuration

To create this SRDF/Star configuration with RDMs:

1. On the primary site, create a datastore called “Non-Replicated-DS1” on non-replicated SAN storage.
The name of the datastore is not important, but choosing a meaningful name such as Non-Replicated-DS1 provides clarity when viewed through the vCenter Server interface.
2. When adding an raw device to a virtual machine, create the mapping file in another location and specify the “Non-Replicated-DS1” datastore.
3. Create similar datastores on the local SAN storage on the secondary and tertiary sites.
Keeping the name consistent across all sites is important. In this example, it must be called “Non-Replicated-DS1” on all three sites.
4. Create unique mapping files on each of the other sites, pointing to the relevant volumes. In continuing the SRDF example, you would use the identifier for the R21 on the secondary site. The most efficient way to do this is through the command line, as it does not require a virtual machine to be present to complete the step, for example:
 - a. Create a directory on the VMFS volume corresponding to the virtual machine name, for example, ClusterNode1.
This is the structure created by Step 2.

b. Run the following command:

```
vmkfstools -z  
/vmfs/devices/disks/naa.60060160435026006ea23bec1505df11  
/vmfs/volumes/Non-Replicated-  
DS1/ClusterNode1/ClusterNode1.vmdk -a lsilogic
```

5. Now that an independent mapping file exists in all three locations, adjust the VMX configuration file for the virtual machine so that it references the “friendly” name for the VMFS datastore that stores the mapping files:

a. Edit the VMX file by using the VI editor on a vSphere host on the primary site.

b. Locate the entry that corresponds to the mapping file, for example:

```
scsil:0.fileName = "/vmfs/volumes/4d340f9d-e83bceb3-e9c6-  
a4badb39a66b /ClusterNode1/ClusterNode1.vmdk"
```

Note In this case, the UID of the datastore is 4d340f9d-e83bceb3-e9c6-a4badb39a66b. This ID will be different on each site. Each Non-Replicated-DS1 is in fact a distinct entity.

c. Edit the entry in the VMX file and replace it with the friendly name of the datastore so that it looks like this:

```
scsil:0.fileName =  
"/vmfs/volumes/Non_Replicated_DS1/ClusterNode1/ClusterNode  
1.vmdk"
```

The key here is that the friendly name of the datastore that holds the mapping files is consistent across all three sites, even though each site has its own version of that datastore.

There are three unique mapping files in this configuration: RDM A, RDM B, and RDM C. The actual name of the mapping file (ClusterNode1.vmdk in the above example) must also be consistent across the three sites, as this will be part of the path included in the VMX file.

This is all done to ensure that when a virtual machine powers up, the virtual machine will reference the version of the Non-Replicated-DS1 local to that site and read the mapping file that points at the correct unique LUN identifier.

Placing and configuring mapping files to allow seamless failover between sites with VPLEX and SRDF/A

Figure 8 shows the mapping-file placement in the VPLEX/SRDF configuration.

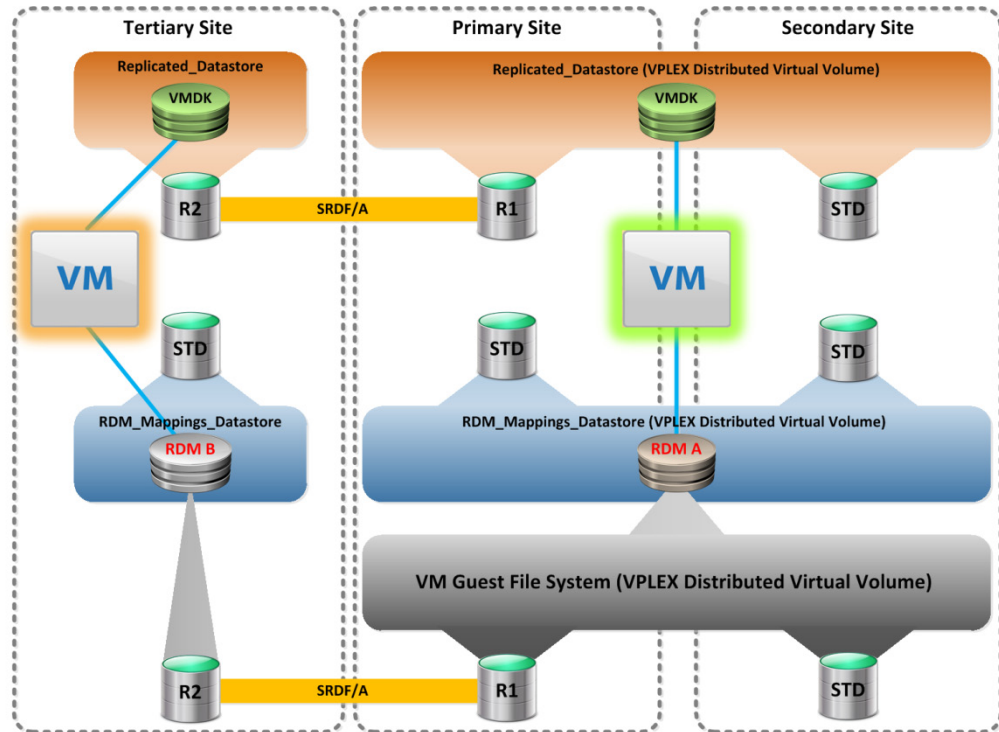


Figure 8. Mapping-file placement in the VPLEX/SRDF configuration

VPLEX by design creates virtual volumes and presents them to both locations with the same ID. This is part of what enables recovery between primary and secondary sites to be achieved by using VMware HA. Effectively, each LUN still has a unique ID, but in this case we are presenting the same distributed LUN to two sites.

In Figure 8, there are only two unique mapping files (RDM A and RDM B). Because the same VPLEX distributed-volume raw device is being accessed on both the primary and secondary sites, these two sites can share a single mapping file (RDM A). The tertiary site is fed by SRDF, and uses a separate mapping file (RDM B).

For the primary and secondary sites, the mapping file should be placed on a datastore created on a VPLEX distributed virtual volume. However, it must be a separate datastore, not the one on which the virtual-machine system volume resides. You need to cater for the lowest common denominator site, which in this case is the tertiary site.

Since the volume the tertiary site will access (the R2) will have its own unique ID, the tertiary site requires a different mapping file (RDM B) to be a separate datastore. It will utilize a local mapping file (and not a replicated mapping file). The full process for creation is as follows:

1. On the primary site, create a datastore called “RDM-Mappings- Datastore” on a VPLEX distributed virtual volume.

The name of the datastore is not important, but choosing a meaningful name such as RDM-Mappings-Datastore provides clarity when viewed through the vCenter Server interface.

2. When adding a raw device to a virtual machine, create the mapping file in another location and specify the “RDM-Mappings- Datastore” datastore.

3. Create a datastore on non-replicated SAN storage on the tertiary site.

Keeping the name of both the VPLEX datastore and the tertiary datastore the same is important. In this example, it must be called “RDM-Mappings- Datastore”.

4. Create unique mapping files on the tertiary site, pointing to the relevant volumes that use the identifier for the R2. The most efficient way to do this is through the command line, which does not require a virtual machine to be present to complete the step, for example:

a. Create a directory on the VMFS volume corresponding to the virtual machine name, for example, ClusterNode1. This is the structure created in step 2.

b. Run the following command:

```
vmkfstools -z  
/vmfs/devices/disks/naa.60060160435026006ea23bec1505df11  
/vmfs/volumes/ RDM-Mappings- Datastore  
/ClusterNode1/ClusterNode1.vmdk -a lsilogic
```

5. Now that an independent mapping file exists at the tertiary site, adjust the VMX configuration file for the virtual machine so that it references the friendly name for the VMFS datastore that stores the mapping files.

a. Edit the VMX file by using the VI editor on a vSphere host on primary site.

b. Locate the entry that corresponds to the mapping file, for example:

```
scsil:0.fileName = "/vmfs/volumes/4d340f9d-e83bceb3-e9c6-  
a4badb39a66b /ClusterNode1/ClusterNode1.vmdk"
```

Note In this case, the UID of the datastore is 4d340f9d-e83bceb3-e9c6-a4badb39a66b. This ID will be different on the tertiary site, because each “RDM-Mappings- Datastore” is in fact a distinct entity.

c. Edit the entry in the VMX file and replace it with the friendly name of the datastore so that it looks like this:

```
scsil:0.fileName =  
"/vmfs/volumes/ RDM-Mappings- Datastore  
/ClusterNode1/ClusterNode1.vmdk"
```

The key here is that the friendly name of the datastore that holds the mapping files is consistent across all three sites, even though the tertiary site has its own version of that datastore.

There are two unique mapping files in this configuration (RDM A, RDM B). The actual name of the mapping file (ClusterNode1.vmdk in the above example) must also be consistent across the three sites.

This is all done to ensure that when a virtual machine powers up, the virtual machine will reference the version of the “RDM-Mappings- Datastore” local to that site and read the mapping file that points at the correct unique LUN identifier.

Mounting datastores in alternate locations

The default behavior of VMware vSphere is to prevent datastores from another vSphere cluster automatically mounting. This prevents that different cluster from inadvertently mounting or manipulating the data on that volume while the original cluster is still operational.

Therefore, the first time a datastore from the primary site is presented to the secondary or tertiary sites, even though the LUNs are present and write-enabled, the vSphere cluster will not automatically mount the volumes. To override this, go to the command line of vSphere host and run the following command:

```
esxcfg-volume -l
```

This command returns a list of the volumes detected as snapshots or replicas, along with a determination of whether they can be mounted and if they can be resignatured.

At this point, you can choose to temporarily or persistently mount a volume by using either of the following commands:

- `esxcfg-volume -m naa.id` (temporary mount)
- `esxcfg-volume -M naa.id` (persistent mount)

Note You can also run these commands from a remote host using VMware vSphere CLI. In that case, use the `vicfg-volume.pl` file rather than `esxcfg-volume`.

Once complete, the volumes are available for use by the vSphere environment.

These commands may also be necessary in a case where you implement a VPLEX and encapsulate existing datastores and present them back to the vSphere hosts. These volumes will be detected as snapshots or replicas and will require the same process to bring them online again.

Adding virtual machines

Adding virtual machines to inventory on alternate sites

In VPLEX, the primary and secondary sites are effectively one site from the perspective of vSphere, so virtual machine inventory control is handled by the VMware HA process. In moving to a tertiary site in the VPLEX scenario, or moving from any site to any other site in the SRDF/Star scenario, it is necessary to find and add all the virtual machines to the target vCenter Server instance.

Any given virtual machine will likely require only additions to inventory once per site (unless for some reason a virtual machine is removed from inventory). There will be situations where virtual machines have been added to the environment since the last failover. These new machines need to be found and added.

Once all virtual machines are added, there is also the issue of addressing power-on prompts for the virtual machines. As virtual machines are moved to a new site, vSphere will recognize this fact and prompt the user for a response as to whether the machine has been moved or copied. In most cases, “I moved it” will be the correct response here, assuming the MAC address and other information as it was on the originating site is to be retained.

Adding virtual machines through PowerCLI cmdlets

The following steps and cmdlets can be used to recover virtual machines at an alternate site. This assumes storage has been presented correctly in advance of these commands:

1. Connect to the relevant vCenter instance:

```
connect-viserver vCenterInstanceName
```

2. Scan each datastore in sequence. Add the discovered virtual machines to inventory on the desired vSphere host:

```
dir  
'vmstores:\vCenterInstanceName@443\DatacenterName\*\*\*.vmx'  
| % \{New-virtual machine -Host vSphereHostname -VMFilePath  
$_.Data storeFullPath}
```

where `**` can be replaced by any string or pattern (including wildcards) matching:

```
\Data storeName\FolderName\VMXFileName
```

3. Power on the virtual machines.

```
Get-virtual machine -Name * | Start-VM
```

4. Answer the virtual machine prompt regarding “msg.uuid.altered”:

```
Get-virtual machine -Name * | Get-VMQuestion | Set-VMQuestion  
-Option “I moved it” -confirm:$false
```

At this point all virtual machines should be powered on.

Allowing for virtual machine power-on sequencing

The syntax mentioned in [Adding virtual machines through PowerCLI cmdlets](#) takes a simplistic view in terms of the powering-on sequence of the virtual machines. They are powered on in the order they are discovered across the datastores.

More sophisticated scripts can search for certain patterns in the name of the virtual machines in order to achieve a specific power-on sequence. For instance, if Active Directory servers were named in the format **ADxx** and SQL servers in the format **SQLxx**, then a full script might look something like this:

```
#Connect to VI Server
connect-viserver vCenterInstanceName

#Find and add all virtual machines from all Data stores to
Inventory
dir 'vmstores:\vCenterInstanceName@443\DatacenterName\*\*\*.vmx' |
%
\{New-virtual machine -Host vSphereHostname -VMFilePath $_.Data
storeFullPath}

#Start all Active Directory Servers and Answer Machine Moved
Prompt
Get-virtual machine -Name AD* | Start-VM
Get-virtual machine -Name AD* | Get-VMQuestion | Set-VMQuestion -
Option "I moved it" -confirm:$false

#Sleep for 3 mins to allow Active Directory Servers to assume
operation
Sleep 180

#Start all SQL Servers and Answer Machine Moved Prompt
Get-virtual machine -Name SQL* | Start-VM
Get-virtual machine -Name SQL* | Get-VMQuestion | Set-VMQuestion -
Option "I moved it" -confirm:$false
```

Conclusion

Summary

When business needs dictate that three sites are required to guarantee the availability of a VMware infrastructure, traditional two-site disaster recovery methods such as cluster enabling technologies or VMware Site Recovery Manager are not sufficient. The two methods described in this white paper offer the customer ways to:

- Establish a three-site disaster recovery
- Automate methods to recover virtual machines in alternate locations
- Cater for RDM usage in three-site configurations

Findings

The two methods of replicating data across three sites by using a combination of EMC technologies shows that:

- Using SRDF/Star technology enables full three-site protection that allows for rolling failures and incremental resynchronizations, but requires three distinct VMware vCenter Server instances.
- Using a combination of VPLEX and SRDF/A technology enables seamless failover and failback between primary and secondary data centers by using VMware HA. The tertiary site operates as a bunker site and requires some full synchronizations on failback to primary or secondary sites.

In all instances, the VMware infrastructure was brought successfully online with low recovery-point objectives by using VMware HA or VMware vSphere PowerCLI.

References

Product documentation

For additional information, see the product documents that are listed below and available on Powerlink. If you do not have access to this content, contact your EMC representative.

- *EMC Symmetrix VMAX Series Product Guide*
- *Symmetrix Customer Procedure Generator utility*
- *VPLEX Procedure Generator utility*
- *VPLEX V4.0 CLI Guide*

Other documentation

For additional information, see [Using VPLEX Metro with VMware HA](#).