Deploy VMware vSphere Metro Storage Cluster on Hitachi Unified Storage VM

Reference Architecture Guide

By Tim Darnell

June 11, 2014
Feedback

Hitachi Data Systems welcomes your feedback. Please share your thoughts by sending an email message to SolutionLab@hds.com. To assist the routing of this message, use the paper number in the subject and the title of this white paper in the text.
# Table of Contents

**Solution Overview** ................................................................. 2
  Logical Design .......................................................................... 3

**Key Solution Components** .................................................... 4
  Hitachi Compute Blade 500 ..................................................... 5
  Hitachi Unified Storage VM ...................................................... 5
  Hitachi Unified Storage ............................................................. 6
  Hitachi Dynamic Link Manager ................................................. 7
  Hitachi TrueCopy Remote Replication Bundle .............................. 7
  Hitachi High Availability Manager ........................................... 7
  Brocade Switches .................................................................... 7
  VMware vSphere 5.5 ................................................................. 8

**Solution Design** ................................................................. 9
  Compute Infrastructure .............................................................. 9
  SAN Infrastructure ................................................................ 14
  Storage Infrastructure ............................................................. 16

**Engineering Validation** .......................................................... 21
  VMware Feature Test Cases ..................................................... 23
  Failure Test Cases ................................................................ 26
  Recovering from Storage Failover (Failback) ................................. 39

**Conclusion** ............................................................................ 42

**Configuring the High Availability Manager Pair** ......................... 43
Deploy VMware vSphere Metro Storage Cluster on Hitachi Unified Storage VM

Reference Architecture Guide

A VMware vSphere Metro Storage Cluster architecture on Hitachi Unified Storage VM with Hitachi Compute Blade 500 provides an ideal solution for maximizing availability and uptime by clustering physical datacenters within metro distances. The metro storage cluster solution from Hitachi Data Systems consists of storage systems presenting replicated storage as a single LUN from different geographically distributed sites. This design enables high availability of services by allowing virtual machine migration between sites with no downtime.

A combination of Hitachi software and hardware provides the following key functions to a vSphere infrastructure:

- Host multipathing
- Internal and external storage provisioning
- Synchronous storage replication across metro cluster distances
- Storage failover

These functions work together with VMware vSphere vMotion, vSphere High Availability, and vSphere Dynamic Resource Scheduler to build this reference architecture for a VMware vSphere Metro Storage Cluster. The advanced functionalities found in Hitachi Unified Storage VM fulfill the requirements of a virtual infrastructure and lessen the need for additional hardware that may be required in traditional Metro Storage Cluster solutions.

This paper is intended for you if you are an IT administrator charged with the storage, deployment, or administration of VMware vSphere infrastructures on Hitachi Unified Storage VM and Hitachi Compute Blade 500. It assumes familiarity with storage area network (SAN)-based storage systems, VMware vSphere, Hitachi data replication technologies, and common IT storage practices.
Solution Overview

This reference architecture uses a VMware infrastructure supported by Hitachi hardware and software to simulate two datacenters in a VMware Metro Cluster Environment.

This solution uses a uniform host access configuration. In a uniform host access configuration, a primary datastore is synchronously replicated to a read-only secondary datastore. ESXi hosts from different sites can access the storage devices on both sites but see the two LUNs as a single datastore.

The following components create a VMware vSphere Metro Storage Cluster environment:

- Hitachi Compute Blade 500
- Hitachi Unified Storage VM
- Hitachi Unified Storage 150
- Hitachi Dynamic Link Manager (HDLM)
- Hitachi TrueCopy Remote Replication bundle
- Hitachi High Availability Manager
- Brocade Storage Area Network and Data Center Networking Switches
- VMware vSphere 5.5
Logical Design

Figure 1 illustrates the high-level logical design of this reference architecture using Hitachi Unified Storage VM (HUS VM), Hitachi Compute Blade 500 (CB 500), and Hitachi Unified Storage 150 (HUS 150).
Key Solution Components

These are descriptions of the key hardware and software components used to deploy this solution.

Table 1 lists information about the hardware components used in this solution.

Table 1. Hardware Components

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Description</th>
<th>Version</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hitachi Unified Storage VM</td>
<td>- Dual controllers</td>
<td>73-03-06</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>- 16 x 8 Gb/sec Fibre Channel ports</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 64 GB cache memory</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 8 x 600 GB 10k RPM SAS disks, 2.5 inch SFF (Replicated Datastores)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 4 x 900 GB 10k RPM SAS disks, 2.5 inch SFF (Command Device)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hitachi Unified Storage 150</td>
<td>- Dual controllers</td>
<td>0977/A</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>- 16 x 8 Gb/sec Fibre Channel ports</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 32 GB cache memory</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 8 x 600 GB 10k RPM SAS disks, 2.5 inch SFF (Quorum Device and vCenter Datastore)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hitachi Compute Blade 500 Chassis</td>
<td>- 8-blade chassis</td>
<td>SVP: A0155-B-7850</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>- 2 Brocade 5460 Fibre Channel switch modules, each with 6 x 8 Gb/sec uplink ports</td>
<td>5460: FOS 7.0.2c</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 2 Brocade VDX 6746 Ethernet switch modules, each with 8 x 10 Gb/sec uplink ports</td>
<td>VDX6746: NOS 3.0.0_dcb3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 2 management modules</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 6 cooling fan modules</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- 4 power supply modules</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Table 1. Hardware Components (Continued)

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Description</th>
<th>Version</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>520HB2 Server Blade</td>
<td>Half blade</td>
<td>Firmware: 04-08 BMC/EFI: 04-06/10-12</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2 x 12-core Intel Xeon E5-2697v2 processors, 2.70 GHz</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>256 GB RAM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16 x 16 GB DIMMs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brocade 6510</td>
<td>SAN switch with 48 x 8 Gb Fibre Channel ports</td>
<td>FOS 7.1.1</td>
<td>4</td>
</tr>
<tr>
<td>Brocade VDX 6720</td>
<td>Ethernet switch with 24 x 10 Gb/sec ports</td>
<td>NOS 3.0.1aa</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2 lists information about the software components used in this solution.

Table 2. Software Components

<table>
<thead>
<tr>
<th>Software</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hitachi Storage Navigator Modular 2</td>
<td>Microcode Dependent</td>
</tr>
<tr>
<td>Hitachi Dynamic Provisioning</td>
<td>Microcode Dependent</td>
</tr>
<tr>
<td>VMware vCenter server</td>
<td>5.5.0, Build 1312298</td>
</tr>
<tr>
<td>VMware Virtual Infrastructure Client</td>
<td>5.5.0, Build 1281650</td>
</tr>
<tr>
<td>VMware ESXi</td>
<td>5.5.0, Build 1331820</td>
</tr>
<tr>
<td>Microsoft® Windows Server® 2008</td>
<td>Enterprise edition, R2</td>
</tr>
<tr>
<td>Hitachi Dynamic Link Manager</td>
<td>7.6.0-00</td>
</tr>
<tr>
<td>VMware vSphere Command Line Interface</td>
<td>5.5</td>
</tr>
<tr>
<td>Command control interface for Hitachi</td>
<td>01-30-03/03</td>
</tr>
</tbody>
</table>

Hitachi Compute Blade 500

Hitachi Compute Blade 500 combines the high-end features with the high compute density and adaptable architecture you need to lower costs and protect investment. Safely mix a wide variety of application workloads on a highly reliable, scalable, and flexible platform. Add server management and system monitoring at no cost with Hitachi Compute Systems Manager, which can seamlessly integrate with Hitachi Command Suite in IT environments using Hitachi storage.

Hitachi Unified Storage VM

Hitachi Unified Storage VM is an entry-level enterprise storage platform. It combines storage virtualization services with unified block, file, and object data management. This versatile, scalable platform offers a storage virtualization system to provide central storage services to existing storage assets.
Unified management delivers end-to-end central storage management of all virtualized internal and external storage on Unified Storage VM. A unique, hardware-accelerated, object-based file system supports intelligent file tiering and migration, as well as virtual NAS functionality, without compromising performance or scalability.

The benefits of Unified Storage VM are the following:

- Enables the move to a new storage platform with less effort and cost when compared to the industry average
- Increases performance and lowers operating cost with automated data placement
- Supports scalable management for growing and complex storage environment while using fewer resources
- Achieves better power efficiency and with more storage capacity for more sustainable data centers
- Lowers operational risk and data loss exposure with data resilience solutions
- Consolidates management with end-to-end virtualization to prevent virtual server sprawl

**Hitachi Unified Storage**

[Hitachi Unified Storage](https://www.hitachi.com) is a midrange storage platform for all data. It helps businesses meet their service level agreements for availability, performance, and data protection.

The performance provided by Hitachi Unified Storage is reliable, scalable, and available for block and file data. Unified Storage is simple to manage, optimized for critical business applications, and efficient.

Using Unified Storage requires a smaller capital investment. Deploy this storage, which grows to meet expanding requirements and service level agreements, for critical business applications. Simplify your operations with integrated set-up and management for a quicker time to value.

Unified Storage enables extensive cost savings through file and block consolidation. Build a cloud infrastructure at your own pace to deliver your services.

Hitachi Unified Storage 150 provides reliable, flexible, scalable, and cost-effective modular storage. Its symmetric active-active controllers provide input-output load balancing that is integrated, automated, and hardware-based.

Both controllers in Unified Storage 150 dynamically and automatically assign the access paths from the controller to a logical unit (LU). All LUs are accessible, regardless of the physical port or the server that requests access.
Hitachi Dynamic Link Manager

Hitachi Dynamic Link Manager, used for SAN multipathing, has configurable load balancing policies. These policies automatically select the path having the least amount of input/output processing through all available paths. This balances the load across all available paths, which optimizes IOPS and response time.

Hitachi TrueCopy Remote Replication Bundle

For synchronous replication up to 190 miles (300 km), Hitachi TrueCopy Remote Replication bundle provides a no-data-loss, rapid restart solution. Real-time copies are the same as the originals. This reduces recovery time to minutes.

Synchronous replication provides very fast recovery time (low RTO) and good data currency (low RPO) between Hitachi Data Systems storage systems.

Hitachi High Availability Manager

Built on the ability of how Hitachi Virtual Storage Platform and Hitachi Unified Storage VM manage virtualized devices, the Hitachi High Availability Manager administers internal storage and externally attached heterogeneous storage with common and integrated management. Its use is in conjunction with storage system-based replication technologies, such as Hitachi Universal Replicator and Hitachi TrueCopy.

Brocade Switches

Brocade and Hitachi Data Systems have collaborated to deliver storage networking and data center solutions. These solutions reduce complexity and cost, as well as enable virtualization and cloud computing to increase business agility.

The Brocade 5460 8Gb SAN Switch for Hitachi Compute Blade 500 delivers an easy to manage embedded Fibre Channel switch with 8 Gbps performance. The Brocade 5460 8Gb SAN Switch hot-plugs into the back of the Hitachi Compute Blade 500 Enclosure. The integrated design frees up rack space, enables shared power and cooling, and reduces cabling and small form-factor pluggables. Enhanced trunking support with external switches provides higher bandwidth to enable demanding applications such as server and desktop virtualization. The Brocade 5460 8Gb SAN Switch saves space and cost, significantly simplifies the SAN environment, enables easier deployment and management, and delivers the performance required for emerging higher throughput applications.
The Brocade 6510 delivers exceptional price/performance for growing SAN workloads through a combination of market-leading throughput and an affordable switch form factor. The 48 ports produce an aggregate 768 Gbps full-duplex throughput; any eight ports can be trunked for 128 Gbps Inter-Switch Links (ISLs). Exchange-based Dynamic Path Selection (DPS) optimizes fabric-wide performance and load balancing by automatically routing data to the most efficient available path in the fabric. It augments ISL trunking to provide more effective load balancing in certain configurations.

The Brocade VDX 6740 Switch is a high-performance 10 Gigabit Ethernet (GbE) switch that supports the most demanding business applications. It is specifically designed to improve network utilization, maximize application availability, increase scalability, and dramatically simplify network architecture in virtualized data centers.

VMware vSphere 5.5

VMware vSphere 5.5 is a virtualization platform that provides a datacenter infrastructure. It features vSphere Distributed Resource Scheduler (DRS), High Availability, and Fault Tolerance.

VMware vSphere 5.5 has the following components:

- **ESXi 5.5** — This is a hypervisor that loads directly on a physical server. It partitions one physical machine into many virtual machines that share hardware resources.

- **vCenter Server** — This allows management of the vSphere environment through a single user interface. With vCenter, there are features available such as vMotion, Storage vMotion, Storage Distributed Resource Scheduler, High Availability, and Fault Tolerance.
Solution Design

Use this reference architecture to create a stretched cluster environment between two virtualized datacenters within a metro distance range. It uses the hardware and software components described in “Key Solution Components” on page 4 to build the following infrastructures:

- Compute Infrastructure
- SAN Infrastructure
- Storage Infrastructure

The architecture uses a synchronous storage replication model.

Compute Infrastructure

This is the compute infrastructure used in this reference architecture:

- VMware Metro Storage Cluster infrastructure components
  - Hitachi Compute Blade 500
  - VMware vSphere 5.5
  - VMware vSphere High Availability
  - VMware vSphere Dynamic Resource Scheduler
- VMware Metro Storage Cluster software components
  - Hitachi Dynamic Link Manager
  - VMware vSphere Command-Line Interface
  - Command control interface on Hitachi products

VMware Metro Storage Cluster Infrastructure Components

A Hitachi Compute Blade 500 chassis with two server blades hosted the VMware vSphere 5.5 Site 1 infrastructure, and another Hitachi Compute Blade 500 chassis with two server blades hosted the VMware vSphere 5.5 Site 2 infrastructure.

For this reference architecture, the Site 1 and Site 2 data centers have two VMware ESXi 5.5 hosts running on each site, totaling four VMware ESXi 5.5 hosts. To effectively leverage VMware High Availability and VMware DRS, and to minimize unnecessary site-to-site failures, a minimum of 2 ESXi hosts on each site is recommended.

A single VMware vCenter Server 5.5 instance manages ESXi hosts on both sites in the cluster with VMware High Availability and Dynamic Resource Scheduler enabled. Virtual machines can take advantage of long distance vMotion and host failover capabilities across sites.
Note — Follow recommended practice by running the VMware vCenter server at a third site. This way any site-wide failure does not affect the management of the virtual environment. For management redundancy, run the vCenter server as a virtual machine in a VMware high availability configuration.

To minimize virtual machine WAN I/O traffic across sites, create DRS groups and affinity rules to ensure virtual machines run on the local site relative to the primary datastore.
VMware Metro Storage Cluster Software Components

Figure 2 shows the overall architecture of the VMware Metro Storage Cluster Software components in a vSphere environment.
The software components handling the management for host path failover and storage replication control are:

- Hitachi Dynamic Link Manager
- VMware vSphere Command Line Interface
- Command control interface for Hitachi products

Hitachi Dynamic Link Manager (HDLM) provides the following:

- Load balancing by distributing loads across multiple paths
- Path failover by automatically switching to another path if there is a failure in the path that is currently being used
- Path failback by manually or automatically bringing a path back online after recovering from an error
- Coordination with Hitachi High Availability Manager to manage the path failover from one storage system to another storage system
- Distinguish between primary volumes and secondary volumes in a High Availability Manager-enabled pair
- VMkernel-level storage brokering functionality

Hitachi Dynamic Link Manager contains the following three functions:

- **HDLM command** — Provides the `dlnkmgr` command for remotely managing paths, displaying path status information, and setting up the Hitachi Dynamic Link Manager operating environment via command line. The `dlnkmgr` command is necessary when executing a failback to restore the metro cluster after a storage failure. (See "Recovering from Storage Failover (Failback)" for more details.) Install it on a remote management client.

- **HDLM driver** — Controls all Hitachi Dynamic Link Manager functions, manages paths, and detects errors. The following plug-ins are installed within the pluggable storage architecture (PSA) from VMware on VMKernel:
  - Storage Array Type Plug-in (SATP)
  - Path Selection Plug-in (PSP)

- **HDLM utility** — Enables you to collect path status information. Install it on a remote management client.

Installing the VMware vSphere Command-Line Interface is a prerequisite to installing Hitachi Dynamic Link Manager. This is necessary so that HDLM command can connect to the ESXi hosts.
For this reference architecture, the VMware vSphere CLI and Hitachi Dynamic Link Manager were installed on the vCenter server to remotely manage the paths on all four ESXi hosts. The HDLM driver offline bundle zip file located in C:\Program Files (x86)\HITACHI\DynamicLinkManagerForVMware\plugin of the vCenter server was uploaded to the ESXi hosts. (For x86 hosts, the path is C:\Program Files\HITACHI\DynamicLinkManagerForVMware\plugin.)

Launch VMware vSphere CLI and run the following command to remotely connect to the ESXi hosts and install the HDLM drivers:

```
esxcli.exe -s <esxi host> -u <user> -p <password> software vib install -d <full path to uploaded HDLM offline bundle file>
```

ESXi hosts use the following installed HDLM drivers to help manage and maintain a metro storage cluster environment:

- hex-hdlm-dlnkmgr
- psp-hdlm-exlbk
- psp-hdlm-exlio
- psp-hdlm-exrr
- satp-hdlm

From vSphere CLI, run the following command to see a sorted view of the status and multipathing configuration for all LUNs on an ESXi host:

```
dlnkmgr.exe -s <ESXi host> -u <user> -p <password> view -path -srt lu
```

Run the `help` command in dlnkmgr.exe or see the *Hitachi Dynamic Link Manager Software User Guide* for additional command line options.

Command control interface on Hitachi products enables you to perform storage system operations by issuing commands to Hitachi Unified Storage VM. There are two command control interface components residing on the following:

- **Storage system** — Command devices and Hitachi TrueCopy volumes (P-VOLs and S-VOLS)
- **Server** — Hitachi Open Remote Copy Manager (HORCM), configuration definition files (for example, `horcm0.conf`), and command control interface commands

Command device logical volumes were created on the Site 1 and Site 2 storage systems. These logical volumes were presented as physical raw device mappings (RDM) to the vCenter server. You cannot map command devices as virtual RDMs.
SAN Infrastructure

This describes the SAN infrastructure used in this reference architecture.

Server to Storage SAN Design

Each server blade at the primary and secondary site used dual-port Fibre Channel mezzanine cards. They connect internally to the Brocade 5460 Fibre Channel switch modules located in the Hitachi Compute Blade 500 chassis.

Two inter-switch links from each of the internal Brocade 5460 Fibre Channel switch module were connected to two Brocade 6510 switches, one link per switch for redundancy. Fabric zoning was configured so that each server had redundant paths to each storage system.

Additionally, two Fibre Channel connections were made from the Hitachi Unified Storage 150 system to each Brocade 6510 switch. The Hitachi Unified Storage 150 is zoned to each Hitachi Unified Storage VM system and presented as virtualized external storage. High Availability Manager uses the externalized storage as a quorum disk to ensure data is consistent between a High Availability Manager pair. Only a single quorum disk is required between a primary and secondary storage system.

Figure 3 shows the storage network configured for each site.
Storage to Storage SAN Design

To facilitate copying data between storage systems, storage replication links were configured between the Hitachi Unified Storage VM at Site 1 and Hitachi Unified Storage VM at Site 2. Each storage system used a total of two initiator ports and two RCU target ports.

A storage replication link consists of an initiator port on Site 1 connected to a remote control unit (RCU) target port defined on Site 2. It represents a one-way remote copy connection from the primary data volume (P-VOL) on Site 1 to the secondary data volume (S-VOL) on Site 2.

For this solution, the configuration of a second storage replication link had an initiator port on Site 2 connected to an RCU target port on Site 1. The configuration enables reversing the direction of replication from Site 2 to Site 1 after a path failover occurs. The ability to reverse the direction ensures data protection of the S-VOL on the remote storage system while it is in an active state. Two additional replication links connect through two Brocade 6510 switches. This totals four paths to provide maximum hardware redundancy, following recommended practice.

**Note** — The supported latency for synchronous storage replication must be less than 5 milliseconds round-trip time (RTT), as defined by VMware.

Figure 4 shows the storage system site-to-site configuration.

- Primary-Secondary replication for Failover
  - (2 Initiator ports – 2 RCU Target ports)
- Secondary-Primary replication for Failback
  - (2 Initiator ports – 2 RCU Target ports)
- Replication Link

![Figure 4]
Storage Infrastructure

The storage infrastructure used in this reference architecture consists of the following:

- Storage Design
- Defining Volume Pair Relationship
- Storage Replication with High Availability Manager and TrueCopy

Storage Design

For this reference architecture, Hitachi Dynamic Provisioning was used on Hitachi Unified Storage VM to simplify management of the storage.

This solution used a dynamic provisioning pool comprised of a single RAID group with eight 600 GB 10k RPM SAS drives in a RAID-6 (6D+2P) configuration for each storage system. Using a RAID-6 configuration lowers the risk of data loss or pool failure, which is a primary concern for virtual machines protected in a stretch cluster environment. While this solution used disks internal to Hitachi Unified Storage VM, external storage virtualized behind Hitachi Unified Storage VM is supported as well.

Figure 5 shows the configuration for the dynamic provisioning pool on each storage system for the test environment.

![Diagram](image-url)
A VMFS LUN is provisioned for storing synchronous replication-protected virtual machines on the storage system on Site 1. The Hitachi Unified Storage VM configuration on Site 2 is identical to the configuration on Site 1. The recovery LUNs on Site 2 are identical in size to their respective protected LUNs, as required for maintaining the P-VOL and S-VOL relationship. Both the P-VOL and S-VOL LUNs are presented to all ESXi hosts in the cluster.

A 46 MB LUN, the smallest LUN that can be created, is provisioned on both sites. These LUNs are converted to command devices and presented to the vCenter server. A command device is a dedicated logical volume on the storage system that functions as the interface to the storage system from a host. In this solution, the command control interface sends replication commands to the command device for execution on the storage system.

In addition, a 47 MB LUN is created on a Hitachi Unified Storage 150 storage system for use as a quorum disk. This LUN was presented to the Site 1 storage system and the Site 2 storage system as externalized storage. The quorum disk stores continually-updated information about data consistency in Hitachi High Availability Manager P-VOLs and S-VOLs for use during failover operations. The information is used by High Availability Manager in the event of failure to direct host operations to the secondary volume.

---

**Note**: Following recommended practice, connect the Hitachi Unified Storage 150 storage system to each Hitachi Unified Storage VM from a third site within metro distance from each site to ensure the quorum disk is accessible from at least one site during a site-wide failure.

---

**Defining Volume Pair Relationship**

A key aspect of this VMware vSphere Metro Storage Cluster reference architecture on Hitachi Unified Storage VM is defining the volume pair relationship for replication between storage systems. Define and manage storage replication relationships through the Hitachi Storage Navigator graphical user interface or a host running command control interface.

Hitachi Open Remote Copy Manager is a component of command control interface. For this solution, the command control interface issues Open Remote Copy Manager commands for managing, monitoring, and controlling the replication process.

Two configuration definition files (horcmx.conf) are defined on the vCenter server, one for each Open Remote Copy Manager instance managing the volumes on the primary and secondary storage systems. This file defines the copy pairs. After changes are made to the configuration definition files, restart the Open Remote Copy Manager daemon to ensure that the most recent changes to the copy pair are reflected.
Following recommended practice, the HORCM configuration file managing the P-VOL was labeled with an even number (`horcm0.conf`). The configuration file managing the S-VOL was labeled with an odd number (`horcm1.conf`). Ensure the port name and number for both HORCM instances is registered in the `C:\Windows\System32\drivers\etc\services` file before starting them.

Figure 6 verifies the pair relationship by running the `pairdisplay` command from the horcm0 instance.

```bash
C:\HORCM\etc>pairdisplay -q UMSC-S1 -IH0 -fex
Group  PeerVol(L/R)  (Port H,TID, LID), Seq LDEV P/S Status. Pense,  %P-LDEV# M
UMSC-S1 UMSC-UMFS-Site1(L)  (CLI-A-1, 1, 101)210203  4.SMPL  ---   ----------- -----------
UMSC-S1 UMSC-UMFS-Site1(R)  (CLI-B-1.28, 201)210081  2.SMPL  ---   ----------- -----------
```

**Figure 6**

Initially, the volumes are in simplex (SMPL) mode. The volumes are not paired and synchronized until running the High Availability Manager paircreate process through the Hitachi Storage Navigator GUI.

---

**Note** — High Availability Manager pair creation is available using the Storage Navigator GUI only. Use command control interface to view and manage storage replication activity after the High Availability Manager pair creation.

---

For more details around the configuration of the High Availability Manager pair relationship, see "Configuring the High Availability Manager Pair."

### Storage Replication with High Availability Manager and TrueCopy

Hitachi High Availability Manager allows virtual machines to failover between Hitachi Unified Storage VM storage systems in the event of an I/O failure to the primary storage system. The combination of High Availability Manager with Hitachi Dynamic Link Manager enables ESXi hosts to see the primary volume and secondary volume as a single LUN. If a host is unable to access the production volume on the primary storage system, I/O from the ESXi hosts redirect to the secondary volume on the remote system.

Hitachi High Availability Manager uses Hitachi TrueCopy Remote Replication bundle to create a synchronous remote copy of a production volume. Prior to creating a High Availability Manager pair, use the Storage Navigator Quorum Disk Operation GUI to configure the 47 MB LUN from the Hitachi Unified Storage 150 storage system as a quorum disk. Configure the primary storage system and the secondary storage system the same way.

After defining volume pair relationships and starting the Hitachi Open Remote Copy Manager (HORCM) daemon, initiate storage replication using the Hitachi Storage Navigator GUI. Disaster recovery operations use a mix of Storage Navigator, command control interface, and the Hitachi Dynamic Link Manager `dlnkmgr.exe` command.
Figure 7 shows the High Availability Manager pair creation option in the Storage Navigator GUI.

Synchronous (real-time) replication provides a high level of data consistency protection. It ensures that the remote data is identical to the local data. In synchronous replication, an input-output update operation from a virtual machine is not complete until confirming the completion at the primary and recovery site. With increased distances, this can cause problems for latency-sensitive applications or virtual machines. The practical distance for synchronous replication ranges from 20 miles (32 km) to 100 miles (161 km), depending on application or virtual machine tolerance. Synchronous replication up to 190 miles (300 km) requires using WAN-optimized controllers.
Note — Once the High Availability Manager pair creation is established, restart all ESXi hosts for the P-VOL and S-VOL volumes to be seen as a single LUN. Validate that all ESXi hosts can see an active and standby path to the storage volume.

Figure 8 shows the expected multipath status from an ESXi host protected by High Availability Manager and Dynamic Link Manager.

![Figure 8](image-url)
Engineering Validation

For testing purposes, the following test cases were run to validate the VMware vSphere Metro Storage Cluster reference architecture.

- **VMware feature test cases**
  - Using VMware vMotion or DRS to migrate virtual machines between Site 1 and Site 2
  - Using VMware HIGH AVAILABILITY to failover virtual machines from Site 1 to Site 2

- **Failure test cases**
  - Single path failure
  - Primary storage system failure
  - All active paths to the P-VOL fail for a single host in the cluster
  - Quorum disk failure
  - Storage replication link failure
  - Wide area network storage connection failure
  - Primary site failure

- Recovering from storage failover (failback)

Table 3 outlines the tested scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>HAM/HDLBM Behavior</th>
<th>Observed VMware behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using VMware vMotion or VMware Distributed Resource Scheduler to migrate virtual machines between Site 1 and Site 2</td>
<td>No impact</td>
<td>Virtual machine migrates to Site 2 hosts and I/O is directed to P-VOL on Site 1. Hosts within the cluster will use active paths to P-VOL</td>
</tr>
<tr>
<td>Using VMware High Availability to failover virtual machines between Site 1 and Site 2.</td>
<td>No impact</td>
<td>Virtual machine fails over to Site 2 hosts and I/O is directed to P-VOL on Site 1</td>
</tr>
<tr>
<td>An active path to the P-VOL fails.</td>
<td>Host I/O is redirected to an available active path via HDLM PSP</td>
<td>Another active path is used</td>
</tr>
<tr>
<td>Scenario</td>
<td>HAM/HDLM Behavior</td>
<td>Observed VMware behavior</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| MCU system fails or all active paths to the MCU system hosting the P-VOL fail. | - Storage failover  
  - Hitachi High Availability Manager verifies data integrity with the quorum disk before failover  
  - High Availability Manager splits pair replication and S-VOL is converted to a write enabled status (SSWS)  
  - Host I/O is redirected via HDLM SATP to the S-VOL | - Active paths to P-VOL are reported dead  
  - Standby paths to S-VOL become active  
  - No disruption to virtual machines |
| All active paths to the P-VOL fail for any ESXi host in the cluster.    | - Same behavior as above (storage failover)  
  - When one host loses access to the P-VOL, all hosts in the cluster fail to the S-VOL to avoid a split-brain scenario | - Same behavior as above |
| Quorum disk fails or all paths to quorum disk removed                  | - Replication between P-VOL and S-VOL stop with P-VOL in PSUE state and S-VOL in SSUS state  
  - Failover does not occur  
  - I/O continues to active P-VOL paths | - No impact |
VMware Feature Test Cases

These are the VMware feature test cases.

Using VMware vMotion or Dynamic Resource Scheduler to Migrate Virtual Machines Between Site 1 and Site 2

Since all ESXi hosts on both sites access the P-VOL on the primary storage system, VMware vMotion and Dynamic Resource Scheduler (DRS) are possible for migrating virtual machines between hosts at different sites.

Table 3. Tested Scenarios (Continued)

<table>
<thead>
<tr>
<th>Scenario</th>
<th>HAM/HDLM Behavior</th>
<th>Observed VMware behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage replication link failure</td>
<td>■ Same behavior as quorum disk failure</td>
<td>■ No impact</td>
</tr>
<tr>
<td>WAN storage connection failure</td>
<td>■ P-VOL cannot process host I/O from site. storage failover occurs&lt;br&gt;■ Site 1: After storage failover, S-VOL cannot process host I/O from Site 1 hosts since Fibre Channel connections across sites are down</td>
<td>■ Site 1: Virtual machines on Site 1 hosts are unable to access their virtual disks on Site 2. Site 1 hosts must be shut down manually for VMware High Availability to restart virtual machines on Site 2 hosts.&lt;br&gt;■ Site 2: After storage failover, S-VOL will process host I/O for Site 2 hosts because local site access remains active. Virtual machines on Site 2 can access the local S-VOL.</td>
</tr>
<tr>
<td>Primary site failure</td>
<td>■ High Availability Manager verifies data integrity with the quorum disk before failover&lt;br&gt;■ High Availability Manager splits pair replication and S-VOL is converted to a write enabled status (SSWS)&lt;br&gt;■ Host I/O is redirected via Hitachi Dynamic Link Manager SATP to the S-VOL</td>
<td>■ VMware High Availability fails over virtual machines to available Site 2 hosts</td>
</tr>
</tbody>
</table>
Figure 9 shows a diagram of the observed behavior.

As mentioned in "VMware Metro Storage Cluster Software Components," host and virtual machine DRS groups are configured with affinity rules to ensure virtual machine migration stays on Site 1 when possible to limit I/O traffic over the WAN.
Using VMware High Availability to Failover Virtual Machines from Site 1 to Site 2

Since all ESXi hosts on both sites access the P-VOL on the primary storage system, VMware High Availability (HA) will restart virtual machines on available ESXi hosts on different sites during host failures.

Figure 10 shows a diagram of the observed behavior.
If host and virtual machine DRS groups are configured with affinity rules to keep virtual machines on Site 1 hosts, VMware High Availability may restart virtual machines on Site 2 hosts but will then use vMotion to move the virtual machines to Site 1 hosts, if available.

**Failure Test Cases**

The following failure test cases are simulated by physically removing cables, taking Fibre Channel ports offline or removing zones from the Fibre Channel fabric.

**Single Path Failure**

When a single active path from an ESXi host to the P-VOL fails, host I/O is redirected to an available active path via the Hitachi Dynamic Link Manager storage array type plug-in (SATP) and path selection plug-in (PSP).
Figure 11 shows a diagram of the failure scenario.

When the path is restored, it will automatically become available and active for use by the ESXi host.
Primary Storage System Failure

When the primary storage system fails or all active paths to the P-VOL hosted on the primary storage system fail, the following occurs:

- High Availability Manager verifies data integrity with the quorum disk before failover.
- High Availability Manager splits pair replication and the S-VOL hosted on the secondary storage system is converted to a write enabled status (SSWS).
- All ESXi host I/O is redirected through the standby path to the S-VOL hosted on the secondary storage system via Hitachi Dynamic Link Manager driver.

The following is observed in the VMware infrastructure:

- Active paths to the P-VOL on the primary storage system are reported dead.
- Standby paths to the S-VOL on the secondary storage system become active.
- No disruption is seen on the virtual machine.
Figure 12 shows a diagram of the failure scenario.

To restore the Metro Storage Cluster to its original state, see "Recovering from Storage Failover (Failback)."
All Active Paths to the P-VOL Fail for a Single Host in the Cluster

When a single ESXi host in the cluster (located in Site 1 or Site 2), loses all active paths to the P-VOL on the primary storage system, storage failover occurs and the same behavior as a primary storage system failure is seen. This prevents a split-brain scenario where some hosts access the P-VOL while other hosts access the S-VOL.
Figure 13 shows a diagram of the failure scenario.

To restore the Metro Storage Cluster to its original state, see Recovering from Storage Failover (Failback).
Quorum Disk Failure
When a quorum disk fails or all paths to the quorum disk are removed, the following occurs:

- Replication between the P-VOL on the primary storage system and the S-VOL on the secondary storage system stops.
- P-VOL is in PSUE state and the S-VOL is in SSUS state.
- Storage failover does not occur.
- All ESXi hosts continue to use active paths to the P-VOL.
Figure 14 shows a diagram of the failure scenario.

Once quorum disk and path issues are resolved, run a `pairresync.exe` command to resume replication.
Storage Replication Link Failure
When the storage replication link fails, the same behavior as a quorum disk failure scenario occurs.

Figure 15 shows a diagram of the failure scenario.

Once replication link issues are resolved, run `pairresync.exe` command to resume replication.
Wide Area Network Storage Connection Failure

**Note** - Permanent Device Loss (PDL) default behavior has changed with the release of vSphere 5.5. It is recommended to disable the PDL AutoRemove feature on vMSC hosts by changing the advanced setting "AutoremoveOnPDL" to 0 (disabled) from 1 (enabled). Please see VMware Knowledge Base article 2059622 for more information.

During a wide area network storage connection failure, where links for replication and access across sites fail but links to the local site and layer 2 network connectivity remain active, the following occurs:

- P-VOL cannot process host I/O from Site 2 so High Availability Manager and Dynamic Link Manager switch the active path to the S-VOL.
- S-VOL on Site 2 is converted to write enabled status (SSWS).
- Site 1: After storage failover, S-VOL cannot process host I/O from Site 1 hosts since Fibre Channel connections across sites are down. Virtual machines on Site 1 hosts are unable to access their virtual disks on Site 2.

**Note** — VMware defines this as an all paths down (APD) condition. If network connections to the virtual machine remain up while storage connectivity is down, VMware High Availability will not automatically restart virtual machines on available hosts. Site 1 hosts must be shut down manually for VMware high availability to restart virtual machines on Site 2 hosts.

- Site 2: After storage failover, S-VOL will process host I/O for Site 2 hosts because local site access remains active. Virtual machines on Site 2 can access the local S-VOL.
Figure 16 shows a diagram of the failure scenario.

This scenario describes a storage connection-only failure across sites. WAN failures involving both storage and network connections will trigger VMware High Availability to automatically restart virtual machines on Site 2 hosts.

To restore the metro cluster to its original state, see "Recovering from Storage Failover (Failback)."
Primary Site Failure
When the entire primary site fails, the following occurs:

- High Availability Manager verifies data integrity with the quorum disk accessible from Site 2 before failover.
- High Availability Manager splits pair replication and the S-VOL hosted on the secondary storage system is converted to a write enabled status (SSWS).
- All ESXi host I/O is redirected to the S-VOL hosted on the secondary storage system via Hitachi Dynamic Link Manager driver.

The following is observed in the VMware infrastructure:

- All Site 1 ESXi hosts fail and VMware High Availability restarts virtual machines on Site 2 ESXi hosts.
- Active paths to the P-VOL on the primary storage system are reported dead.
- Standby paths to the S-VOL on the secondary storage system become active.
Figure 17 shows a diagram of the failure scenario. To restore the metro cluster to its original state, see "Recovering from Storage Failover (Failback)."
Recovering from Storage Failover (Failback)

The following assumes that High Availability Manager has triggered a storage failover to the S-VOL on Site 2. After Site 1 storage has been restored and is ready to serve production data, the following steps are required for restoring the metro storage cluster to its original state:

1. Run the `pairresync -swaps` command from the horcm1 instance in command control interface or through the Hitachi Storage Navigator GUI (Site 2 to Site 1).

   - This reverses the roles of the two sites and synchronizes the volume. Any changes written to Site 2 storage (P-VOL) are replicated back to Site 1 storage (S-VOL). All owner paths for Site 1 storage are changed to Online(S) within Dynamic Link Manager.

Figure 18 shows a diagram of the host paths and replication state before and after running the `pairresync -swaps` command.
2. Bring the owner paths for P-VOLs to an Online(D) status by issuing the `dlnkmgr online -dfha` command from the Dynamic Link Manager command line interface on each ESXi host in the Metro Storage Cluster.

- The `-dfha` option is a new option with Dynamic Link Manager version 7.6 and higher. This introduces a new path status of "Online(D)" in addition to "Online(S)". Online(D) stands for Online Deferred status, and Online(S) stands for Online Suspended status. This new path status does not allow I/O which is similar to Online(S), but allows for automatic failback when the host loses access to the S-VOL during execution of the `dlnkmgr online -hapath` command below. As a result, VMs will run without disruption during High Availability Manager failback without the need to migrate to other hosts.

3. Run the `dlnkmgr online -hapath` command from the Dynamic Link Manager command line interface on a single ESXi host accessing the High Availability Manager pair.

- This allows hosts to begin writing back to Site 1 storage (S-VOL). I/O to the Site 1 storage S-VOL triggers a storage failover event by High Availability Manager in the reverse direction. Replication from Site 2 storage to Site 1 storage is split and Site 1 storage is converted to write enabled status (SSWS).
Figure 19 shows a diagram of the observed behavior.

1. Run the `pairresync -swapp` command from the horcm0 instance in command control interface or through Storage Navigator GUI (Site 1 to Site 2).
   - This reverses the roles of the two sites. Replication is re-enabled from Site 1 to Site 2, returning the metro cluster to its original state.
Conclusion

This reference architecture guide documents creating and validating a VMware vSphere Metro Storage Cluster environment with VMware vSphere 5.5 on Hitachi Unified Storage VM. Multiple components work together to provide an infrastructure with continuous availability:

- Hitachi Unified Storage VM provides a flexible, robust, and high performance storage environment.
- Hitachi Dynamic Link Manager provides host multipathing with deep storage integration.
- Hitachi TrueCopy Remote Replication bundle provides synchronous mirror replication to ensure zero data loss.
- Hitachi High Availability Manager enables host-based applications to failover between Hitachi Unified Storage VM storage systems with no disruption.

The above components are tightly integrated with VMware virtual infrastructures to protect the data center from various failure scenarios. You can also leverage the benefits of synchronous data replication beyond traditional disaster recovery protection to include workload balancing across sites and coordinated site-wide maintenance with no down time.
Configuring the High Availability Manager Pair

The High Availability Manager pair definition used these HORCM configuration files.

Figure 20 shows the horcm0.conf contents used in the test environment:

```
HORCM_MON
172.17.172.158 horcm0 1000 3000
HORCM_CMD
\\\PhysicalDrive2
HORCM_LDEV
UMSC-S1 UMSC-UMFS-Site1 210203 00:04
HORCM_INST
UMSC-S1 172.17.172.158 horcm1
```

**Figure 20**

Figure 21 shows the horcm1.conf contents used in the test environment:

```
HORCM_MON
172.17.172.158 horcm1 1000 3000
HORCM_CMD
\\\PhysicalDrive3
HORCM_LDEV
UMSC-S1 UMSC-UMFS-Site1 210081 00:02
HORCM_INST
UMSC-S1 172.17.172.158 horcm0
```

**Figure 21**

Figure 20 and Figure 21 show the logical device or LUN associated to device group VMSC-S1. In the VMSC-S1 pair, the following is true:

- Horcm0 instance manages the P-VOL defined as LUN 00:04 on a Hitachi Unified Storage VM registered with serial number 210203.
- Horcm1 instance manages the S-VOL defined as LUN 00:02 on a Hitachi Unified Storage VM registered with serial number 210081.

Table 4 shows the pair configuration defined by the configuration files in Figure 20 and Figure 21.

### Table 4. Replication Pair Defined by HORCM Configuration Files

<table>
<thead>
<tr>
<th>Device Group</th>
<th>HUS VM Serial Number</th>
<th>LUN</th>
<th>Volume Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMSC-S1</td>
<td>210203</td>
<td>00:04</td>
<td>P-VOL</td>
</tr>
<tr>
<td></td>
<td>210081</td>
<td>00:02</td>
<td>S-VOL</td>
</tr>
</tbody>
</table>
For More Information

Hitachi Data Systems Global Services offers experienced storage consultants, proven methodologies and a comprehensive services portfolio to assist you in implementing Hitachi products and solutions in your environment. For more information, see the Hitachi Data Systems Global Services website.

Live and recorded product demonstrations are available for many Hitachi products. To schedule a live demonstration, contact a sales representative. To view a recorded demonstration, see the Hitachi Data Systems Corporate Resources website. Click the Product Demos tab for a list of available recorded demonstrations.

Hitachi Data Systems Academy provides best-in-class training on Hitachi products, technology, solutions and certifications. Hitachi Data Systems Academy delivers on-demand web-based training (WBT), classroom-based instructor-led training (ILT) and virtual instructor-led training (vILT) courses. For more information, see the Hitachi Data Systems Services Education website.

For more information about Hitachi products and services, contact your sales representative or channel partner or visit the Hitachi Data Systems website.