Dell PowerStore: Metro Volume

May 2024

H19163.5

White Paper

Abstract

This white paper provides in-depth coverage of Dell PowerStore Metro Volume and its features.

DELL Technologies

Copyright

Contents

Executive summary

Overview

Preventing the loss of data or transactions requires a reliable method of continuous data protection and high availability. For disaster avoidance and planned outages, applications and services must be made available at an alternate site. Various data-mobility methods, including asynchronous and synchronous replication, can accomplish the task of providing offsite replicas. Dell PowerStore provides native and non-native solutions to protect data and help organizations meet business goals for both data availability and protection. PowerStore native replication solutions can replicate data to other systems, whether they are at the same site or at a remote facility. Synchronous replication with Metro Volume differentiates itself from the other methods by guaranteeing transactional consistency between PowerStore clusters during normal operation. This white paper focuses on PowerStore Metro Volume and provides in-depth coverage of its features.

Audience

This white paper is intended for Dell Technologies customers, partners, and employees who are considering using PowerStore Metro Volume. The document assumes familiarity with the PowerStore system and management software.

Revisions

Date	Part number/ revision	Description
July 2022	H19163	Initial release: PowerStoreOS 3.0
August 2022	H19163.1	Update initial synchronization process
October 2022	H19163.2	Minor updates
May 2023	H19163.3	Updates for PowerStoreOS 3.5
October 2023	H19163.4	Updates for PowerStoreOS 3.6
May 2024	H19163.5	Updates for PowerStoreOS 4.0 Removed references to PowerStore X

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: Robert Weilhammer, Jason Boche, Marty Glaser, Henry Wong

Note: For links to other documentation for this topic, see the PowerStore Info Hub.

Introduction

Business continuity planning

Data is one of the most valuable assets to an organization. Users (and sometimes their customers) access data constantly, directly and indirectly, using various applications. This makes data a crucial part of day-to-day operations. Outages can occur at any time and can be restricted to a single system or an entire data center or location. Whether they are planned outages such as regular maintenance, or unplanned events such as a power outage, it is a top priority to ensure that critical data is always available. A business-continuity plan for critical data can prevent these costly outages. To protect against different outage scenarios, an organization should plan and implement a data-protection strategy that includes a data-replication solution.

To protect against a storage-system outage, you can use asynchronous or synchronous replication to create a copy of data on a remote system. Replication is a software feature that synchronizes data to a remote system within the same site or a different location. Replicating data helps to provide data redundancy and safeguards against storage-system failures at the main production site. Having a remote disaster recovery (DR) site protects against system and site-wide outages. It also provides a remote location that can resume production and minimize downtime due to a disaster. The PowerStore platform offers many data-protection solutions that can meet DR needs in various environments.

Asynchronous replication is primarily used to replicate data over long distances, but you can use it to replicate to systems within the same location also. The asynchronous replication for PowerStore is designed to have minimal impact on host I/O latency. Host writes are acknowledged once they are saved to the local storage resource, and no other writes are required for change tracking. Because write operations are not immediately replicated to a destination resource, all writes are tracked on the source. This data is replicated during the next synchronization. With protection policies, asynchronous replication uses the concept of a recovery point objective (RPO). The RPO is the acceptable amount of data, which is measured in units of time, that may be lost due to an outage. This delta of time also affects the amount of data that must be replicated during the next synchronization. It also reflects the amount of potential data loss in a disaster scenario.

Synchronous replication is similar to asynchronous replication in that it replicates data between arrays. Synchronous replication differentiates itself from asynchronous in that it can offer an RPO of zero - meaning zero data loss during planned or unplanned outages. Synchronous replication achieves this by always keeping the volume data between two arrays in sync. Over longer distances, synchronous replication can be associated with higher volume latency. For this reason, synchronous replication is primarily used to replicate data over shorter distances or in situations where volume latency is not a concern.

A block Metro Volume provides synchronous replication, spanned across two individual PowerStore clusters for a VMware vSphere Metro Storage Cluster configuration as well as Microsoft Windows Server Failover Clustering (WSFC), Hyper-V, and Linux hosts. A Metro Volume provides disaster avoidance and a load-balancing solution, when participating PowerStore clusters are in the same building or in metro distance. Concurrent active/active host I/O is possible on both sides and is replicated to the remote system. Hosts can connect to both PowerStore clusters simultaneously with active paths

for more redundancy (Uniform host configuration). The stretched architecture allows workload mobility and cross-site automated load balancing together with fast recovery for optimized data-center utilization and downtime avoidance.

You can configure all PowerStore native replication features using PowerStore Manager, PowerStore CLI, or REST API. PowerStore can also integrate with Dell metro node, VPLEX, and RecoverPoint for Virtual Machines. While metro node and VPLEX provide a metro volume feature across different storage arrays, RecoverPoint for Virtual Machines supports VM replication for PowerStore. You can configure it using the Unisphere Manager for RecoverPoint user interface.

PowerStore overview

PowerStore achieves new levels of operational simplicity and agility. It uses a container-based microservices architecture, advanced storage technologies, and integrated machine learning to unlock the power of your data. PowerStore is a versatile platform with a performance-centric design that delivers multidimensional scale, always-on data reduction, and support for next-generation media.

PowerStore brings the simplicity of public cloud to on-premises infrastructure, streamlining operations with an integrated machine-learning engine and seamless automation. It also offers predictive analytics to easily monitor, analyze, and troubleshoot the environment. PowerStore is highly adaptable, providing the flexibility to host specialized workloads directly on the appliance and modernize infrastructure without disruption. It also offers investment protection through flexible payment solutions and data-in-place upgrades.

The PowerStore platform is available as a bare-metal, unified storage array which can service block, file, and VMware vSphere Virtual Volumes (vVols) resources along with numerous data services and efficiencies.

Terminology

The following table provides definitions for some of the terms that are used in this document.

Table 1. Terminology

Term	Definition
Appliance	Solution containing a base enclosure and attached expansion enclosures. The size of an appliance could be only the base enclosure or the base enclosure plus expansion enclosures.
Asynchronous Logical Unit Access (ALUA)	PowerStore uses implicit ALUA which allows PowerStore to provide a recommended active optimized path to a storage resource for the hosts.
Asynchronous replication	Replication method that allows replicating data over long distances and maintaining a replica at a destination site. Updates to the destination image can be issued manually, or automatically based on a customizable RPO.
Bandwidth	Amount of data, represented in MB/s, which can be transferred in a given period.
Common base	Pair of snapshots that are taken on a replication source and destination storage resource that have the same point-in-time image.

Term	Definition
Destination storage resource	Storage resource that is used for disaster recovery in a replication session. This term is also known as a target image.
Fibre Channel (FC) protocol	Protocol used to perform IP and SCSI commands over a Fibre Channel network.
File system	Storage resource that can be accessed through file-sharing protocols such as SMB or NFS.
Internal snapshot (replication snapshot)	The system creates unified snapshots and is part of an asynchronous replication session. These snapshots are only visible in the PowerStore CLI or PowerStore REST API, and manual modification is not possible. Each asynchronous replication session uses up to two internal snapshots that are taken on the source and destination storage resources. Each session also takes up one read/write snapshot on destination storage system. The last successful internal read-only (RO) snapshots for source and destination storage resources and are used as a common base.
iSCSI	Provides a mechanism for accessing block-level data storage over network connections.
Metro Volume	Synchronous replicated PowerStore block volume or volume group that provides active-active access for the connected hosts.
Network-attached storage (NAS) server	File-level storage server used to host file systems. A NAS server is required to create file systems that use SMB or NFS shares.
Network File System (NFS)	An access protocol that allows data access from Linux or UNIX hosts on a network.
PowerStore base enclosure	Enclosure containing both nodes (node A and node B) and 25 NVMe drive slots
PowerStore CLI	Tool that can be installed on an operating system to manage a PowerStore system.
PowerStore cluster	A group of one or more appliances. Up to four PowerStore appliances can be clustered by adding appliances as required.
PowerStore Command Line Interface (PSTCLI)	Tool which can be installed on an operating system to manage a PowerStore system. It allows a user to perform tasks on the storage system by typing commands instead of using the graphic user interface.
PowerStore expansion enclosure	Enclosure that can be attached to a base enclosure to provide additional storage.
PowerStore Manager	Web-based management interface for creating storage resources and configuring and scheduling protection of stored data on PowerStore. PowerStore Manager can be used for all management of PowerStore native replication.
PowerStore node	Storage controller that provides the processing resources for performing storage operations and servicing I/O between storage and hosts. Each PowerStore appliance contains two nodes.

Term	Definition
PowerStore Representational State Transfer (REST) API	Set of resources (objects), operations, and attributes that provide interactive, scripted, and programmatic management control of the PowerStore cluster.
PowerStore Q model	Container-based storage system that is running on purpose-built hardware. This storage system supports unified (block and file) workloads, or block-optimized workloads. The PowerStore Q model supports Quad-Level Cell (QLC) NVMe SSDs for data storage.
PowerStore T model	Container-based storage system that is running on purpose-built hardware. This storage system supports unified (block and file) workloads, or block-optimized workloads. The PowerStore T model supports Triple-Level Cell (TLC) NVMe SSDs for data storage.
RecoverPoint for Virtual Machines	Protects virtual machines (VMs) in a VMware environment with VM-level granularity and provides local or remote replication for any point-in-time recovery. This feature is integrated with VMware vCenter and has integrated orchestration and automation capabilities.
Recovery point objective (RPO)	Acceptable amount of data, which is measured in units of time, that may be lost due to a failure. For example, if a storage resource has a one-hour RPO, data that is written to the storage resource within the last hour may be lost when the replication session is failed over to the destination storage resource.
Recovery time objective (RTO)	Duration of time in which a business process must be restored after a disaster. For example, an RTO of one hour requires restoring data access within one hour after a disaster occurs.
Remote systems	Relationship that is configured between two PowerStore systems to establish a replication session.
Replication session	Relationship that is configured between two storage resources of the same type on different systems, and automatically synchronizes data from one resource to another.
Snapshot	Also called a unified snapshot, a snapshot is a point-in-time view of a storage resource. When a snapshot is taken, it creates an exact copy of the source storage resource and shares all blocks of data with it. As data changes on the source, new blocks are allocated and written to. Unified snapshot technology can be used to take a snapshot of a block or file storage resource.
Storage resource	Top-level object that a user can provision, which is associated with a specific quantity of storage. All host access and data-protection activities are performed at this level. In this document, storage resources refer to resources that support replication such as volumes, volume groups, and thin clones.
Synchronous replication	Replication method in which the host initiates a write to the system at the local site. The data must be successfully stored in both the local and destination systems before an acknowledgment is sent back to the host. Guarantees zero-data-loss RPO while source and destination are synchronized.

Term	Definition
Thin clone	Read-write copy of a thin block storage resource (volume, volume group, or VMware vSphere VMFS datastore) that shares blocks with the parent resource.
Unisphere Manager for RecoverPoint	Web-based interface for managing RecoverPoint replication. It serves as a single pane of glass for replicating storage resources of multiple storage systems that are configured to use RecoverPoint. Consistency groups are created, replicated, and recovered through this interface.
User snapshot	Snapshot that the user creates manually, or a protection policy creates with an associated snapshot rule. This snapshot type is different than an internal snapshot, which the system with asynchronous replication takes automatically.
Virtual Volumes (vVols)	VMware storage framework which allows VM data to be stored on individual Virtual Volumes. This ability allows data services to be applied at a VM-granularity level while using Storage Policy Based Management (SPBM).
Volume	Block-based storage resource that a user provisions. It represents a SCSI logical unit.
Volume group	Storage instance that contains one or more volumes within a storage system. Volume groups can be configured with write-order consistency and help organize the storage that is allocated for particular hosts.
vStorage API for Array Integration (VAAI)	VMware API that allows storage-related tasks to be offloaded to the storage system.
vSphere API for Storage Awareness (VASA)	VMware API that provides additional insight about the storage capabilities in vSphere.
vSphere Metro Storage Cluster (vMSC)	VMware based solution that combines replication with array-based clustering.
Witness	An independent service deployed at a third site which adds additional resiliency to Metro Volume storage availability scenarios.

Metro Volume

Introduction

A Metro Volume provides synchronous replication of spanned block storage volumes or volume groups across two PowerStore clusters in metro distance. This section describes PowerStore native block Metro Volume support in a vSphere Metro Storage Cluster architecture (vMSC).

Note: PowerStoreOS 4.0 extends Metro Volume support to Microsoft and Linux. For more information about Microsoft, see Metro Volume support for Microsoft. For more information about Linux, see Metro Volume support for Linux.

A vMSC provides fully active and workload balanced data centers with resources in a stretched vSphere cluster. The stretched cluster infrastructure can be in same data center, or even across site borders in two different sites in metro distance. Metro Volume configuration provides disaster and downtime avoidance with vSphere High Availability

(HA). For more information about high availability and disaster recovery in a vMSC configuration, see VMware vSphere Metro Storage Cluster (vMSC).

When enabled, this feature allows an operator to configure fully active-active Metro Volumes across two PowerStore clusters running PowerStoreOS 3.0 or later. The stretched architecture allows workload mobility and cross-site automated load balancing. When a Metro Volume is fully synchronized, it provides concurrent host read and write access to the same Metro Volume on both participating PowerStore clusters. An embedded polarization mechanism protects against a split-brain situation during a failure scenario. After the failure has been corrected, PowerStore starts a self-healing process to turn the Metro Volume back into an active-active state.

The PowerStoreOS 3.0 release supports configuring Metro Volume with standard volumes only. The Metro Volume feature uses the common remote-system configuration for replication management and replication data traffic through an Ethernet (LAN) connection. The following sections show the configuration in PowerStore Manager, although using the PowerStore CLI and REST API are also supported. The following subsections discuss these topics:

- Features
- Licensing requirements for the Metro Volume
- How Metro Volume feature works
- Configurations supported for Metro Volume
- PowerStore Manager: Metro Volume configuration and management

Features

Symmetric active-active Metro Volume architecture: Read and write I/O can occur directly on either PowerStore cluster hosting the Metro Volume. Bi-directional synchronous replication ensures the volumes on both clusters are synchronized during normal operation.

Non-disruptive life cycle management: You can provision and decommission Metro Volumes non-disruptively. Other configuration tasks, such as modifying a preferred role, do not interfere with application I/O.

Snapshot support: User snapshots created by a Protection Policy on Metro Volumes are available on each PowerStore cluster hosting a Metro Volume. You can use snapshots for various reasons such as creating thin clones, or performing volume-refresh and restore operations. During the active-active state of a Metro Volume, the snapshots are taken simultaneously on both PowerStore clusters and result in near-identical snapshots. For snapshots taken when Metro Volume is not in the active-active state, the snapshots are replicated to the peer system after self-healing reestablishes the Metro Volume.

Starting with PowerStoreOS 3.5, the secure snapshot setting can be enabled for snapshots on volumes and volume groups. With secure snapshots enabled, the snapshots and parent resource are protected from accidental or malicious deletion and serve as a cost-effective line of defense against ransomware attacks. If an unauthorized user gains access to a system, the attacker cannot delete secure snapshots and cause data loss.

Host connectivity: Metro Volume supports VMware ESXi FC or iSCSI front-end connectivity with Uniform or Non-Uniform storage presentation.

VMware integration: Metro Volume is designed to work in tandem with vSphere High Availability (HA). Also, it is compliant with ALUA states and integrates with vSphere storage multipathing.

Uniform topology path cost awareness: When using uniform storage presentation, Metro Volume supports granular ALUA states for both equidistant and non-equidistant paths.

Proactive use cases: Metro Volume supports proactive use cases such as planned maintenance, disaster avoidance, load balancing, and migration.

Pause: You can pause Metro Volume protection for maintenance operations and resume it later.

Self-healing: Metro Volumes are synchronized and recovered to an active-active state automatically.

Licensing

Metro Volume replication is included in the basic license at no extra cost for supported PowerStore clusters.

Theory of operation

You can use each standard volume on a PowerStore cluster to create a Metro Volume that spans to a remote PowerStore system. The option to configure Metro Volume is available in PowerStore Manager > Volume overview, or in the subtab Protection > Metro Volume > Volumes details, and it requires a remote system configuration. When you select the Remote System and the Metro Volume creation starts, PowerStore configures a metro replication session and starts the initial sync of the volume to the remote system. When the Remote System is a multi-appliance PowerStore cluster, PowerStore can place the volume automatically on the best applicable appliance on the remote PowerStore cluster. Alternately, PowerStore Manager allows you to select a specific appliance in the Metro Volume configuration. During initial synchronization, even though the volume is not usable on a remote PowerStore for host I/O, the new Metro Volume could already be mapped to hosts. For initial synchronization, PowerStore uses the same snapshot-based asynchronous replication process as described in the document PowerStore: Replication Technologies. While the Metro Volume is not fully synchronized between the PowerStore clusters, asynchronous replication cycles are continuously running. For final synchronizations, the replication changes from an incremental-based replication to a differential synchronization.

When the Metro Volumes are fully synchronized, PowerStore changes into the active-active state and enables host I/O on the remote PowerStore cluster. The duration for initial synchronization depends on the amount of data on the source and host write activity to the volume. A single internal snapshot is created after initial synchronization. The internal snapshot is used as a common base when a Metro Volume must be resynchronized, for instance, after a pause or fracture. If a new sync is required, PowerStore uses the common base for an asynchronous replication until Metro Volume is synchronized on both PowerStore clusters and switches into the active-active state. When the status shows **Operating Normally (active-active),** hosts can perform concurrent I/O on both PowerStore clusters simultaneously. PowerStore controls the active path with

implicit Asynchronous Logical Unit Access (ALUA). Each Metro Volume on an appliance has ALUA active-optimized paths to the volume on one node. It also has ALUA active-non-optimized paths to the same volume on the other nodes. However, a host can use multiple paths to both nodes for I/O. The node affinity of the individual volume is the determining factor to select the active-optimize path for an appliance.

Polarization

One of the most critical situations for a Metro Volume is a **split-brain** scenario when both volumes of a Metro Volume are active simultaneously. This situation can lead to having different data on both volumes and may require manual intervention. To prevent a split-brain situation, PowerStore uses polarization for failure handling. For each Metro Volume, one volume is assigned with a preferred role while the peer side is configured with a non-preferred role. The initial preferred side of a Metro Volume is the side where the Metro Volume replication session configuration was performed. The following figure shows an example of the distributed, preferred side of Metro Volumes.

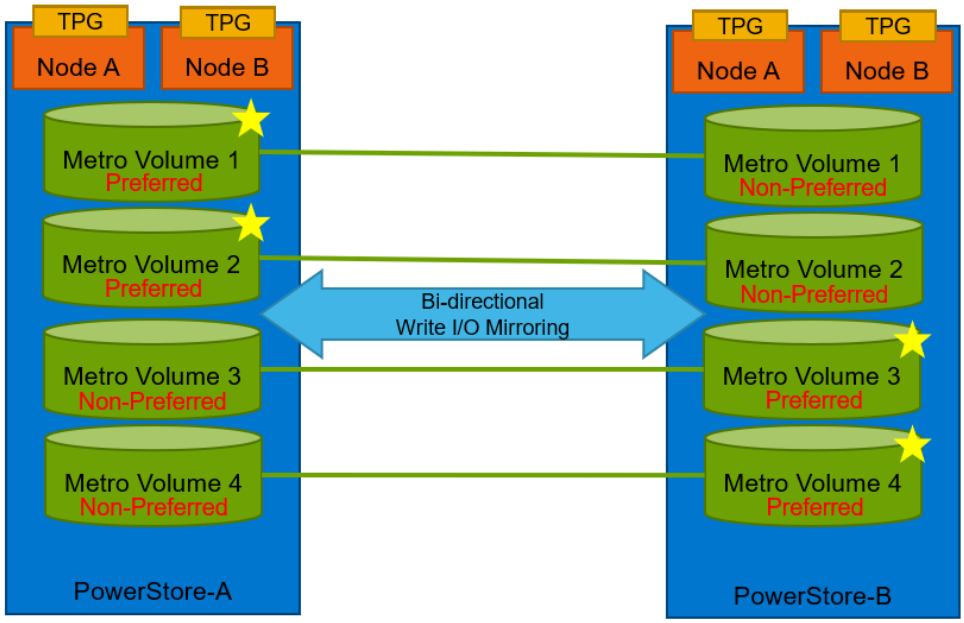


Figure 1. Distributed Metro Volume roles

You can change the preferred role to non-preferred, and conversely, in PowerStore Manager. This action is non-disruptive for the hosts when a Metro Volume is in the status **Operating Normally**. While the Metro Volume session is running with the status active-active, the PowerStore cluster must ensure I/O can be replicated to the peer PowerStore appliance. Both clusters constantly check the peer cluster and only serve I/O on hosts at the non-preferred side when the preferred side is available. Polarization occurs when a side is not responding, and it is leading to a status of **Fractured** for affected Metro Volumes. When polarization is invoked and the preferred side determines that the non-preferred is not available, the replication traffic is stopped and host I/O on the preferred side continues. On the non-preferred side, hosts do not receive acknowledgments for their writes, the I/O stops, and paths become ALUA unavailable. In this situation when paths are unavailable, the ESXi server shows an All Path Down (APD) event for affected datastore volumes. Depending on the host connectivity, the affected VMs continue or VMware HA must restart VMs on hosts with paths to the preferred volume.

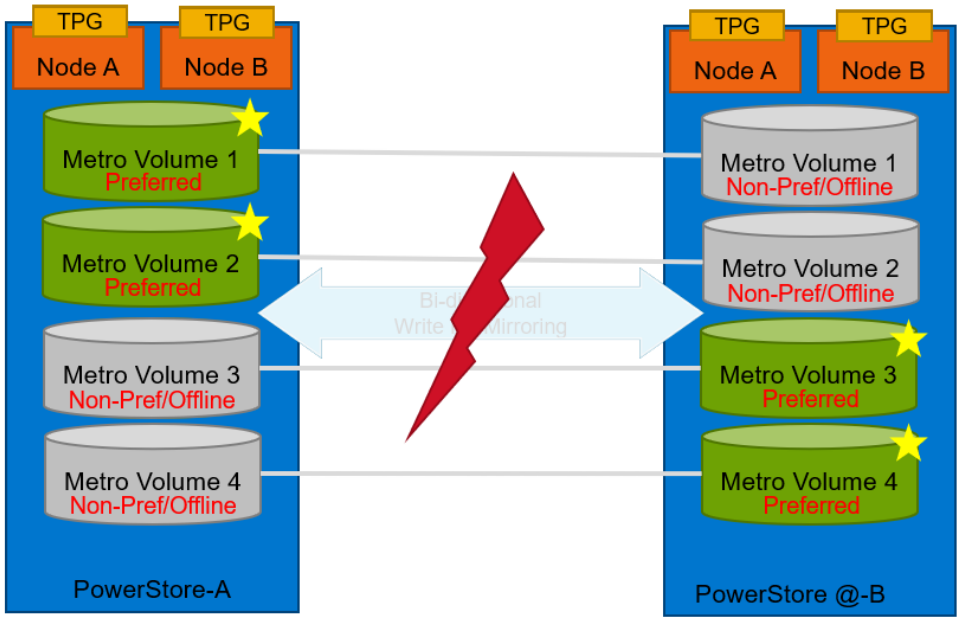


Figure 2. Polarization during failure scenario

After the connection is reestablished, PowerStore initiates self-healing of the Metro Volume with data on the most current volume. Without manual intervention, the preferred volume is seen as the most current, and self-healing initiates the synchronization from the preferred side to the non-preferred side. Self-healing uses the last common base and does not require a full synchronization.

During polarization, only hosts with paths to the preferred side can access the volume. When a Metro Volume is not in an expected state for host access, you can use promote and demote operations to override the preferred or non-preferred role for a Metro Volume. The promote operation enables host I/O on the non-preferred side of a Metro Volume, and the demote operation disables host I/O on the preferred side. After a promote operation, PowerStore determines the data on the promoted side to be more current than data on preferred side, and starts self-healing from the promoted volume to demoted volume. An unwanted manual intervention with the promote operation could lead to a split-brain situation. It could lead to data loss because production data on the preferred could be overwritten by data on promoted side during self-healing. PowerStore helps to mitigate the potential risk of data loss by performing a snapshot before running the promote and demote actions, and when self-healing begins.

Host connectivity

In a Metro Volume configuration, PowerStore supports uniform and non-uniform host access. During host registration, the operator can choose the host connectivity which also sets the ALUA states for the individual paths.

In a non-uniform host connectivity configuration (shown below), each host has paths to only a single PowerStore appliance. For instance, hosts in Site-A are mapped to only the PowerStore cluster in Site-A, and hosts in Site-B are mapped to only the PowerStore cluster in Site-B. Since the Metro Volume is spanned across both PowerStore clusters, the mapped Metro Volume is the same to all hosts in the cluster. The following sections use the terms Site-A and Site-B to identify the different sites when running Metro Volume across different locations (two fault domains). However, PowerStore Metro Volume is also

supported when the participating PowerStore systems are installed in a single location side by side (single fault domain).

Non-uniform host connectivity

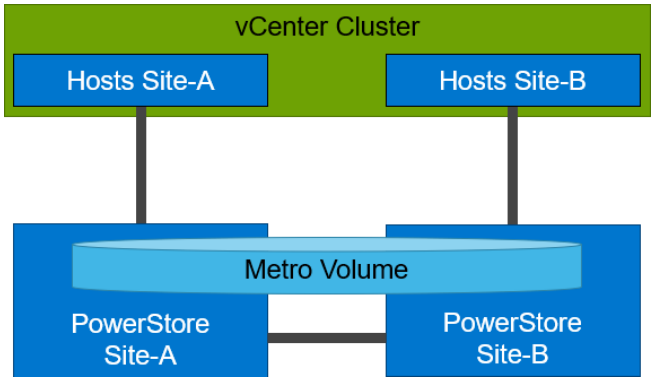


Figure 3. Non-uniform host connectivity

In a setup with uniform host connectivity (shown below), each host has active paths to both PowerStore clusters. For instance, all hosts in Site-A and Site-B are mapped to PowerStore in Site-A and Site-B. This configuration provides extra resilience. If there is an array failure or link-loss situation, VMs can continue running on any hosts by using paths to the non-affected array with the preferred volume.

Uniform host connectivity

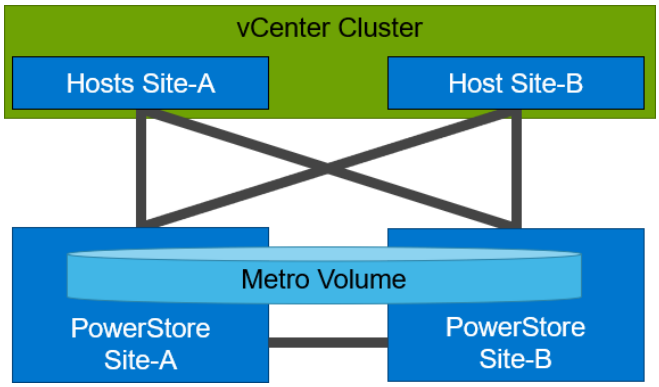


Figure 4. Uniform host connectivity

Host configuration

PowerStore Manager enables you to select the host connectivity type when registering a new host. The following diagrams show the possible host configuration.

Local connectivity: This configuration provides host and application access to standard volumes and to the Metro Volume exclusively in this storage system. Use this configuration for standard volumes, and for Metro Volumes on the PowerStore in a non-uniform host setup. The following example shows that host H1 is connected to the PowerStore cluster PS1 in a non-uniform host configuration.

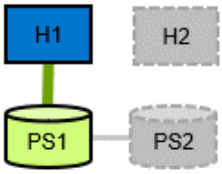


Figure 5. Local connectivity

Host is co-located with local system: Use this option with Metro Volumes that have non-equidistant paths in a uniform storage presentation when the host and PowerStore are local to each other. In other words, the host and PowerStore are in the same data center or are connected optimally. Use this option for **local** hosts with lower latency than the **remote** hosts. In this configuration, hosts always attempt to send I/O to the metro volume on this system except in failure situations, and it results in an active-optimized path for the hosts. As shown below, you can use this configuration for a uniform host connection on PowerStore cluster PS1 for host H1 which is in the same location.

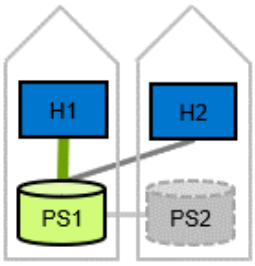


Figure 6. Host co-located with local system

Host is co-located with the remote system: Use this option with Metro Volumes that have non-equidistant paths in a uniform storage presentation when the host and PowerStore are not local to each other. In other words, the host and PowerStore are in different data centers with significant latency in between. Use this option for the hosts which are remote to the PowerStore cluster with a higher latency than the **local** hosts. The host only sends I/O to the Metro Volume on this system in failure situations and results in active-non-optimized paths for the hosts. As shown below, you can use this configuration for a uniform host connection on PowerStore cluster PS1 for host H2 which is in the remote location with higher latency.

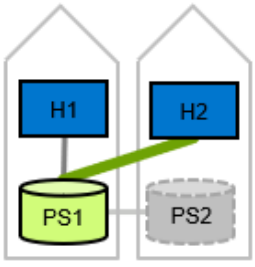


Figure 7. Host co-located with remote system

Host is co-located with both systems: Use this option with Metro Volumes that have equidistant paths in a uniform storage presentation when the host and both PowerStore clusters are local to each other. In other words, the host and both PowerStore clusters are in the same data center or are connected optimally. The host uses its own multipath configuration to determine the best path for I/O. This option is useful for configurations

when all hosts have same latency to the PowerStore cluster. If you use this option for both PowerStore clusters, it results in active-optimized paths to both ends of a Metro Volume. You can use this option when hosts H1 and H2 and PowerStore clusters PS1 and PS2 are in the same location with the same latency.

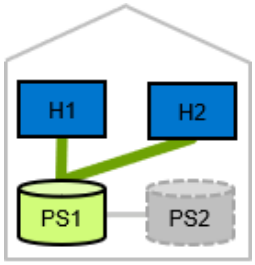


Figure 8. Host co-located with both systems

With different metro connectivity options for hosts in a uniform setup, the paths are presented to the hosts with the following ALUA states. PowerStore uses implicit ALUA to provide optimized-path information for connected hosts. For load-balancing, PowerStore selects one node of the appliance as the optimized node for a volume with node affinity. The assigned node affinity information for each volume is available in the Volume overview after adding you add the column **Node Affinity**. The PowerStore CLI and PowerStore REST API enable you to manually change the node affinity setting.

Note: It is required to use consistent LUN numbers for standard volume mappings and Metro Volume mappings across hosts within the same vSphere cluster, hosts within other vSphere clusters, or hosts not in a cluster. For additional information, please see the following references:

Dell KB Article 000191503 PowerStore: Inconsistent Logical Unit Numbers between hosts could result in data access or data consistency issues

VMware vSphere Product Documentation: vSphere Storage | Setting LUN Allocations | Storage provisioning

---

Starting with PowerStoreOS 3.5, user interface enhancements in PowerStore Manager provide diagrams for each of the host connectivity options. The diagrams describe the locality of the host and the I/O pattern that can be expected relative to the PowerStore appliances configured for Metro Volume. A solid line between host and PowerStore indicates an optimized path. A dotted line between host and PowerStore indicates a non-optimized path.

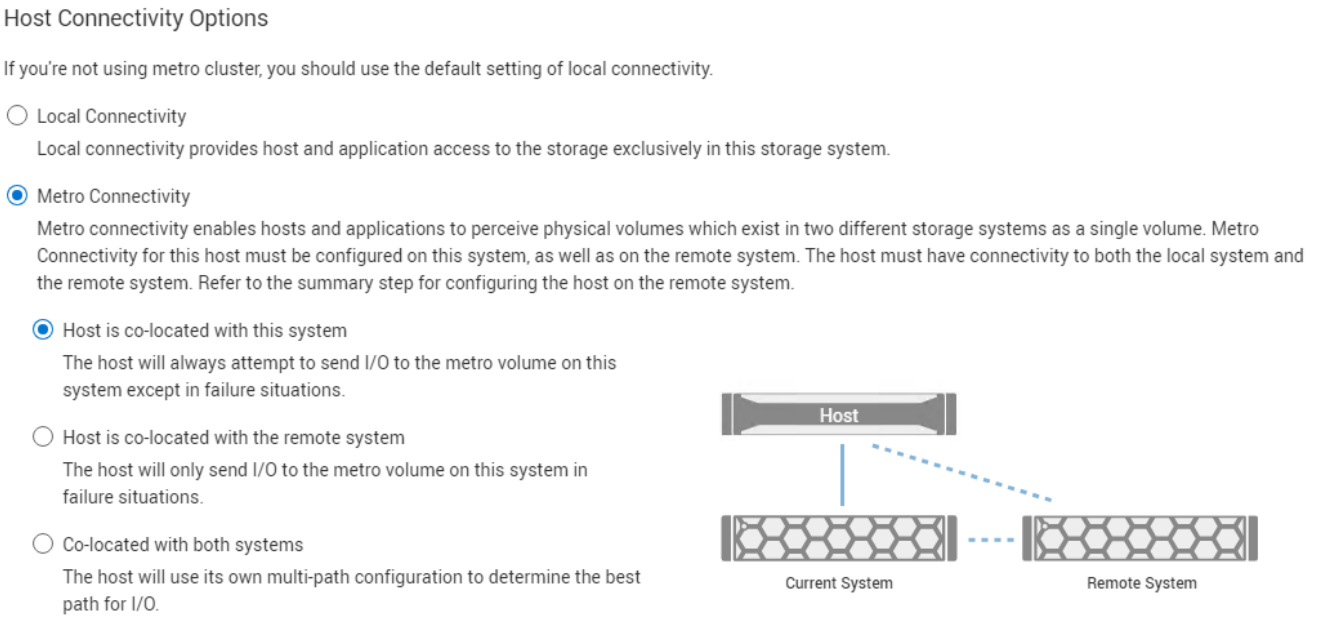


Figure 9. Host Connectivity Options shown in PowerStoreOS 4.0

The following table outlines the valid metro host connectivity configuration possibilities.

Table 2. Metro host connectivity configuration possibilities

Physical/Virtual HostA instance Definition		
HostA registered with specific connectivity type on PStoreA (Preferred role for MetroVolume1)	**HostA registered with specific connectivity type on PStoreB (Non-Preferred role for MetroVolume1)**	**Connectivity**
Uniform-Local	Uniform-Remote	Valid
Uniform-Remote	Uniform-Local	Valid
Uniform-Equidistant	Uniform-Equidistant	Valid
Non-Uniform	--- not mapped to same physical/virtual host instance as metro peer	Valid
--- not mapped to same physical/virtual host instance as metro peer	Non-Uniform	Valid

The configuration in the following table assumes that the volume has selected node A for volume node affinity with Metro Volume role preferred on site B.

Table 3. ALUA path states for uniform metro connectivity for a metro

Metro connectivity		PowerStore cluster site-A		PowerStore cluster site-B	
Host site-A	Host site-B	Node A	Node B	Node A	Node B
Co-located with **both** systems	Co-located with **both** systems	**Active-optimized**	**Active-non-optimized**	**Active-optimized**	**Active-non-optimized**
Co-located with **local** system	Co-located with **remote** system	**Active-optimized**	**Active-non-optimized**	**Active-non-optimized**	**Active-non-optimized**
Co-located with **remote** system	Co-located with **local** system	**Active-non-optimized**	**Active-non-optimized**	**Active-optimized**	**Active-non-optimized**
During failure scenario, or standard volumes[1]					
Co-located with **local** system	Co-located with **remote** system	**Unavailable**	**Unavailable**	**Active-optimized**	**Active-non-optimized**
During node failure PowerStore node A at site-A					
Co-located with **local** system	Co-located with **remote** system	**Unavailable**	**Active-non-optimized**[2]	**Active-non-optimized**	**Active-non-optimized**
Co-located with **remote** system	Co-located with **local** system	**Unavailable**	**Active-non-optimized**	**Active-optimized**	**Active-non-optimized**

1. When Metro Volumes and standard volumes are shared on the same system with host setting **Co-Located with remote system,** ALUA paths from the host to standard volumes operate like local connectivity.
2. To mitigate latency issues, we recommended enabling NMP-latency-based round robin which is covered in the next section.

ESXi path selection policies (PSP)

vSphere ships with three MPIO path selection policies (PSP): Most Recently Used (MRU), Round Robin (RR), and Fixed. MRU is the default policy for most active-passive storage devices and is not used with PowerStore.

When the host connectivity feature is configured correctly, the best path selection policy to choose with Metro Volume is Round Robin. Round Robin is the easiest PSP to configure and provides both optimal I/O performance and a balance of the storage fabric in uniform or non-uniform storage presentation. Also, it provides an automatic failover capability to non-optimal paths and an automatic failback capability to recovered optimal paths.

For the ESXi host, the working path state is available in Configure > Storage Device.

The following figure shows how vCenter displays the paths for a uniform host configuration with the settings of Co-Located with Local (target ending with …fnm102135) and Co-Located with Remote (target ending with …fnm102139). The number of total paths shown in the vSphere UI is determined by multiplying the number of host initiators by the number of target port groups on PowerStore.

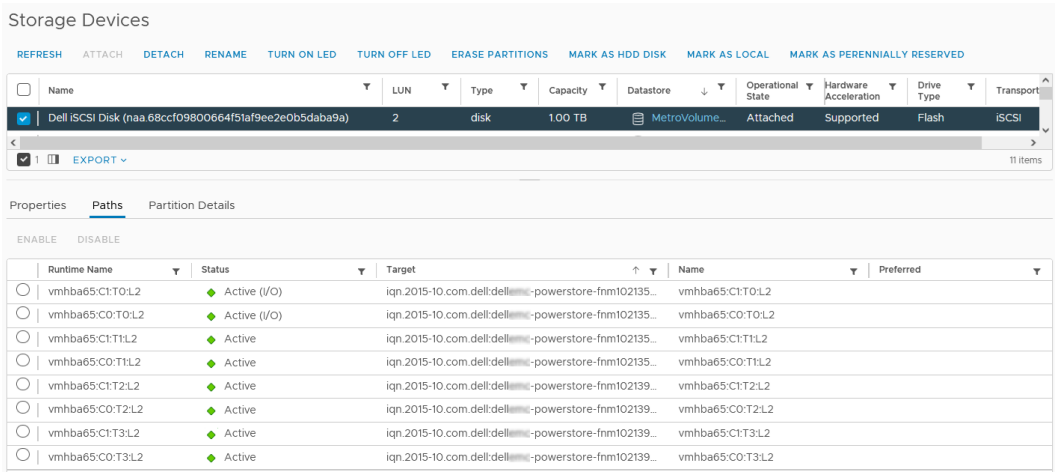


Figure 10. vCenter > Storage Devices > Paths view

VMware NMP chooses the working path and indicates the status in vCenter. The NMP selected working path may not always be the Active-Optimized path. Possible path states in vCenter are described in the table below.

Table 4. Path states in vSphere Client

State in vCenter	Description
Active (I/O)	Indicates the working path for a volume. It might be the Active-Optimized path, or another active path selected by NMP
Active	Indicates an active path for a volume.
Dead	Path is not available for host I/O to the volume.

The esxcli provides more detail for the paths used for a device. For a particular device mapped to an ESXi host, you can use the following command to show the path details. The device id (naa.68ccf…) is available either in the PowerStore Manager volume overview, in vSphere, or by using esxcli. The following figure shows two commands how to access the information. While **device list** gives a comprehensive overview of paths for the device, the option path list shows details for all paths to a volume and includes the runtime name as used in vCenter UI.

```
[root@esx-c:~] esxcli storage nmp device list -d naa.68ccf09800664f51af9ee2e0b5daba9a
naa.68ccf09800664f51af9ee2e0b5daba9a
   Device Display Name: Dell[illegible] iSCSI Disk (naa.68ccf09800664f51af9ee2e0b5daba9a)
   Storage Array Type: VMW_SATP_ALUA
   Storage Array Type Device Config: {
                implicit_support=on;
                explicit_support=off;
                explicit_allow=on;
                alua_followover=on;
                action_OnRetryErrors=off;
                {TPG_id=1,TPG_state=AO}
                {TPG_id=2,TPG_state=ANO}
                {TPG_id=61441,TPG_state=ANO}
                {TPG_id=61442,TPG_state=ANO}}
   Path Selection Policy: VMW_PSP_RR
   Path Selection Policy Device Config: {
                policy=rr,iops=1,bytes=10485760,useANO=0;
                lastPathIndex=3: NumIOsPending=0,numBytesPending=0}
   Path Selection Policy Device Custom Config:
   Working Paths: vmhba65:C0:T0:L2, vmhba65:C1:T0:L2
   Is USB: false
[root@esx-c:~] esxcli storage nmp path  list -d naa.68ccf09800664f51af9ee2e0b5daba9a
iqn.1998-01.com.vmware:esx-c:1071653994:65-00023d000002,iqn.2015-10.com.dell:dellemc-
powerstore-fnm10213500520-a-53f52a7b,t,1-naa.68ccf09800664f51af9ee2e0b5daba9a
   Runtime Name: vmhba65:C1:T0:L2
   Device: naa.68ccf09800664f51af9ee2e0b5daba9a
   Device Display Name: Dell[illegible] iSCSI Disk (naa.68ccf09800664f51af9ee2e0b5daba9a)
   Group State: active
   Array Priority: 0
   Storage Array Type Path Config: {TPG_id=1,TPG_state=AO,RTP_id=27,RTP_health=UP}
   Path Selection Policy Path Config: PSP VMW_PSP_RR does not support path configuration.

iqn.1998-01.com.vmware:esx-c:1071653994:65-00023d000001,iqn.2015-10.com.dell:dellemc-
powerstore-fnm10213500520-a-53f52a7b,t,1-naa.68ccf09800664f51af9ee2e0b5daba9a
   Runtime Name: vmhba65:C0:T0:L2
   Device: naa.68ccf09800664f51af9ee2e0b5daba9a
   Device Display Name: Dell[illegible] iSCSI Disk (naa.68ccf09800664f51af9ee2e0b5daba9a)
   Group State: active
   Array Priority: 0
   Storage Array Type Path Config: {TPG_id=1,TPG_state=AO,RTP_id=27,RTP_health=UP}
   Path Selection Policy Path Config: PSP VMW_PSP_RR does not support path configuration

. . . detailed information for remaining paths removed . . .
```

Figure 11. Esxcli mapping TPG_id to Runtime Name

Note: **TPG state** identifies the active-optimal and active-non-optimal ALUA state for each target port group. PowerStore has one target port group per node. **Working Paths** identifies each active-optimal path in use as derived from the target port groups. **Working Paths** corresponds with **Active (I/O)** in the vSphere Client UI.

The following table shows the different possible TPG_states which also represent the PowerStore provided ALUA path information.

Table 5. Esxcli path states

TPG_state in esxcli	ALUA path state	Description
AO	Active-optimized	Indicates the optimized path for host I/O to the volume and can be controlled with Volume Node-Affinity setting in PowerStore CLI or REST API.
ANO	Active-non-optimized	Path is available, but not optimized for host I/O to the volume.

TPG_state in esxcli	ALUA path state	Description
UNAVAIL	Unavailable	Path is not available for host I/O to the volume.

If there is a single-node failure, PowerStore does not switch the Active-Optimized path to the remaining node. With the default iops-based NMP, NMP round-robin selects all remaining active-non-optimized as working path (see below).

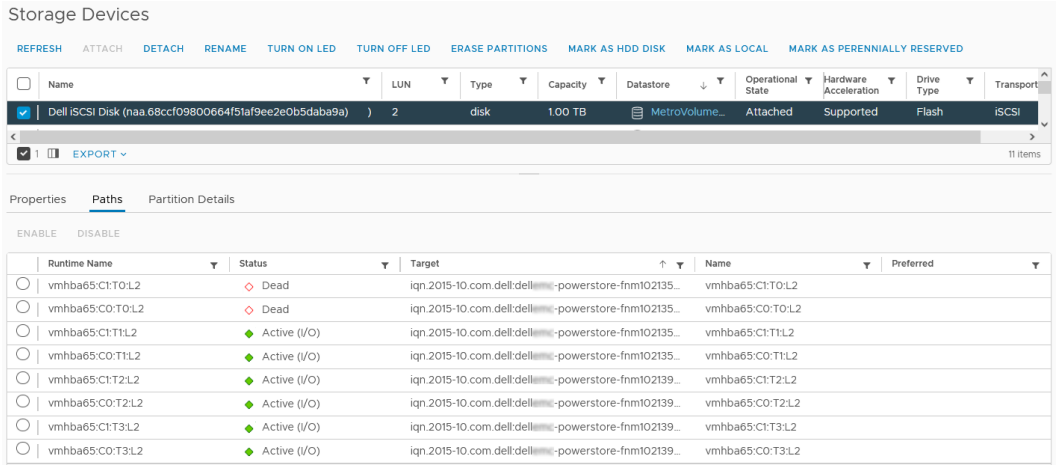


Figure 12. vCenter path states during a node failure

This scenario could lead to a performance impact because there is probability that NMP chose the remote PowerStore cluster with higher latency for I/O. For mitigation, you can change the device configuration to NMP round-robin with the latency mechanism. The commands in the figure below show an example of the commands to check (deviceconfig get) and change (deviceconfig set) the NMP round-robin mechanism to latency-based round-robin.

Note: The below command is applied per-host and per-device.

```
[root@esx-c:~] esxcli storage nmp psp roundrobin deviceconfig get -d naa.68ccf0980083cb0d76feb73e91b908cf
   Byte Limit: 10485760
   Device: naa.68ccf0980083cb0d76feb73e91b908cf
   IOOperation Limit: 1
   Latency Evaluation Interval: 0 milliseconds
   Limit Type: Iops
   Number Of Sampling IOs Per Path: 0
   Use Active Unoptimized Paths: false

[root@esx-c:~] esxcli storage nmp psp roundrobin deviceconfig set --type=latency -d naa.68ccf0980083cb0d76feb73e91b908cf

[root@esx-c:~] esxcli storage nmp psp roundrobin deviceconfig get -d naa.68ccf0980083cb0d76feb73e91b908cf
   Byte Limit: 0
   Device: naa.68ccf0980083cb0d76feb73e91b908cf
   IOOperation Limit: 0
   Latency Evaluation Interval: 180000 milliseconds
   Limit Type: Latency
   Number Of Sampling IOs Per Path: 16
   Use Active Unoptimized Paths: false
```

Figure 13. Change NMP round-robin to latency mechanism

You can use a claim rule to implement the latency policy for PowerStore devices. You must add this claim rule to each ESXi host. After you apply the claim rule, each newly discovered device has the claim rules applied to it (devices already discovered before the claim rule is applied will not).

Note: The following commands are for vSphere 7/8 ESXi hosts. ESXi 6.7 hosts should also include the **disable_action_OnRetryErrors** option. See the *PowerStore Host Configuration Guide* for more information.

```
esxcli storage nmp satp rule add -c tpgs_on -e "PowerStore" -M
PowerStore -P VMW_PSP_RR -O "policy=latency" -s VMW_SATP_ALUA -t
vendor -V DellEMC
```

You can also add the claim rule to ESXi hosts using the PowerCLI.

```
# Add or remove a claim rule on each vSphere host
$esxlist | ForEach-Object {
$esxcli = Get-EsxCli -VMHost $_ -V2

# Fill the hash table (optional params are not required)
$sRule = @{
   satp = 'VMW_SATP_ALUA' #esxcli: -s
   psp = 'VMW_PSP_RR' #esxcli: -P
   pspoption = 'iops=1' #esxcli: -O
   claimoption = 'tpgs_on' #esxcli: -c
   #option = 'disable_action_OnRetryErrors' #esxcli: -o
   vendor = 'DellEMC' #esxcli: -V
   model = 'PowerStore' #esxcli: -M
   description = 'PowerStore' #esxcli: -e
}

# Call the esxcli command to add/remove the rule
Write-Host $selection "rule on" $_
$esxcli.storage.nmp.satp.rule.$selection.Invoke($sRule)
```

The Fixed PSP may be desirable when a preferred path on the storage fabric should be used. As shown in the figure below, you should set the MPIO policy on the vSphere host to Fixed with the preferred path leading to vmhba3:C0:T1:L3. Using the Fixed PSP generally requires more administrative effort to implement, maintain, and document.

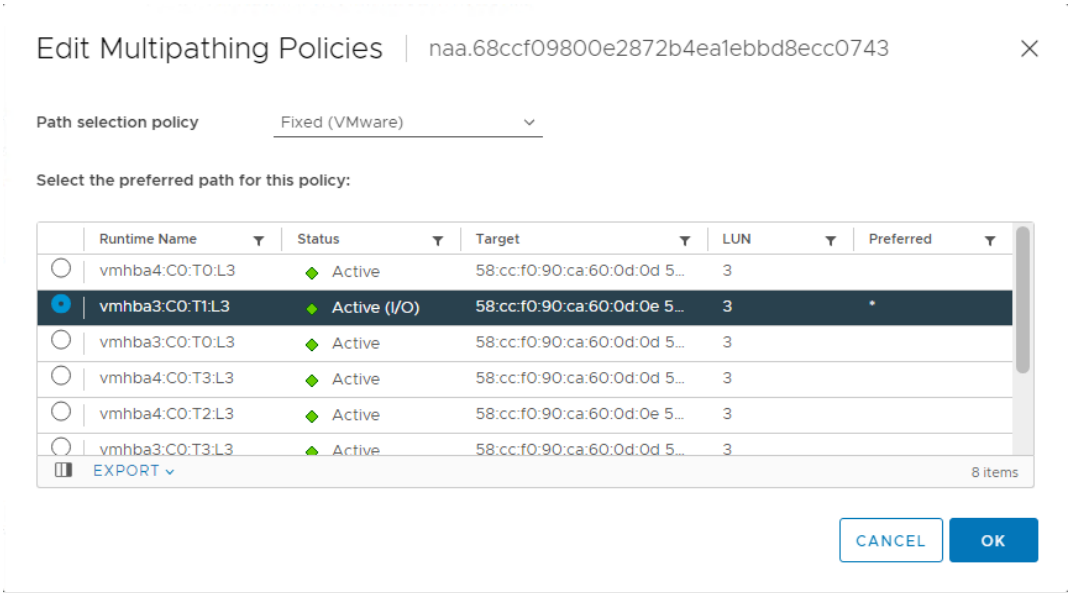


Figure 14. Fixed path selection policy

If host connectivity is modified to a new value, causing I/O to be sent down an active/non-optimal path, PowerStore is designed to report the ALUA path state changes to the vSphere host using the following process:

1. Host connectivity is modified to a new value causing I/O to be sent down an active-non-optimal path.
2. The vSphere hosts send Round Robin I/O to an active/non-optimal path.
3. PowerStore fails the I/O with a Unit Attention Check condition.
4. The vSphere host requests Report Target Port Groups.
5. PowerStore responds with new ALUA state changes.
6. vSphere begins Round Robin I/O over new optimal paths.

Note: You can view the **Unit Attention Check Condition** in **/var/log/vmkernel.log** in the following example. For more information, see the VMware KB article, Interpreting SCSI sense codes in VMware ESXi and ESX (289902).

```
2022-04-28T01:53:52.332Z cpu11:2097789)NMP:
nmp_ThrottleLogForDevice:3867: Cmd 0x89 (0x45b8c3883988, 2097177)
to dev "naa.68ccf0980080631b71144f5e582abe31" on path
"vmhba3:C0:T0:L2" Failed:

2022-04-28T01:53:52.332Z cpu11:2097789)NMP:
nmp_ThrottleLogForDevice:3875: H:0x0 D:0x2 P:0x0 Valid sense data:
0x6 0x2a 0x6. Act:FAILOVER. cmdId.initiator=0x430541297ec0 CmdSN
0x1b191
```

In previous testing, we discovered that vSphere did not consistently or reliably begin using the new optimal paths immediately after an ALUA state changed had occurred. Instead, up to five minutes elapsed before vSphere recognized the ALUA path state change. This

means that vSphere may continue to follow the Round Robin policy over non-optimal paths for up to five minutes.

If you observe this condition and are concerned about this behavior occurring in your environments, there are two workarounds:

- After a known ALUA state change, perform a vSphere storage rescan.

  Reduce the vSphere host advanced setting **Disk.PathEvalTime** from the default of 300 seconds down to an acceptable automatic storage rescan interval. If you choose this approach, you must perform this action on each vSphere host where the host connectivity was changed.

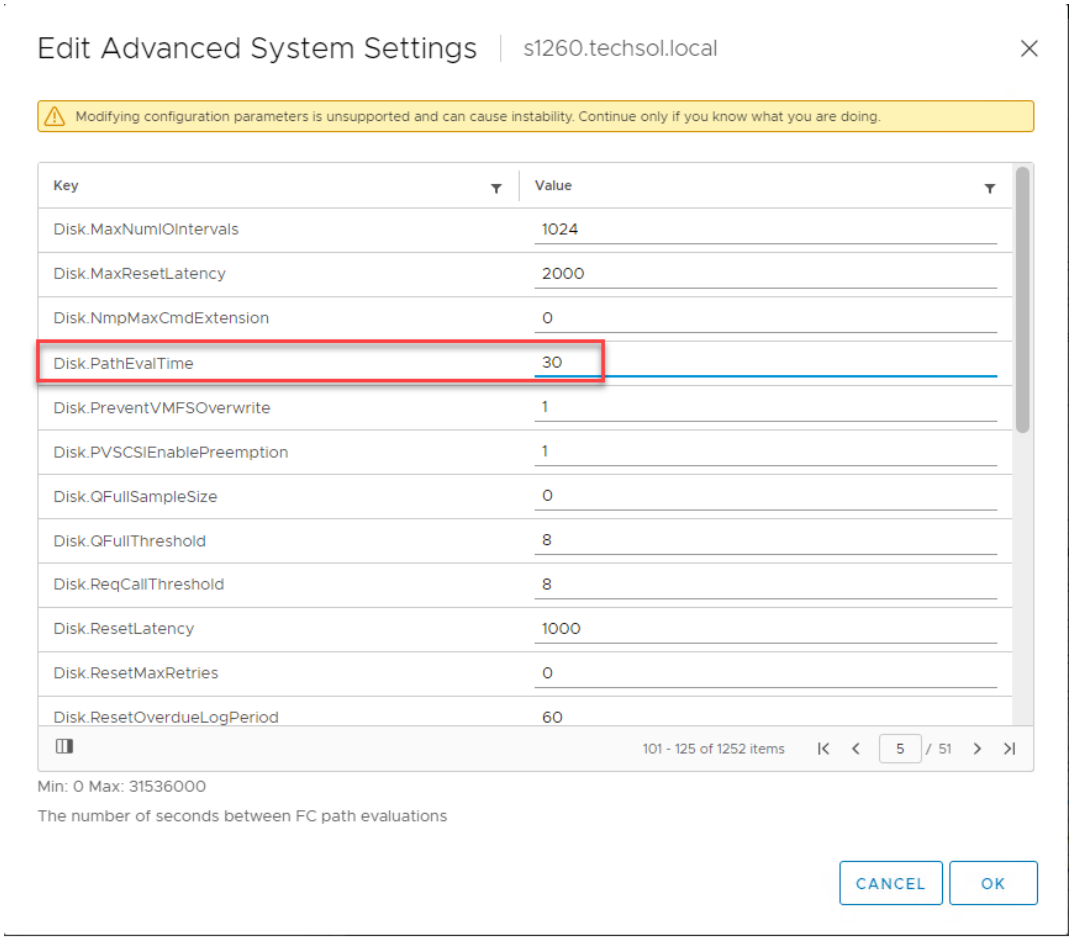


Figure 15. Example: Reducing the Disk.PathEvalTime setting to 30 seconds on a vSphere host

Replication states

This section describes the different major states for a Metro Volume. Some of the following states show more information in parentheses in the UI.

Operating Normally

In this state, PowerStore Metro Volume is fully synchronized and active-active. Both sides of the Metro Volume serve host I/O, and a host only receives an acknowledgment from the array when the I/O is committed on both sides of the Metro Volume. Possible sub-states are **Active-Active**, and **Asynchronously Synchronizing** when reprotecting starts from a promoted volume.

Switching to Metro Sync

PowerStore orchestrates the copy and mirror operations. This state only allows host I/O to the preferred or promoted side of a Metro Volume. Possible substates are **Asynchronously Synchronizing**, and **Committing to Active-Active.**

Fractured

The peer PowerStore systems could not synchronize I/O. This state could be caused by an array failure or link failure required for replication. This state only allows host I/O to the preferred or promoted side of a Metro Volume.

Paused

In this state, the Metro Volume is paused. The state is triggered by the operation **Pause** in PowerStore Manager. This state only allows host I/O to the preferred or promoted side of a Metro Volume.

Resuming

This is a transient state after you resume a paused Metro Volume and trigger the state **Switching to Active-Active.** This state only allows host I/O to the preferred or promoted side of a Metro Volume.

Reprotecting

This state indicates that PowerStore initiated the Metro Volume self-healing recovery for Metro Sync replication to return to the active-active state. This state only allows host I/O to the preferred or promoted side of a Metro Volume.

Modify role in progress

Replication is in this state when it is swapping the roles of preferred and non-preferred volumes of metro sync session.

Deleting

This state occurs after a Metro Volume is being unconfigured after an **End Metro Volume** state in PowerStore Manager.

Replication Metrics

For each Metro Volume replication session, PowerStore Manager provides performance metrics. One graph shows the replication traffic bandwidth for the selected Volume (shown at the top of the following figure). A second graph shows the remaining data to be synchronized when running an initial synchronization or self-healing (shown at the bottom of the following figure). During the active-active state, the remaining data graph shows only a flat line.

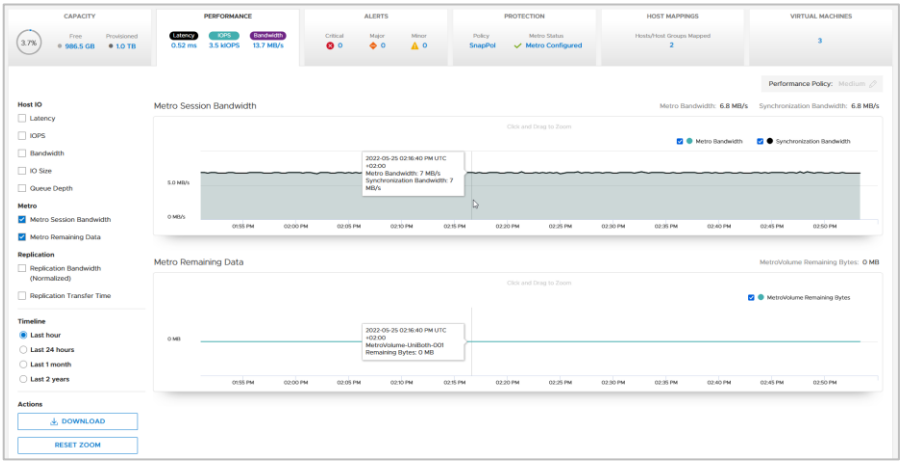


Figure 16. Metro session performance metrics

Metro Volume operations

This section describes the operations in PowerStore Manager for a Metro Volume. You can run an operation for a Metro Volume in various screens. Apart from the actions to Configure and End Metro Volume, which are also available in the Volumes overview screen, the operations are available in Volume details view. This view is available in Volumes > [Volume Name] > Volume details > Protection > Metro Volume. The operations are also available in the Metro session overview (Protection > Metro). When an operation with impact on host I/O is supposed to be performed, a dialog box appears which shows a detailed description of the effective Metro Volume state after the operation is executed.

Configure Metro Volume

You can start a new Metro Volume configuration in Volumes Overview or in the Volume details view in the subtab for Protection > Metro Volume. The wizard only requires the remote system to define the peer PowerStore cluster. (See more details about the remote system and replication protocol in the document PowerStore: Replication Technologies.) When the peer is a multi-appliance cluster, there is an extra field to specify the appliance. Without selecting an appliance, PowerStore balances the Metro Volumes across available appliances. When no migration or replication is configured for the volume, configuring Metro Volume is possible in any state of a standard volume. When a Metro Volume is configured, a Metro replication session is created. The Metro replication session initiates the initial synchronization for the Volume.

Set local role to preferred or modify preferred role

These operations swap the preferred role for a Metro Volume and could be performed once a Metro Volume reached the state **Operating normally (Active-Active).** Changing the role could be required to set the same role for volumes relating to the same application to a single PowerStore cluster. If there is a failure scenario, all volumes of the same application could be affected by polarization. Another use case would be changing the role before a planned maintenance in one of the Metro Volume sites or pausing a Metro Volume.

Pause

To modify some Metro Volume properties, you must pause a Metro Volume. The following operations require a Metro Volume to be paused: volume resize, volume restore from

snapshot, volume refresh from an object within the clone family (clone or snapshot), and volume rename.

Pausing a Metro Volume could also be used during a planned maintenance. Be aware that pausing a Metro Volume disables host I/O on the non-preferred side of a Metro Volume. For non-uniform host configuration, we recommended ensuring that hosts with running VMs are using the preferred volume to avoid an unwanted restart of VMs.

The Metro Volume data synchronization between PowerStore-A and PowerStore-B will be paused. On the preferred system, production access will continue for applications. The local protection policy will remain active. On the non-preferred system, production access will remain disabled for hosts and applications. The local protection policy will remain paused.

Resume

This operation resumes a paused Metro Volume and starts switching into an active-active state. The host writes and snapshots created on the preferred volume while the metro volume was paused, will be synchronized to the peer system.

Promote

A promote action enables host and production access on a non-preferred Metro Volume and is only available in states **Paused** and **Fractured**. A promote action can result in data corruption if the remote system is online and serving I/O. If only the connection between the current PowerStore system and the remote PowerStore system is down, verify that the remote system is no longer online.

After the promote action, production access will be enabled for hosts and applications on the current system. The local protection policy becomes enabled, and snapshot creation resumes.

Demote

A demote action disables host and production access on a preferred Metro Volume in the paused state. To prevent data corruption, the demote action is restricted when there is network connectivity between the two PowerStore systems and the remote Metro Volume also has production access. Once the demote operation completes, the non-preferred metro volume can be promoted.

After the demote action, production access will be disabled for hosts and applications. The local protection policy will be disabled, and scheduled snapshot creation will be halted.

End metro

This operation ends a Metro Volume replication session for the volume. The wizard to end this session provides two options:

- **End metro and keep the volume on both systems.**
  - On **current,** the volume properties and host access will remain unchanged.
  - On **remote,** this option unmaps the Metro Volume, deletes the Metro replication session, and assigns different SCSI WWN to the volume.

- **End metro and delete the volume and any associated snapshots on the remote system.**
  - On **current**, the volume properties and host access will remain unchanged.
  - On **remote**, this option unmaps the Metro Volume, unconfigures the Metro replication session, and deletes the volume.

Operation overview

Table 6. Metro sync session operations overview

	Status	Reconfig allowed	Modify role	Promote	Demote	Pause	Resume	End Metro
On preferred	Operating Normally		Yes			Yes		Yes
	Paused	Yes			Yes		Yes	Yes
	Fractured	Yes			Yes	Yes		Yes
	Synchronizing					Yes		Yes
On non-preferred	Operating Normally		Yes			Yes		Yes
	Paused	Yes		Yes			Yes	Yes
	Fractured	Yes		Yes		Yes		Yes
	Synchronizing					Yes		Yes

Create and manage a Metro Volume

This section shows how to create a Metro Volume configuration in PowerStore Manager. To create and manage a Metro Volume in the PowerStore REST API or PowerStore CLI, see the API and CLI guides at PowerStore Product Documentation & Videos.

The examples use the following:

- Two PowerStore clusters PowerStore-A, and PowerStore-B with remote system configuration for replication
- A single vCenter appliance vcsa.lab
- ESXi hosts esx-a, and esx-b connected to local PowerStore only for Non-Uniform host access, mapped to Metro Volume **Sales**
- ESXi hosts esx-c, and esx-d connected to local and remote PowerStore for Uniform host access, mapped to Metro Volume **Engineering**
- Heartbeat volumes are mapped from each PowerStore to all ESXi Servers to satisfy vSphere HA

- A standard Volume **Technical Marketing Engineering** is mapped to hosts esx-c, and esx-d and prepared for Metro Volume configuration as VMFS datastore in vSphere vCenter

Figure 17 shows the host connectivity options in PowerStore Manager, and Figure 18 represents the host configuration with different host connectivity settings for PowerStore-A. For uniform host connectivity, host esx-c is local to PowerStore-A. The host connectivity for esx-c, and esx-d is swapped on PowerStore-B

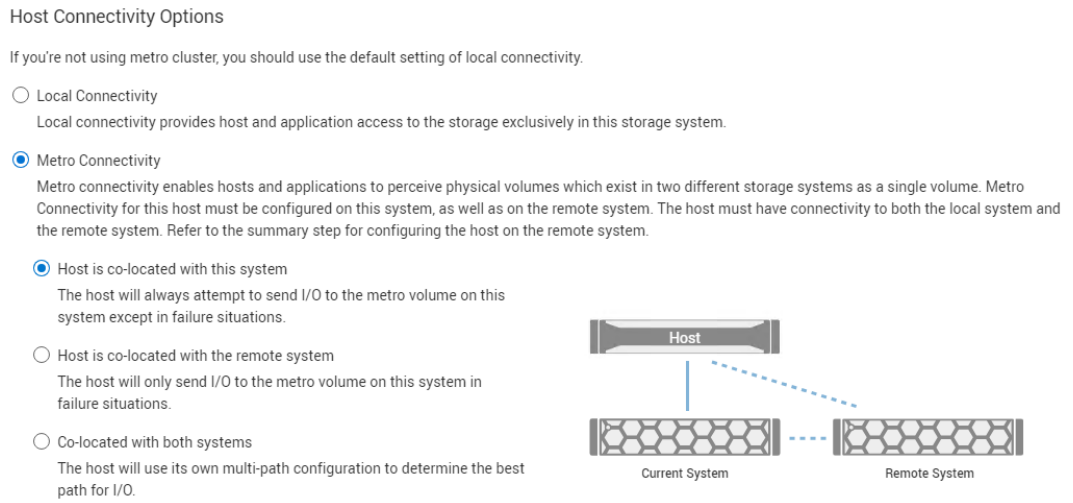


Figure 17. Host connectivity setting

Name	Host/Host Group	OS	Initiator Type	Initiators	Volume Mappings	Host Connectivity
esx-a.lab	Host	ESXi	iSCSI	1	2	Local Only
esx-b.lab	Host	ESXi	iSCSI	1	1	Local Only
esx-c.lab	Host	ESXi	iSCSI	1	2	Host is co-located with this system
esx-d.lab	Host	ESXi	iSCSI	1	2	Host is co-located with the remote system

Figure 18. PowerStore-A > Metro host overview

Create a Metro Volume replication

In the PowerStore Manager UI, you can create a Metro Volume in the Storage > Volumes overview page after selecting the appropriate Volume in drop-down menu **Protect**. Alternately, you can create it in the Volume details view under the **Protection** > **Metro** volume subtab with a link to configure Metro Volume.

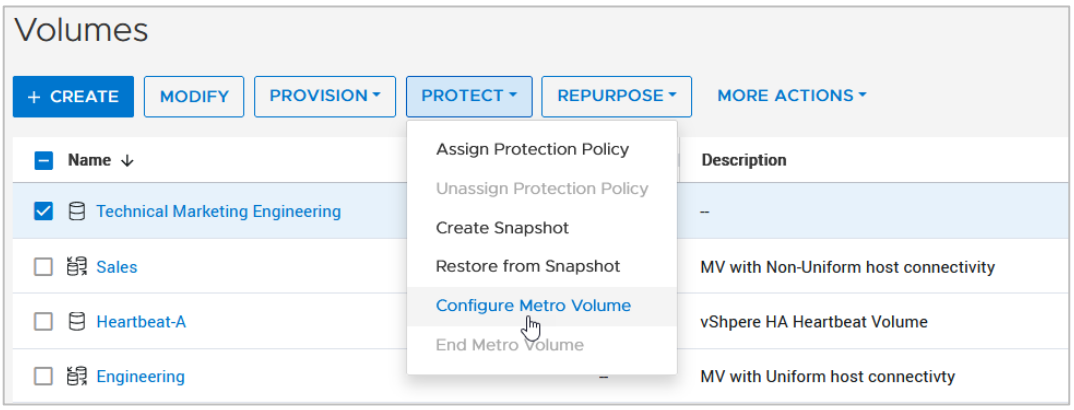


Figure 19. Configure Metro Volume in Volumes overview

The wizard prompts for the Remote System. When no remote system is set up, a link leads to the new remote systems wizard.

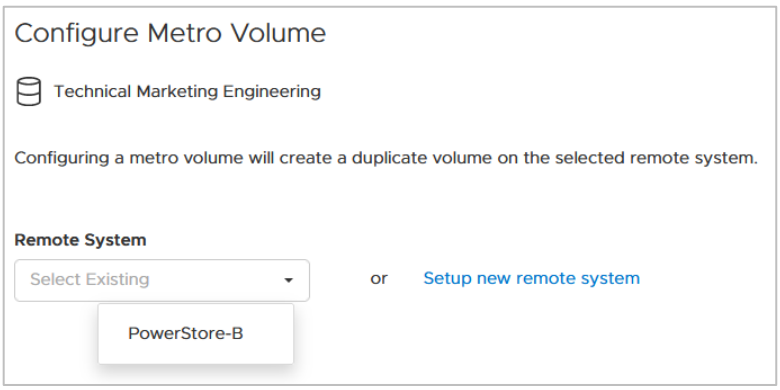


Figure 20. Configure Metro Volume wizard, single-appliance remote system

When the Remote System is a multi-appliance cluster, an extra drop-down menu allows you to select the appliance. With the **Auto** option, PowerStore Manager selects the best applicable appliance on remote PowerStore cluster for the volume.

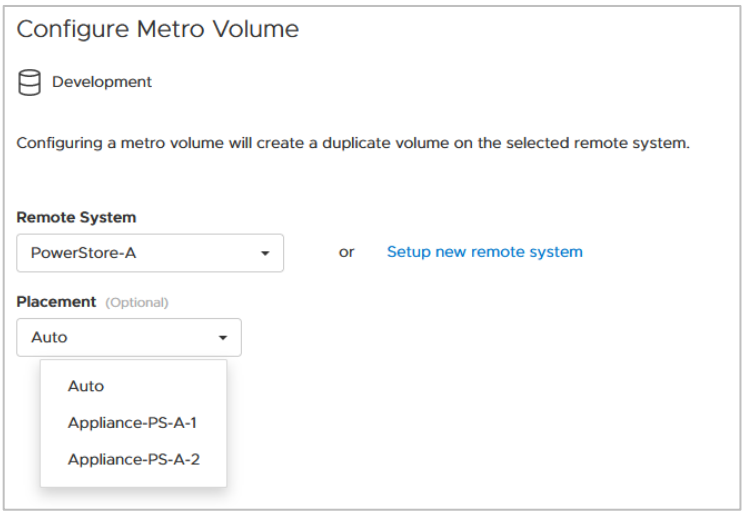


Figure 21. Configure Metro Volume wizard, multi-appliance remote system

Once a Metro Volume is set up, the icon in volume overview changes to indicate it is a Metro Volume. Also, the Metro Volume appears in Metro overview screen **Protection > Metro**. Depending on the current state, operations for the selected Volume are enabled.

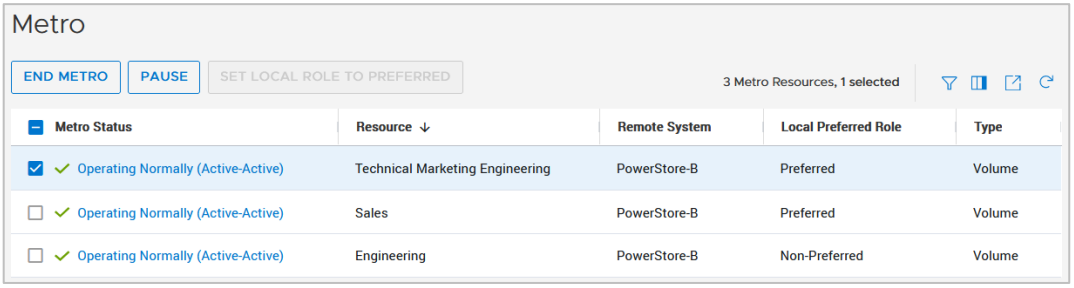


Figure 22. Protection > Metro overview

To view more details about the Metro Volume replication session, click the Metro Status for an individual Metro Volume:

Figure 23. Metro Volume in Active-Active status

Pause and Resume

A Metro Volume pause operation pauses the replication traffic to the peer, and this operation is possible in almost all situations. A dialog box details the status after the Metro Volume is in paused state. While the Metro Volume is in paused state, the non-preferred volume appears offline to mapped hosts and no replication traffic is allowed.

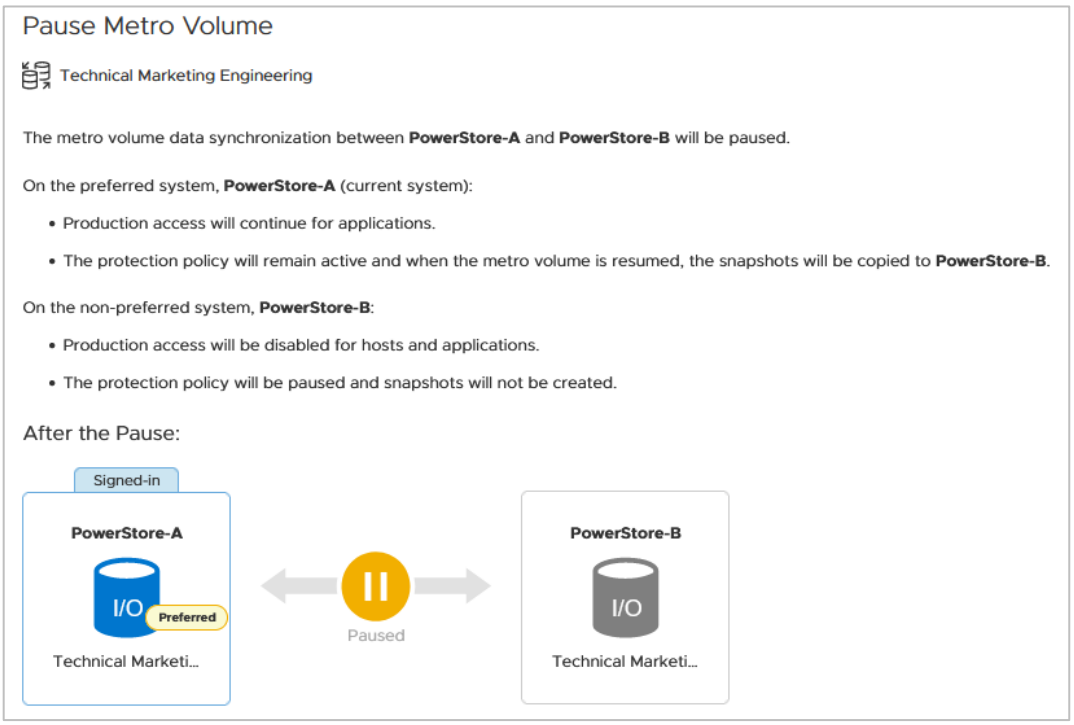


Figure 24. Pause Metro Volume

When the Metro Volume is in the paused state, the Metro Volume session details show the status and allows the demote and resume operations.

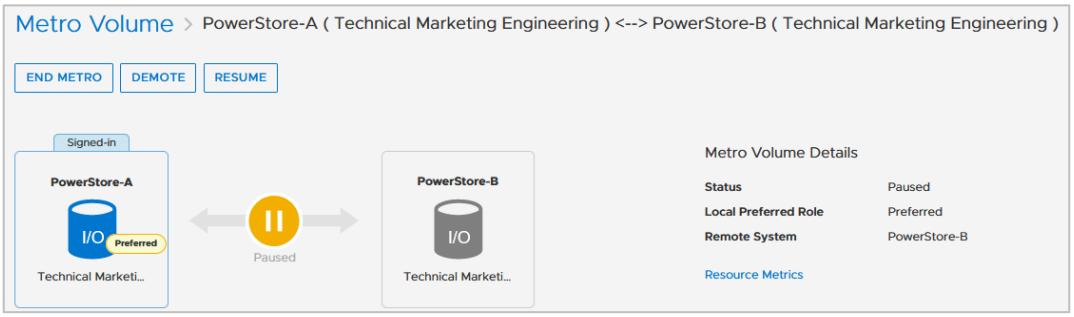


Figure 25. Metro Volume in Paused state

After you perform the **Resume** operation, the Metro Volume starts synchronization of the Metro Volume and enables host access after switching to active-active.

Modify preferred role

After a new Metro Volume configuration, the preferred side for the volume is the same as the original volume. In this example, the preferred side is PowerStore-A. The Metro Volume details page highlights the current system where the browser session is logged in and indicates the preferred side of the Metro Volume. The operation **Modify Preferred Role** is only possible in active-active state.

Figure 26. Metro Volume detail view

After you select **Modify Preferred Role**, a dialog box shows the new target configuration after the modification is finished. Modifying the preferred role does not have any influence on the host path ALUA setting.

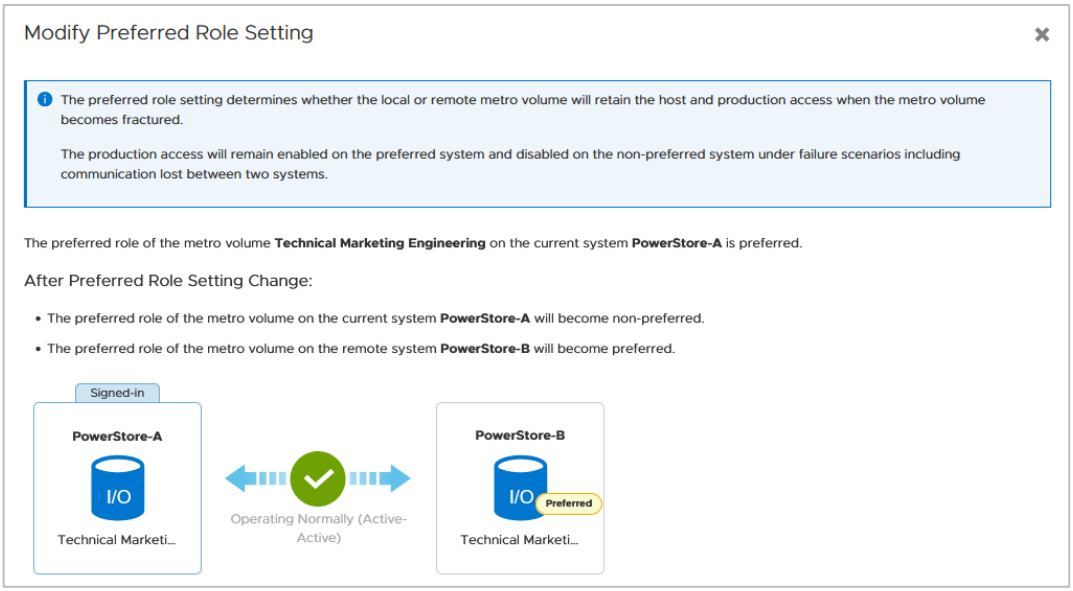


Figure 27. Modify Preferred Role Setting

Metro Volume in vSphere vCenter

When the Metro Volume sync session is set up, the volume is available on the peer PowerStore to be mapped to the hosts. After mapping and rescanning storage in vCenter, the paths appear in hosts under Storage Devices view. The shown path depends on the chosen host connectivity setup. The following figure shows a Metro Volume in a uniform connectivity after the initial sync when the volume reached active-active state. In this example, PowerStore-A is set up as co-located to the ESXi host, and PowerStore-B is the remote array. Active (I/O) indicates the active working path.

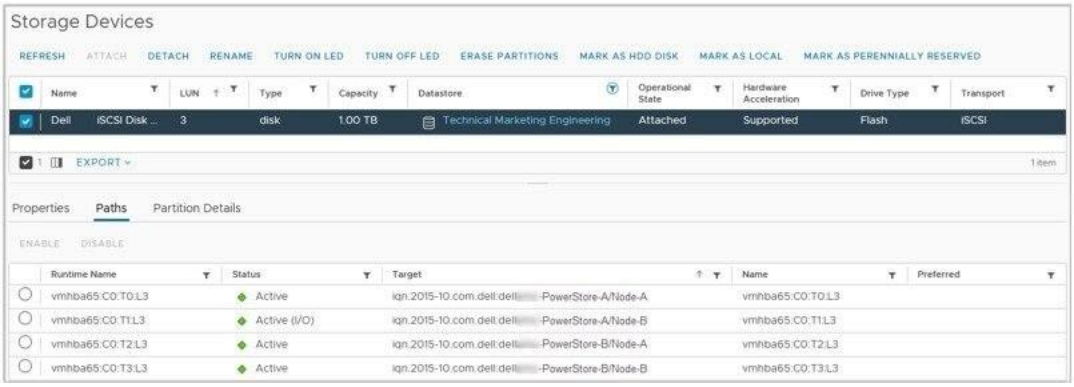


Figure 28. Uniform Metro Volume in vCenter

If you are pausing the Metro sync session or if there is a failure scenario that affects the replication, the path to the non-preferred side of the Metro Volume will show the status of dead. After a previous Modify Preferred Role operation, the Metro Volume would be offline on PowerStore-A and shows the dead paths while the active-optimized path switched over to PowerStore-B/Node-B with volume node affinity. The status Active (I/O) in vCenter shows the changed working path which is the new ALUA active optimized path presented by PowerStore-B.

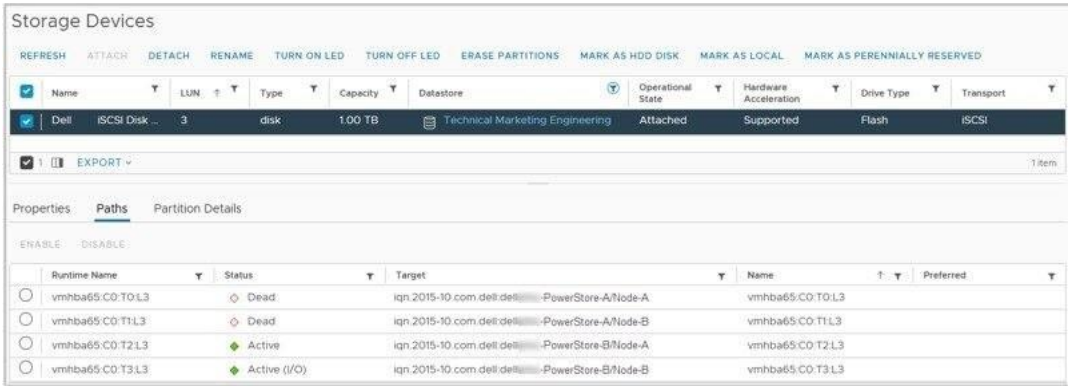


Figure 29. Uniform Metro Volume in vCenter after pause or during failure scenario

In any case, after Metro Volume resumes from a paused state or a failure scenario is solved, PowerStore starts the self-healing process of Metro Volume. After the Metro Volume is switched back into active-active, ESXi hosts can reestablish the paths.

Promote or demote after a failure situation

In a non-uniform host configuration when using the non-preferred volume, affected hosts lose all active paths to the Metro Volume. In this case, vSphere HA or vSphere FT must switch production to a host with paths to the preferred side of the Metro Volume. Path information for a Uniform connected Volume is identical for all ESXi hosts. In the example, host esx-a has a non-uniform Metro Volume configured and shows only the path to the local PowerStore cluster **PowerStore-A**.

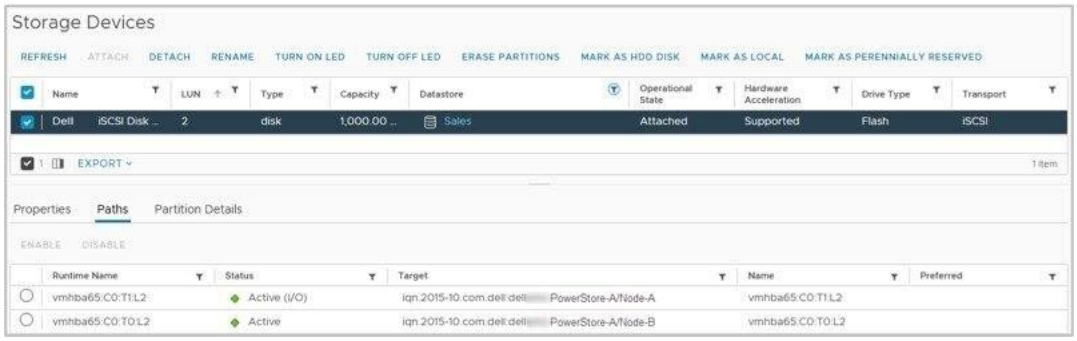


Figure 30. Non-uniform Metro Volume in vCenter

For instance, during a power outage of the site with the preferred volume, the hosts with connectivity to non-preferred volume will lose all paths to the volume. To enable production on the surviving array, it is possible to promote the non-preferred volume to enable host access to resume operations. The following figure shows the non-preferred Volume on PowerStore-B after a failure.

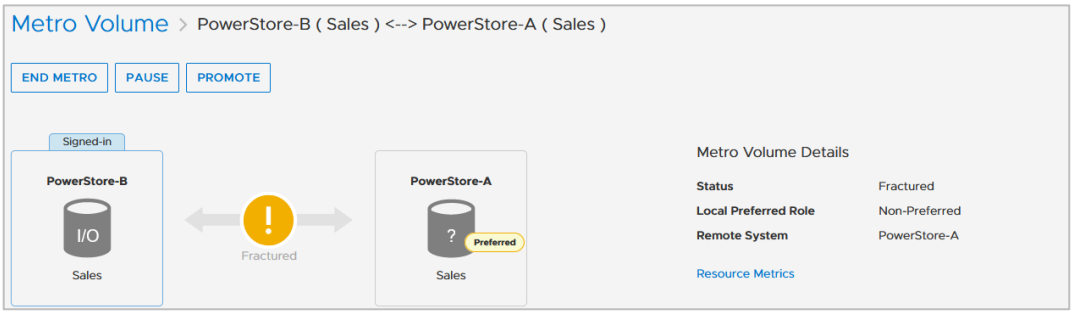


Figure 31. Fractured state on Non-Preferred

With the promote operation**,** it is possible to manually enable production host I/O for this volume even it is non-preferred. To mitigate a possible split-brain situation when both sides are online for the hosts, a promote action is only possible when the peer array is offline, or the volume is demoted on the preferred side. A dialog box shows the situation after **Promote** is performed (see below).

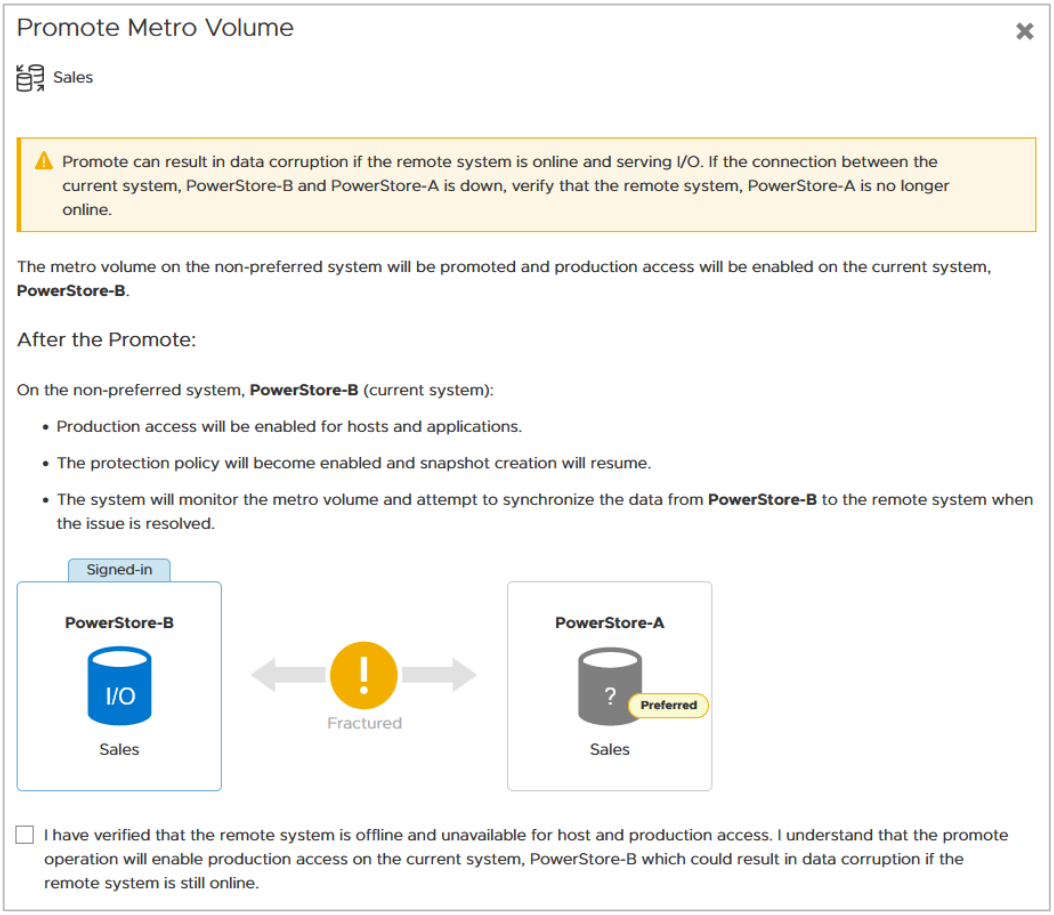


Figure 32. Promote Metro Volume

When the promote operation finishes, the paths to the non-preferred volume on ESXi hosts change to active for production I/O. Once the problem causing the fracture is solved, PowerStore starts the self-healing process and synchronizes the I/O from the promoted side to the peer.

A demote operation allows disabling the production host I/O on preferred volume. This action might be required when a promote is planned, but the PowerStore with the preferred volume is still available. After you click the demote operation, the dialog box shows the Metro Volume after the operation is finished (see below).

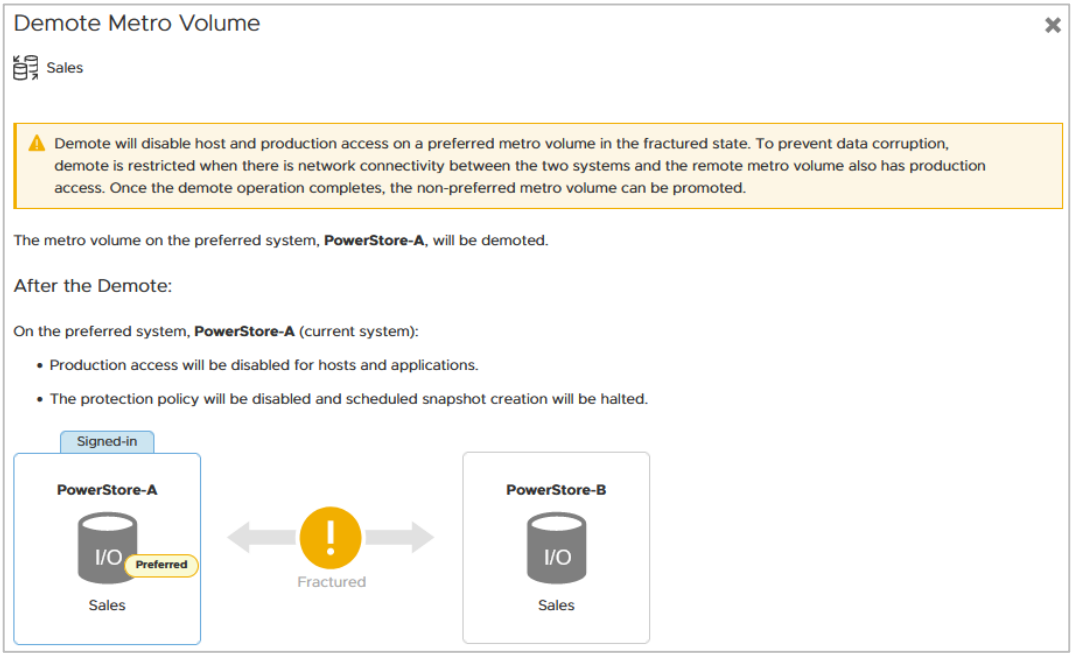


Figure 33. Demote Metro Volume

Failure scenarios

This Section covers possible failure scenarios with non-uniform and uniform host connectivity. Due to the polarization mechanism, a Metro Volume is only active-active across both PowerStore clusters when the PowerStore clusters are operational, and the links for replication management traffic and replication data traffic are established. During a link failure required for replication, the non-preferred side loses communication with preferred side of the Metro Volume and switches to offline for the hosts while preferred volumes can still be served from the same PowerStore cluster. During an array failure, only preferred volumes remain accessible without a manual promote of non-preferred volumes on surviving array. The following subsections show failure scenario in non-uniform and uniform host configuration.

Failure scenario in a non-uniform configuration

In this scenario, VMs on host 2 access two different Metro Volumes on PowerStore-B. VM1 is on Metro Volume 1 with the non-preferred role, and VM 2 is on Metro Volume 2 with the preferred role (see below).

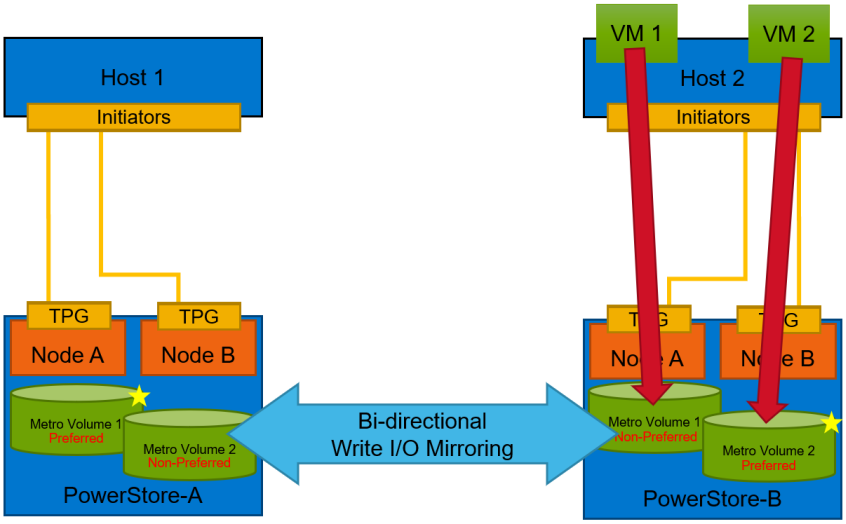


Figure 34. Non-Uniform active-active

In the following figure, due to a link failure, Host 2 loses access to Metro Volume 1 with an all path down (APD) event, and VM 1 is interrupted. VM 2 continues to run on the preferred Metro Volume 2 mapped to host 2. Host 1 has access to the preferred volume of Metro Volume 1, vSphere HA recognizes that the Volume is still available, and it restarts VM 1 on host 1.

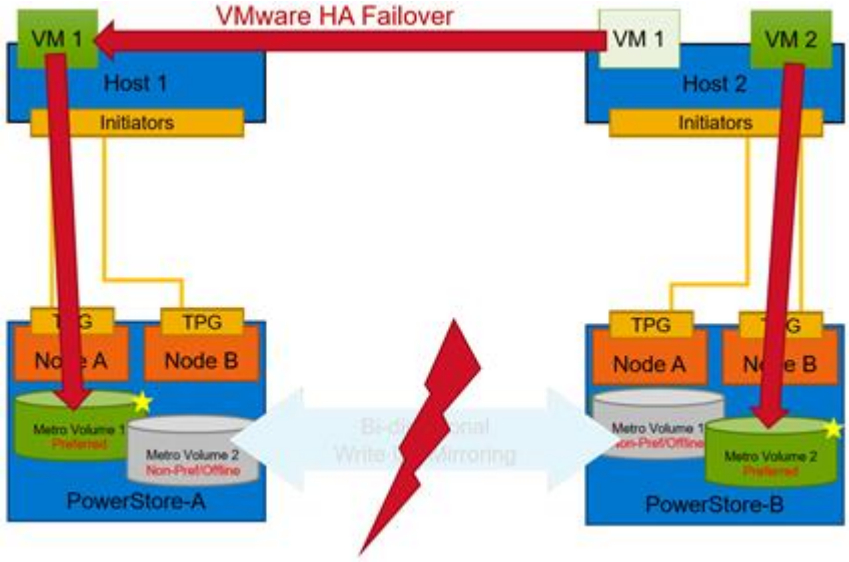


Figure 35. Non-uniform during a link failure

With the same configuration when PowerStore-B is down during a failure situation (array failure), Metro Volume 2 is not accessible until a manual promote operation completes on PowerStore-A. After the promote action of Metro Volume 2 on host 1, the volume and active paths are enabled, and VM 2 could start on host 1.

Failure scenario in a uniform configuration

The example configuration for a uniform configuration is identical to the previous example with extra paths from the hosts to the remote PowerStore cluster. The redundant active paths to local and remote PowerStore clusters give an extra level of high availability.

Depending on the host configuration as described in the Host configuration section, the paths to the volume served by remote PowerStore can be ALUA active-non-optimized only, or also active-optimized with active-non-optimized paths dependent on volume node affinity setting. In example as shown below, VM 1, and VM 2 are running on host 2 using different Metro Volumes. While VM 1 uses Metro Volume 1 in the non-preferred role which goes offline during a failure, VM 2 uses Metro Volume 2 in the preferred role.

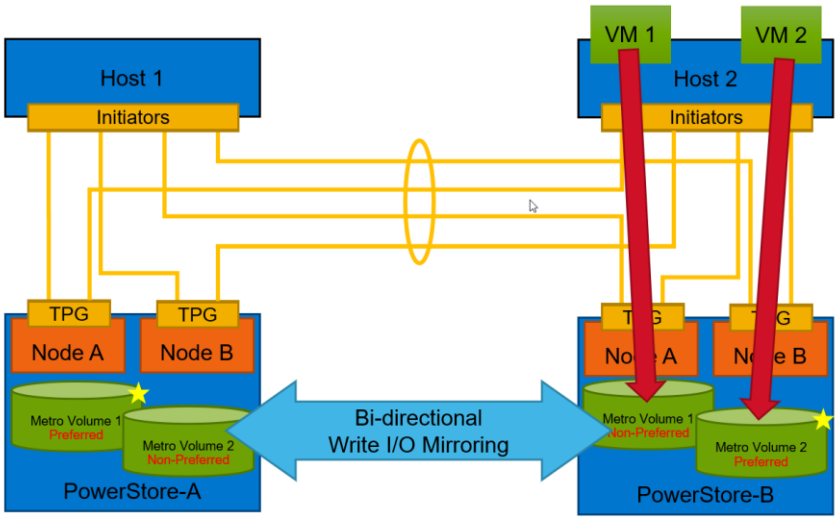


Figure 36. Uniform active-active

In a failure scenario, when the link between PowerStore-A and PowerStore-B is affected, host 2 would lose the volume for VM 1. As active paths are still available to the same volume on PowerStore-A, and VM 1 remains online without interruption. VMware Native Multipathing (NMP) handles the failover to an available active path.

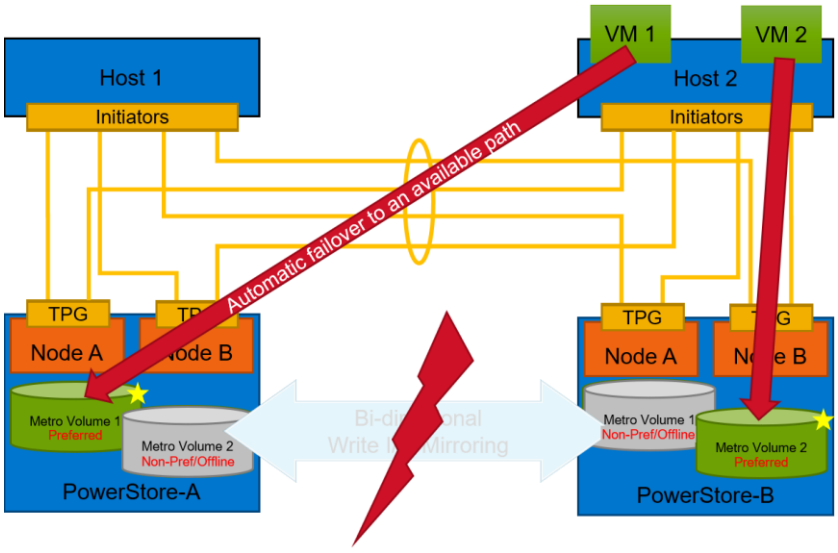


Figure 37. Uniform during a link failure

During an array failure of PowerStore-B, VM 2 would also lose the active path to Metro Volume 2, but it cannot continue on PowerStore-A because Metro Volume 2 is in the non-preferred role. VM 2 can only start with Metro Volume 2 on PowerStore A after a manual Metro Volume promote action performs on PowerStore-A.

Failure scenario table

The following table outlines tested design and component failure scenarios with Metro Volume enabled with vSphere HA.

Table 7. Failure scenarios with uniform and non-uniform storage presentation (no witness)

Event scenario	Metro Volume behavior	vSphere HA behavior
Uniform: Outage of non-preferred Metro Volume	Preferred Metro Volume remains available	None. VMs have uniform paths to surviving preferred Metro Volume.
Non-uniform: Outage of non-preferred Metro Volume	Preferred Metro Volume remains available	vSphere HA restarts impacted VMs on preferred Metro Volume.
Uniform: Outage of synchronous replication link	Preferred Metro Volume remains available	None. VMs have uniform paths to surviving preferred Metro Volume.
Non-uniform: Outage of synchronous replication link	Preferred Metro Volume remains available	vSphere HA restarts impacted VMs on preferred Metro Volume.
Uniform: Outage of preferred Metro Volume	Non-preferred Metro Volume becomes unavailable. Manually promote non-preferred Metro Volume.	vSphere HA attempts to restart impacted VMs if a Metro Volume becomes available in time.
Non-uniform: Outage of preferred Metro Volume	Non-preferred Metro Volume becomes unavailable. Manually promote non-preferred Metro Volume.	vSphere HA attempts to restart impacted VMs if a Metro Volume becomes available in time.

vSphere DRS, HA, and Metro Volume

VMware Distributed Resource Scheduler (DRS) is a cluster-centric configuration that uses VMware vSphere vMotion to automatically move virtual-machine compute resources to other hosts in a vSphere cluster. This action is performed without local, metro, or stretched-site awareness. Also, vSphere is unaware of storage virtualization that occurs in Metro Volume. When Metro Volume is configured with vSphere, consider placing tier 1 or business-critical virtual machines on the preferred Metro Volume if there could be an unplanned synchronous replication link outage.

If DRS is enabled on a vSphere cluster, DRS could automatically move virtual machines around. This situation could be a problem in a non-uniform storage presentation design. DRS host groups and VM groups can be used to apply rules resulting in boundaries for virtual machine mobility.

- VM groups: Virtual machines can be placed into VM groups. VM groups can be pinned to hosts in a Host group where the VMs should run.

  Host groups: Hosts that share a common preferred or non-preferred Metro Volume can be placed into Host groups. Once the host groups are configured, they can represent locality for the preferred or non-preferred Metro Volume.

You can assign VM groups to host groups using the vSphere Client. This practice ensures virtual machines follow predictable placement rules on vSphere hosts backed by Metro Volumes. You can also design the infrastructure with separate DRS-enabled clusters existing at both sites, keeping automatic migration of virtual machines within the respective site where the preferred or non-preferred Metro Volume resides.

vSphere Availability (HA) is also a cluster-centric configuration. If there is a host or storage failure, HA attempts to restart virtual machines on a host candidate within the same cluster that still has access to the Metro Volume. In a non-uniform storage presentation, if an outage impacts non-preferred Metro Volume availability, you can configure vSphere HA to restart impacted virtual machines on the preferred Metro Volume.

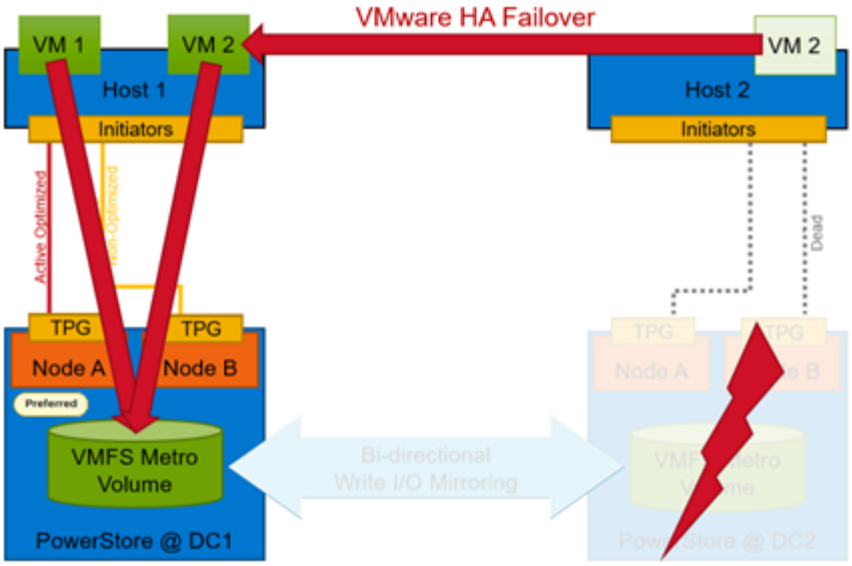


Figure 38. Failure of non-preferred role in non-uniform storage presentation

The following vSphere advanced tuning should be configured for non-uniform stretched cluster configurations. This tuning allows HA to power off and migrate virtual machines to the preferred Metro Volume if the non-preferred Metro Volume becomes unavailable.

Configure the following:

- vSphere 6.0+: Disk.AutoremoveOnPDL = 1 (VMware recommended default advanced setting on each host)
- Besides supporting HA restart for Permanent Device Loss (PDL) events, HA restart can also react to All-Paths-Down (APD) events. We recommend configuring VM Component Protection for PDL and APD events. For PDL events, select **Power off and restart VMs**. For APD events, VMware recommends selecting **Power off and restart VMs** (conservative). For the advanced APD settings, see the document *VMware vSphere Metro Storage Cluster Recommended Practices*.

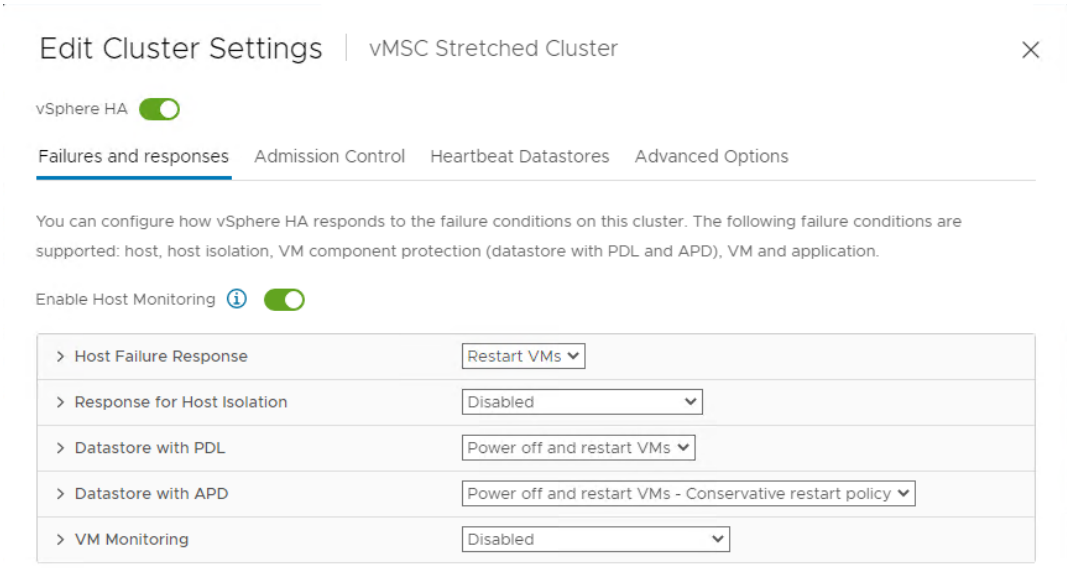


Figure 39. Configuring vSphere HA for PDL and APD conditions

The Disk.AutoremoveOnPDL advanced setting is not configurable in VMCP and should remain at its default value of 1 for each vSphere 6 host in the cluster. For more information about the Disk.AutoremoveOnPDL feature, see the VMware KB article 2059622, PDL AutoRemove feature vSphere 6.x/7.x.

If an unplanned outage impacts the preferred Metro Volume availability, the Metro Volume becomes unavailable. Virtual machines running on this Metro Volume will lose access to storage.

VMware vSphere monitors SCSI sense codes sent by an array to determine if a device is in a PDL state. These SCSI sense codes are outlined in VMware KB article 2004684, Permanent Device Loss (PDL) and All-Paths-Down (APD) in vSphere 6.x and 7.x. PowerStore supports the following SCSI sense codes (see table below) which will be sent to vSphere hosts when a PDL condition is met.

Table 8. SCSI sense codes

SCSI sense code	Description
0x02/0x04/0x0c	ALUA UNAVAILABLE

vSphere vMotion and Metro Volume

For stretched clusters or data centers, consider the vMotion and Metro vMotion latency requirements between hosts. vMotion requires a round-trip latency of 5 ms or less between hosts on the vMotion network. vSphere Metro Storage Cluster boosts the allowable round-trip latency 5 ms–10 ms on the vMotion network between sites with Enterprise Plus licensing.

Note: Metro Volume requires 5ms or less latency.

When host connectivity is properly configured along with the appropriate path selection policy, migrating virtual machines between PowerStore clusters in a Metro Volume configuration should work as designed. However, there are a few things to keep in mind:

- Stretched Layer 2 networking is required between sites in a stretched cluster architecture. Virtual machines that are migrated from one site to another will expect to be on the same Layer 2 network after the vMotion migration is complete.
- For higher availability, virtual machines should run on a vSphere host which is local to the preferred Metro Volume. The reason for this practice is because Metro Volume failures are handled using polarization. If there is a synchronous replication link outage, the preferred volume remains available to hosts while the non-preferred volume becomes unavailable to hosts.

Supported configurations

Metro Volume connectivity requirements may vary depending on intended and actual use. For example, requirements for using Metro Volume to migrate a workload of powered off virtual machines is going to be significantly different than migrating virtual machines that are powered on and highly active.

A native PowerStore Metro Volume can be configured on all PowerStore models running PowerStoreOS 3.0 or later in a vSphere Metro Storage Cluster (vMSC) configuration. A vMSC can provide a vSphere stretched cluster architecture. For the host configuration PowerStore supports non-uniform host access which uses only local path, and Uniform host access which provides extra cross connectivity to the remote PowerStore cluster.

We recommend using dedicated VLANs or fabrics to isolate IP-based storage traffic from other types of general-purpose LAN traffic, especially when spanning data centers. While this recommendation is not a requirement for Metro Volume, it is a general best practice for IP-based storage.

Besides the existing requirements for PowerStore replication, the following restrictions and requirements apply for Metro Volume:

- PowerStore Q/T model cluster only
- VMware ESXi in a vSphere Metro Storage Cluster configuration
- Scalability as of PowerStoreOS 4.0:
  - 500T model: 128 Metro Volumes per appliance
  - 1000T model: 128 Metro Volumes per appliance
  - 1200T/3000T model: 256 Metro Volumes per appliance
  - 3200T/3200Q/5000T model: 256 Metro Volumes per appliance
  - 5200T/7000T model: 512 Metro Volumes per appliance
  - 9200T/9000T model: 512 Metro Volumes per appliance
- Support for FC/SCSI or iSCSI VMFS volumes and physical or virtual Raw Device Mappings (RDMs)
- Support for FC VMware shared/clustered VMDKs
- PowerStore TCP-based replication Ethernet link

- Maximum distance between sites: 100 km (60 miles)
- Minimum bandwidth: 250Mbps per concurrent vMotion + replication traffic
- Maximum RTT latency: 5 milliseconds
- Volume migration feature within a PowerStore cluster is not allowed with Metro Volumes

Note: The PowerStore TCP-based replication protocol uses TCP ports in the range of 13333-13337 depending on the network latency configuration of the remote system. Ensure firewall rules between PowerStore clusters with Metro Volumes allow these ports to pass traffic.

Note: Support for SCSI-3 persistent reservations, vSphere Raw Device Mappings (RDMs), shared/clustered VMFS VMDK virtual disks, and PowerStore volume groups with Metro Volume were added in the PowerStoreOS 4.0 release.

System limits

For the most up-to-date system limits, see the Simple Support Matrix on Dell.com/powerstoredocs.

Witness

Introduction

The release of PowerStoreOS 3.6 introduces a witness component to the Metro Volume architecture. The functional design of the witness adds more resiliency to new and existing Metro Volume deployments and further mitigates risk of split-brain situations. This is accomplished by more intelligent decision making across a wider variety of infrastructure outage scenarios, including unplanned outages. In addition, the witness component satisfies requirements and recommendations found in vSphere Metro Storage Cluster (vMSC) designs.

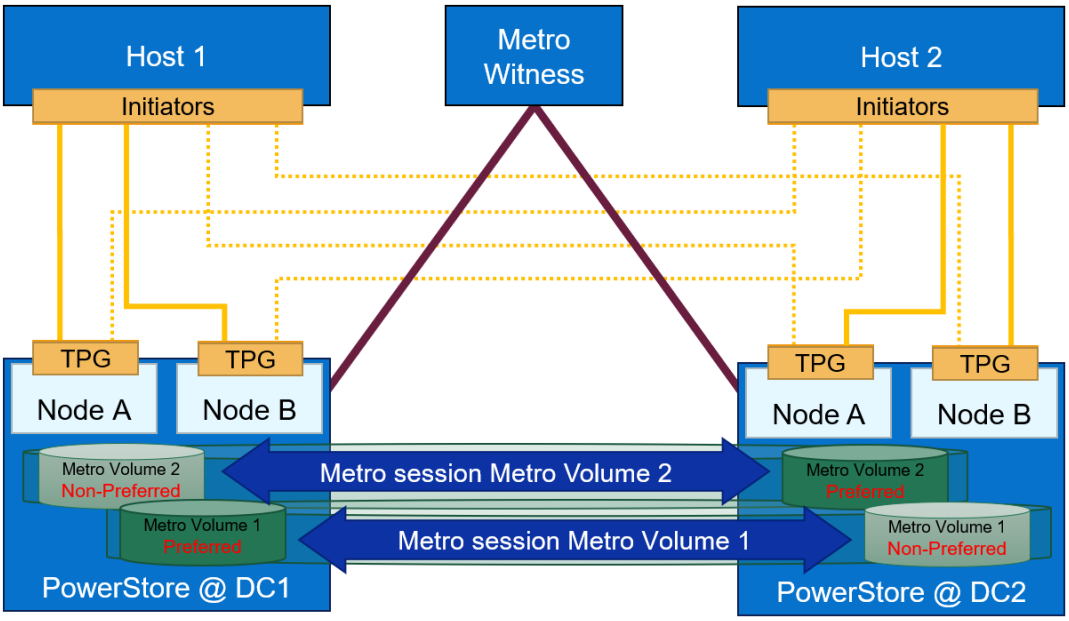


Figure 40. Metro Volume witness adds more resiliency to new and existing Metro Volume deployments

Requirements and installation

The witness is distributed as an installable RPM package which can be deployed on a virtual or physical system running a supported version of Linux. See the Dell PowerStore Simple Support Matrix on the Dell E-Lab Navigator portal for a list of supported Linux hosts. The package may be installed using RPM-Tool if required dependencies have already been met (mainly Java and SQLite), or a package management utility (such as yum or zypper) which will resolve all required dependencies.

```
[root@psmvwitl tmp]#
[root@psmvwitl tmp]# firewall-cmd --permanent --zone=public --add-service=https
success
[root@psmvwitl tmp]# firewall-cmd --reload
success
[root@psmvwitl tmp]# firewall-cmd --permanent --zone=public --list-services
cockpit dhcpv6-client https ssh
[root@psmvwitl tmp]#
```

```
[root@psmvwitl tmp]#
[root@psmvwitl tmp]# systemctl stop firewalld
[root@psmvwitl tmp]# systemctl disable firewalld
Removed /etc/systemd/system/multi-user.target.wants/firewalld.service.
Removed /etc/systemd/system/dbus-org.fedoraproject.FirewallD1.service.
[root@psmvwitl tmp]#
```

Figure 41. Dell recommends adding an https allow rule to allow https traffic to the witness. Alternatively, the firewall may be stopped and disabled but this is not recommended.

```
[root@psmvwitl tmp]#
[root@psmvwitl tmp]# rpm -i dell-witness-service.noarch.rpm
[root@psmvwitl tmp]#
```

Figure 42. Installing the witness using RPM-Tool

Note: TCP ports 22 and 443 must be open for remote SSH login and https witness communication respectively. In addition, the witness time must be synchronized with the time on PowerStore. Time synchronization can be automated using NTP or PTP configuration for the witness operating system and NTP configuration in PowerStore manager.

Hostnames and certificates

The SSL certificate for the witness service is setup during installation of the RPM and is using the configured hostname, the primary Ipv4 and Ipv6 (if configured) address as valid names for the generated certificate. For the hostname it's recommended to use a FQDN (Full Qualified Domain Name) to get the FQDN as valid name into the certificate. For registration and use of a Metro Volume witness service in PowerStore, it's important to get the correct name or IP address into the certificate subject alternate name (SAN). After installing metro witness service, the openssl tool could be used to verify the SAN of the used certificate.

```
# echo "" | openssl s_client -connect metro-witness.lab:443 2>/dev/null |
openssl x509 -noout -subject -ext subjectAltName

subject=CN = www.dell.com, OU = PowerStore, O = Dell, L = Hopkinton, ST =
Massachusetts, C = US

X509v3 Subject Alternative Name:

    IP Address:192.168.8.220, IP Address:FE80::FE81:758F, DNS:metro-
witness.lab
```

The example below shows the output for metro witness service running on host “metro-witness.lab”. The valid addresses to register the witness in PowerStore manager are the Ipv4 address 192.168.8.220, the Ipv6 address FE80::FE81:758F, and the DNS name for the server metro-witness.lab. On a shared OS installation or when multiple IP’s or network cards are used, the certificate may not contain a valid IP or DNS name to register the witness service in PowerStore manager successful. The Dell witness service RPM contains the script /opt/dell-witness-service/scripts/replace_certificate.sh which allows to change the certificate SAN entries for the witness service certificate. The following examples shows the command to create SSL certificates for the witness service containing multiple IP and DNS names which could be required when running the witness serivce in different DNS zones or in a NAT environment.

With that example, PowerStore manager can register the witness service by using one of the entries listed as Subject Alternative Name (SAN) in the output:

- 192.168.8.220
- 172.16.8.220

metro-witness.lab

- metro-witness.lab.dell.com

```
# /opt/dell-witness-service/scripts/replace_certificate.sh -i "192.168.8.220
172.16.8.220" -f "metro-witness.lab metro-witness.lab.dell.com"

# echo "" | openssl s_client -connect metro-witness.lab:443 2>/dev/null |
openssl x509 -noout -subject -ext subjectAltName

subject=CN = www.dell.com, OU = PowerStore, O = Dell, L = Hopkinton, ST =
Massachusetts, C = US

X509v3 Subject Alternative Name:

    IP Address:192.168.8.220, IP Address:172.16.8.220, DNS:metro-witness.lab,
DNS:metro-witness.lab.dell.com
```

For additional reading, full coverage of Metro Volume witness requirements and installation is provided in the document Dell PowerStore: Protecting Your Data.

Witness registration

Each PowerStore can register only a single witness service which is valid for all metro sessions on the PowerStore cluster. When metro sessions are already configured when metro witness is registered successful on participating PowerStore clusters, all existing metro sessions and new metro sessions are protected with the witness service. It’s not possible to enable or disable witness protection for individual metro sessions. The required registration token to register the witness service in PowerStore manager is generated with a SSH session to the metro witness server by using the script generate_token.sh as shown in the example below.

```
# /opt/dell-witness-service/scripts/generate_token.sh

The generated token is: LyFTkq9r
```

The generated token is valid for 10 minutes to register the witness service for one or multiple PowerStore clusters.

The following figure provides an example of registering the Witness. This is found in PowerStore Manager > Protection > Metro Witness > ADD

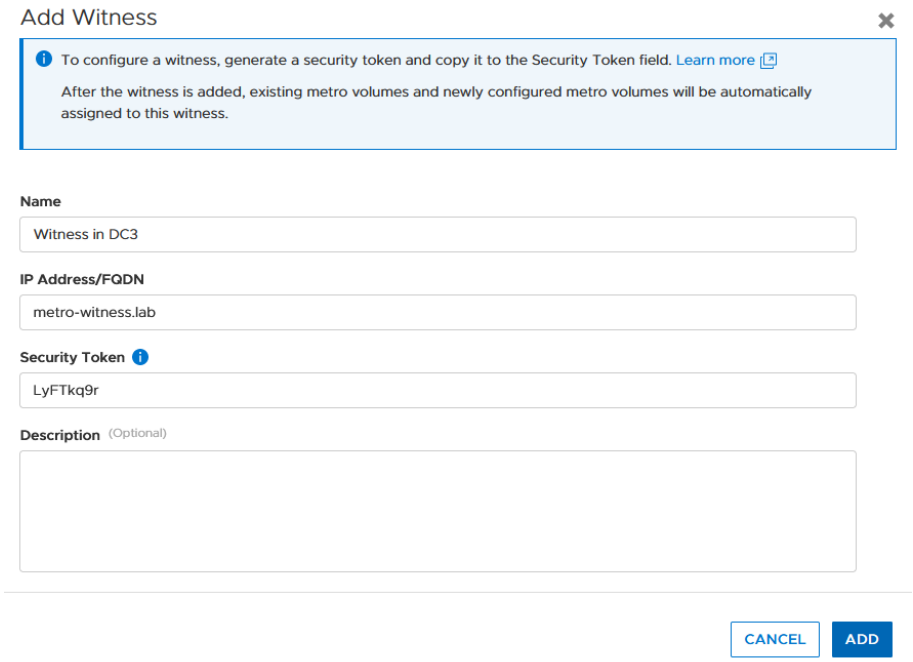


Figure 43. PowerStore Manager – Add Witness

In the following screen the SSL certificate for the witness service with the certificate thumbprint is shown. For an optional validation of the shown certificate thumbprint in

```
# /opt/dell-witness-service/scripts/thumbprint.sh

The certificate thumbprint is:
0562664c08cd16d953dcad5331d461d161cb0794bb8994cf1f4999c51b9f29d9
```

PowerStore manager can be compared with the thumbprint used by the witness service. The thumbprint can be shown in a SSH shell on the witness service server with script thumbprint.sh as shown in the example below.

Upgrading

Upgrading the witness is much like installing the witness and can be completed in a few easy steps:

1. Uninstall the old witness package (rpm -e dell-witness-service)
2. Install the new witness package (rpm -i dell-witness-service.noarch.rpm)
3. Unregister the witness from each PowerStore cluster
4. Generate the token
5. Re-register the witness on each PowerStore cluster

Note: When OS or Java upgrades are installed, either manually or automatically, it may require a restart of the witness service (systemctl restart dell-witness-service).

Witness components

The witness design consists primarily of two component areas: The witness service running on the Linux machine, and the witness logic residing in PowerStoreOS. These components work together to provide a robust set of predictable outcomes and application uptime in the event of an unplanned infrastructure outage.

Witness service:

- Installed outside of PowerStoreOS at a third site
  - Registered in PowerStore Manager
  - Communicates using the PowerStore cluster management network
  - No active checks and only responds to requests from a PowerStore
  - Only responds to requests from PowerStore and does not trigger a metro session operation
- Is not involved in application data path and doesn't get any host workload
  - Maintains a witness status database
- Witness logic in PowerStoreOS:
  - Exchanges the witness session information with metro witness
  - Runs periodic keep-alive check against metro witness
  - Uses a preferred/non-preferred role to prioritize surviving volume
  - Gets the winner or loser role from metro witness during a failure scenario
  - Allows to keep preferred or non-preferred volume online after a decision with metro witness

Communication flow – healthy

The communication between PowerStore and witness service uses 2048bit SSL encryption with password-less authentication using SSL client certificates. During registration of the witness, the client certificates are exchanged with a token in PowerStore manager, and the witness session is established (1). Each PowerStore appliance pings the witness regularly with an HTTP GET request to confirm connectivity and healthy availability of the witness (2). Each PowerStore appliance with a preferred Metro Volume will send an HTTP POST request once per metro session per day to renew the witness session. The metro witness session will expire after seven days if not renewed. Lastly, in a short interval, each PowerStore appliance with a non-preferred Metro Volume requests and receives a grant from its peer with the preferred Metro Volume (3). This mechanism helps assure its peer is healthy and that the non-preferred Metro Volume may still fulfill read and write I/O requests.

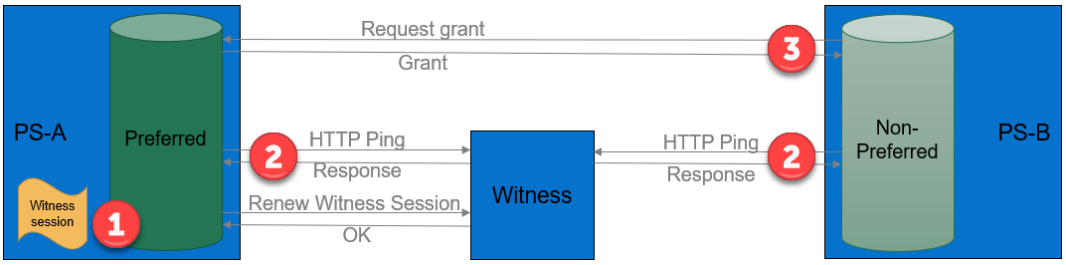


Figure 44. Communication flow – healthy

In all cases throughout its lifecycle, the witness acts like a backboard and only responds to inward communication requests from PowerStore appliances. The witness does not outwardly initiate communication with PowerStore appliances.

Metro Volume witness connection states

One Metro Volume witness may be configured per PowerStore appliance. The witness related configuration is available in PowerStore manager *Protection* > *Metro Witness*. This page allows the addition of a new witness and shows the current configured witness service.

There are five possible Metro Volume witness connection states:

- **OK** – This is the normal operating condition. All nodes on all appliances can communicate with the witness.
- **Partially Connected** – Some nodes on some appliances can communicate with the witness. The same witness may not be registered on the peer PowerStore.
- **Disconnected** – All nodes on all appliances cannot communicate with the witness.
- **Deleting** – The witness is currently being unregistered and deleted from the cluster. This state may persist if the delete operation fails. In that case, the delete can be retried.
- **Initializing** – All nodes on all appliances are initializing their connection to the witness. This is the initial state of the witness when first added to the PowerStore.

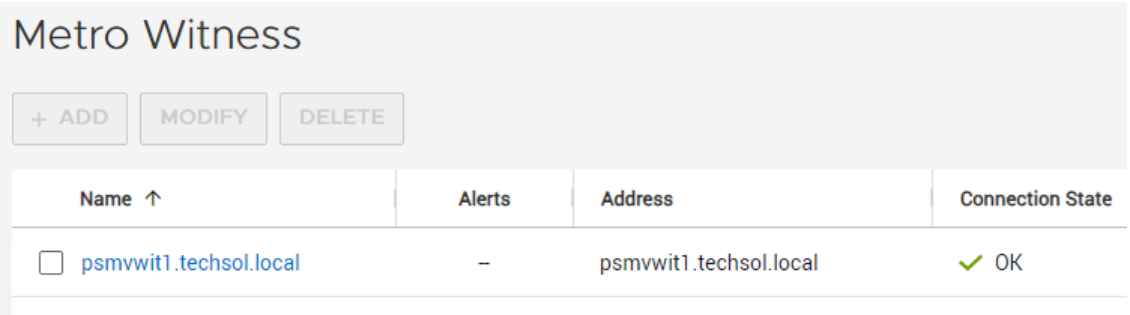


Figure 45. Metro Volume witness connection state of OK

Metro Volume session witness states

The Metro Volume session related configuration is available in PowerStore manager *Protection* > *Metro*. This page allows the management of each Metro Volume session and shows the current session witness and corresponding witness state.

There are six possible Metro Volume session witness states:

- **Initializing** – The witness is being initialized, but not engaged.
- **Disengaged** – The witness is initialized, but not engaged. It's the valid state during a fracture as witness is not involved. While in this state on both sides, Metro Volume will resort to utilizing the polarization mechanism for Metro Volume availability.
- **Engaged** – The witness is engaged. This is the normal condition, and it indicates the witness will be leveraged by the Metro Volume session in an infrastructure failure scenario.
- **Disengaged Invalid Configuration** – Not engaged due to incorrect witness configuration on the preferred system of the Metro Volume session.
- **Disengaged Failed To Initialize** – There is a problem with the Metro Volume session because it failed to initialize with the witness.
- **Unconfigure In Progress** – The witness is in the process of being unconfigured for the Metro Volume session.

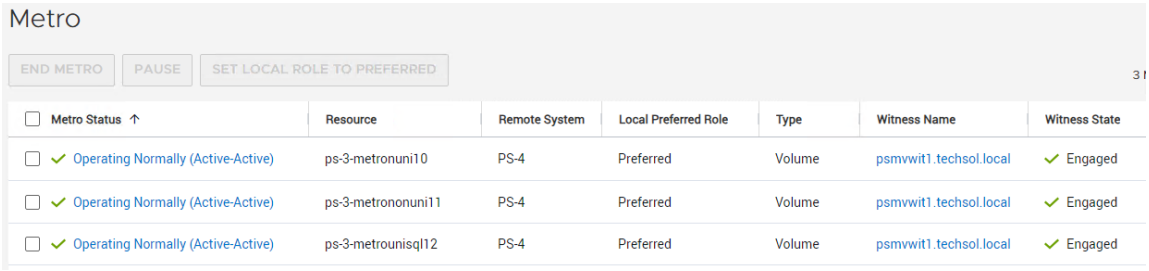


Figure 46. Metro Volume session witness states of Engaged

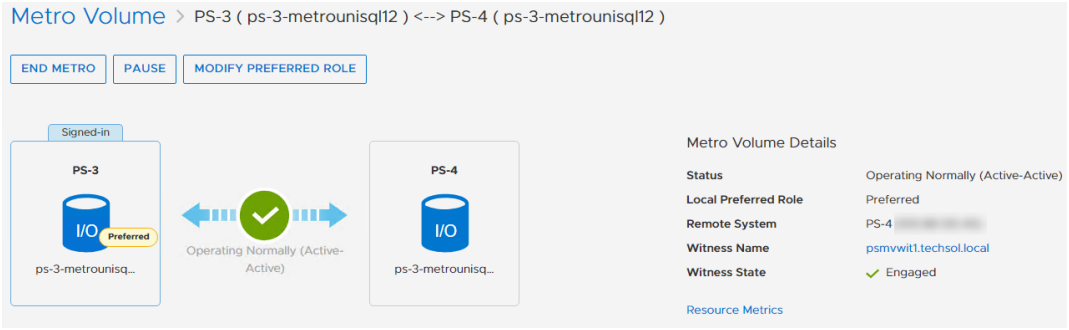


Figure 47. The session witness state can also be observed while drilling down into the Metro Volume detail

Communication flow – fracture

If PowerStore detects a communication problem with its peer system (due to missing TTL request or grant), a fracture request is sent to the witness for each Metro Volume session, with a short delay from the non-preferred side. Whichever request is received first will be declared as the winner and the losing request, if received by the witness, will receive an error. The declared winner continues to serve read and write host I/O and the loser will demote itself.

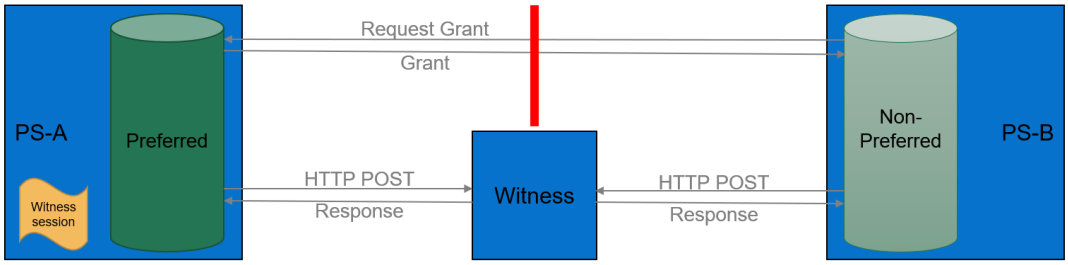


Figure 48. When PowerStore cannot communicate with its peer, a fracture will occur

Failure scenarios with witness

The witness adds resiliency to Metro Volume by handling an expanded scope of failure scenarios while mitigating risk of a split-brain situation. Some of the additional outcomes available with the witness are not possible solely with the polarization mechanism. The table below explores infrastructure failure scenarios for preferred and non-preferred and the expected outcomes when the witness is deployed. These resources should be used in the planning process when deploying Metro Volume and determining the location of the preferred and non-preferred roles as well as the workloads residing on them.

Note: If there are multiple infrastructure failures, the expected outcomes are dependent on the failures happening concurrently. If there is an ordered delay between failures, the actual outcome may not match the expected outcome.

Table 9. Infrastructure failure scenarios with witness

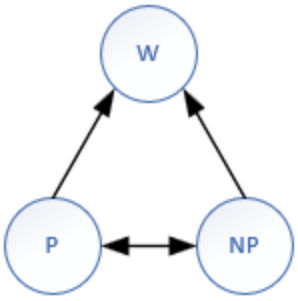
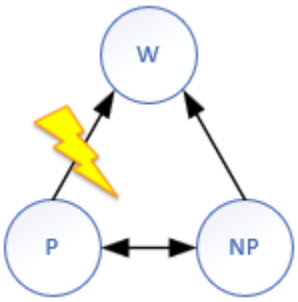
Figure	Preferred	Non-preferred	Link: Replication	Link: Preferred to witness	Link: Non-preferred to witness	Witness	Expected Outcome
	OK	OK	OK	OK	OK	OK	Volume is online on both arrays
	OK	OK	OK	X	OK	OK	

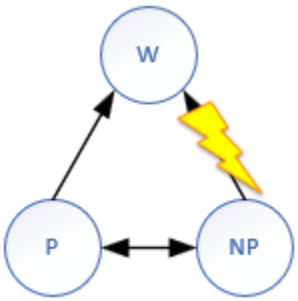
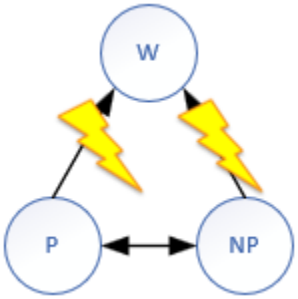
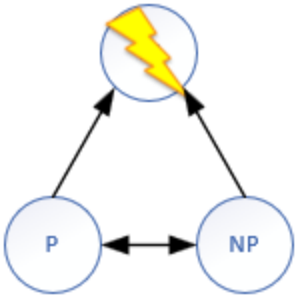
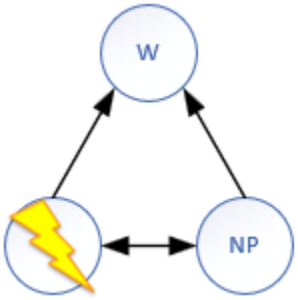
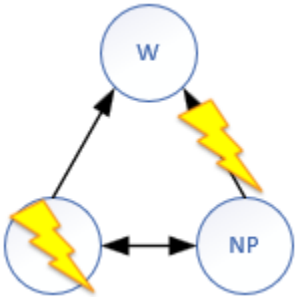
Figure	Preferred	Non-preferred	Link: Replication	Link: Preferred to witness	Link: Non-preferred to witness	Witness	Expected Outcome
	OK	OK	OK	OK	X	OK	
	OK	OK	OK	X	X	OK	
	OK	OK	OK	OK	OK	X	
	X	OK	OK	OK	OK	OK	Volume is online on the Non-preferred *
	X	OK	OK	X	OK	OK	

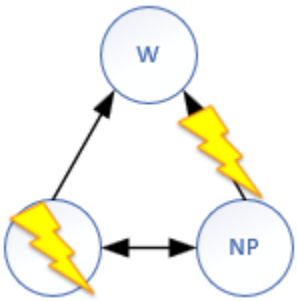
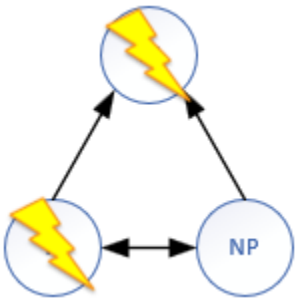
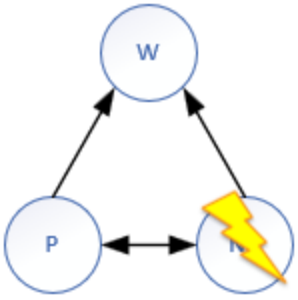
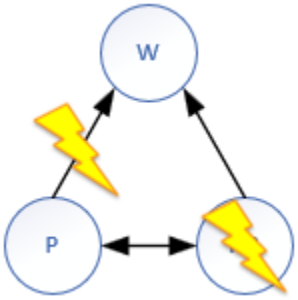
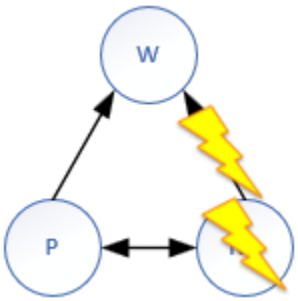
Figure	Preferred	Non-preferred	Link: Replication	Link: Preferred to witness	Link: Non-preferred to witness	Witness	Expected Outcome
	X	OK	OK	OK	X	OK	Volume is offline on both arrays
	X	OK	OK	OK	OK	X	
	OK	X	OK	OK	OK	OK	Volume is online on the Preferred
	OK	X	OK	X	OK	OK	Volume is offline on both arrays
	OK	X	OK	OK	X	OK	Volume is online on the Preferred

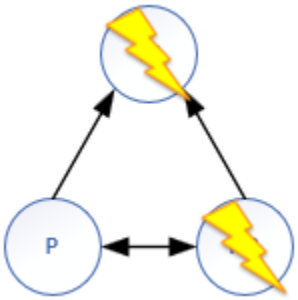
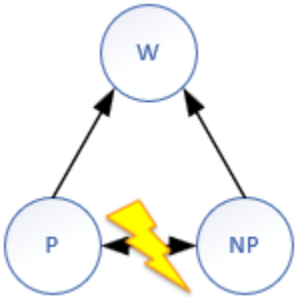
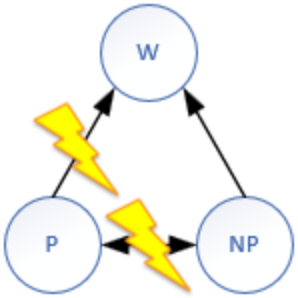
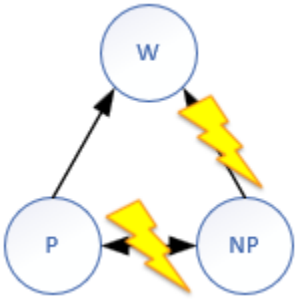
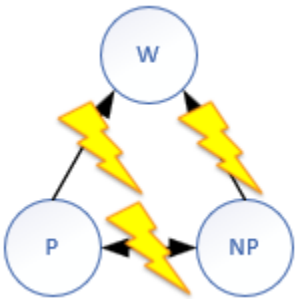
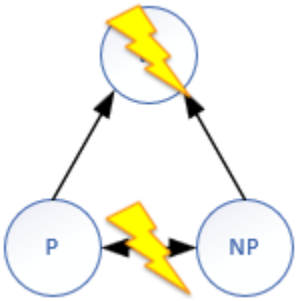
Figure	Preferred	Non-preferred	Link: Replication	Link: Preferred to witness	Link: Non-preferred to witness	Witness	Expected Outcome
	OK	X	OK	OK	OK	X	Volume is offline on both arrays
	OK	OK	X	OK	OK	OK	Volume is online on the Preferred
	OK	OK	X	X	OK	OK	Two scenarios depending on the order of the failures **
	OK	OK	X	OK	X	OK	Volume is online on the Preferred
	OK	OK	X	X	X	OK	Volume is offline on both arrays

Figure	Preferred	Non-preferred	Link: Replication	Link: Preferred to witness	Link: Non-preferred to witness	Witness	Expected Outcome
	OK	OK	X	OK	OK	X	

* Improvement in Metro Volume availability due to witness

** Two scenarios depending on the order of the failures

Scenario #1 Failure Order:

1. Witness link down from preferred system
2. Replication link lost
3. Outcome: Non-preferred stays online

Scenario #2 Failure Order:

1. Replication link lost
2. Metro session fractured – polarizes taking witness input
3. Witness link down from preferred system
4. Outcome: Preferred stays online

Note: Host connectivity status is not considered in any failure scenario response.

The multi failure situations as shown in Table 9 and **Error! Reference source not found.** expect the individual failures at almost the same time. In failure situations where both Preferred and Non-Preferred are not able to access the witness over some time, the metro sessions revert to Preferred (bias) mode and the expected outcome mirrors a configuration without a witness involved. The illustration below shows situations where the Preferred will remain online after a series of failure events.

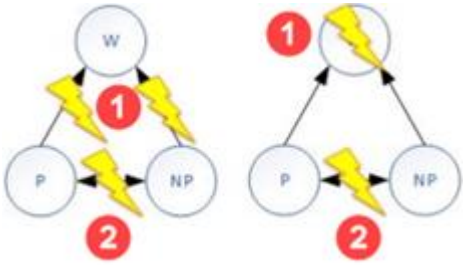


Figure 49. An ordered sequence of failure events with time in between could result in falling back to Preferred mode failure handling whereby Preferred will remain online for host access

Metro Volume use cases

Introduction

This section provides more examples of how you can use Metro Volume in various environments.

Zero-downtime SAN maintenance and data migration

With Metro Volume, you can perform maintenance activities without resulting in downtime on a PowerStore cluster. These tasks include taking a cluster offline to move its location, perform service-affecting enclosure or disk firmware updates, and migrate the volumes to a new cluster.

The requirements for this operation include:

- MPIO installed and appropriately configured on the host computers
- Host or hosts properly zoned to both PowerStore clusters
- Host connectivity configured on both clusters
- Synchronous replication link between clusters

Summary: Before a planned outage, Metro Volume can non-disruptively migrate volumes from one PowerStore cluster to another, enabling continuous operation for all workloads—even after one cluster has completely powered down.

Operation: In an on-demand, operator-driven process, Metro Volume can transparently move volumes from one PowerStore cluster to another. The workloads operate continuously. This ability enables several options for improved system operation:

- Redefine Metro Volumes at remote site as preferred role
- Shut down local site
- Reverse process after planned outage is completed

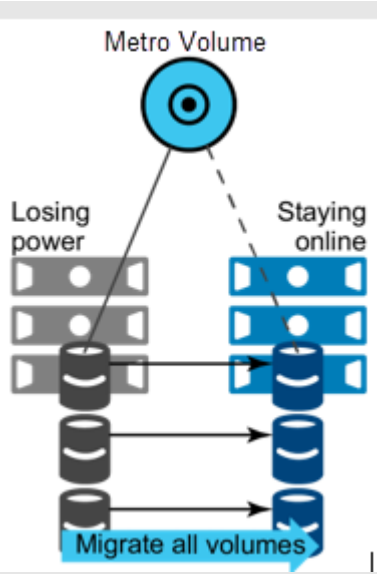


Figure 50. Maintaining availability of applications and services during scheduled maintenance

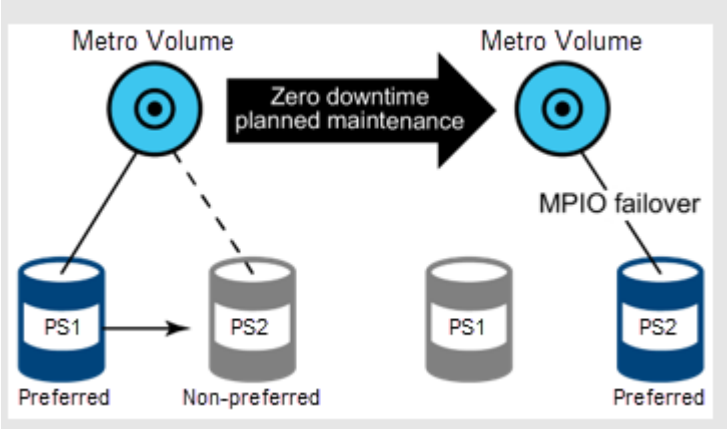


Figure 51. Applications remain continuously available on PS2 while PS1 undergoes maintenance

Storage migration for virtual machine migration

As VMware virtual machines are migrated from data center to data center, you can use Metro Volume to migrate the virtual machine disk files in bulk seamlessly.

This action involves the following:

1. Enable Metro Volume on the source cluster.

Migrate the virtual machines using vMotion.

2. End Metro Volume.

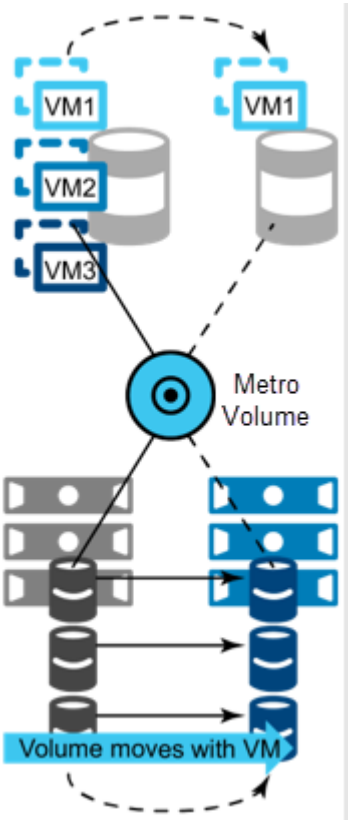


Figure 52. Storage follows the application (server virtualization)

The requirements for this operation include the following:

- Host or hosts properly zoned to both PowerStore clusters
- Host connectivity configured properly on both arrays
- Stretched Layer 2 networking between source and destination sites

  5 ms or less latency on synchronous replication link between sites

Disaster avoidance and continuous availability

In anticipation of an outage (for instance, an approaching hurricane in a coastal region) or after an unplanned outage has occurred, Metro Volume migrates applications, workloads, and data to recovery systems. Metro Volume used in this manner can prevent data loss and minimize application downtime.

Operation: Metro Volume can transparently move volumes from one PowerStore cluster to another. The applications operate continuously. This enables several options for improved system operation and continuous availability:

1. Enable Metro Volume on the source cluster.

Migrate the virtual machines using vMotion.

2. Manually or automatically recover applications at remote site as needed.
3. Use this same process to move applications seamlessly back to their original locations using Metro Volume.

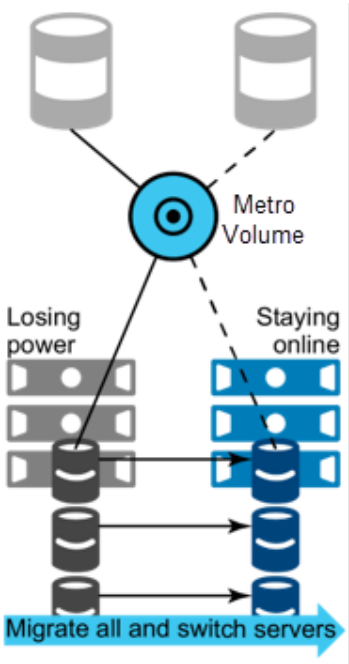


Figure 53. Disaster avoidance and continuous availability

Note: Leveraging Metro Volume for autonomous storage recovery or continuous uptime requires the witness component introduced in PowerStoreOS 3.6.

On-demand load distribution

In this use case, Metro Volume transparently distributes the workload, balances storage utilization, or balances I/O traffic between two PowerStore clusters.

Operation: In an on-demand, operator-driven process, Metro Volume can transparently move volumes from one PowerStore cluster to another. The applications operate continuously. This ability enables several options for improved system operation:

- Distribution of I/O workload
- Distribution of storage
- Distribution of front-end traffic load
- Reallocation of workload to match capabilities of heterogeneous systems

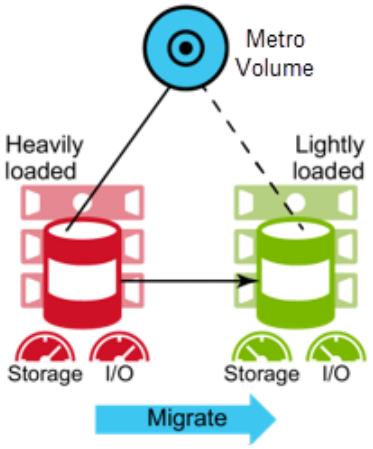


Figure 54. On-demand load distribution

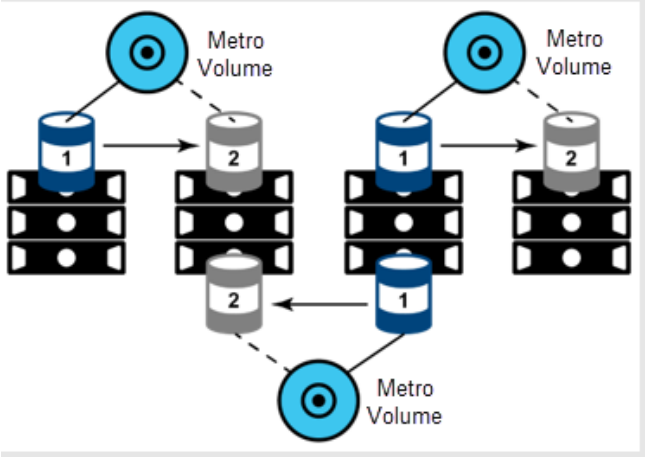


Figure 55. On-demand load distribution between multiple clusters

Metro Volume support for Microsoft

Required reading

Review the following documentation before proceeding with this section.

- Review the general sections of this document to learn more about PowerStore Metro Volume features, Metro Volume witness, and use cases.
- Dell PowerStore: Microsoft Hyper-V Best Practices white paper on the Dell PowerStore Info Hub.

Introduction

PowerStoreOS 4.0 extends Metro Volume migration, disaster avoidance, and load-balancing solution options to Windows Server and Hyper-V environments. Participating PowerStore clusters must be local (in the same building or data center) or within metro distance.

Metro Volume provides many benefits to Microsoft environments:

- Increased workload mobility
- Cross-site load balancing
- Fast recovery at a remote site within metro distance

- Disaster avoidance for planned downtime

Metro Volume supports:

- Long-term servicing channel (LTSC) versions of Windows Server 2016 (build 14393.2395 and later), Server 2019, and Server 2022, with local boot.
  - Stand-alone Windows Server hosts and stand-alone Hyper-V hosts.
  - Windows failover cluster nodes and clustered Hyper-V nodes.
  - Local clusters or stretched clusters over metro distance with uniform or nonuniform server mappings.
- Microsoft in-box Device-Specific Module (DSM) for MPIO.
- SCSI data volumes formatted as NTFS or ReFS (iSCSI and FC).
- SCSI cluster volumes formatted as NTFS or ReFS (iSCSI and FC).
  - Failover cluster volumes.
  - Cluster shared volumes (CSV).
- Pass-through disks on Hyper-V guest VMs

Note: PowerStore does not support Metro Volumes configured as boot-from-SAN (BfS) disks, in-guest iSCSI disks, or virtual Fibre Channel (vFC) disks.

MPIO

Windows behavior

When all data paths presented to a Windows server are optimal, the server identifies the paths as Active Optimized (AO) and will default to the Round Robin MPIO policy.

Round Robin spreads read and write I/O evenly over all available AO data paths. Path selection with Round Robin does not consider latency or bandwidth. All available AO data paths are considered optimal.

MPIO with PowerStore

PowerStore leverages asymmetrical logical unit access (ALUA) to inform hosts whether data paths are optimal or nonoptimal. Data paths presented from a PowerStore cluster to a Windows host will consist either of a combination of optimal paths and nonoptimal paths, or of all nonoptimal paths, depending on the host connectivity option selected.

When a Windows server detects optimal and nonoptimal paths to a data volume, the server will default to the Round Robin with Subset MPIO policy. A Windows server identifies the optimal paths as **Active Optimized** (AO), and the nonoptimal paths as **Active Unoptimized** (AU).

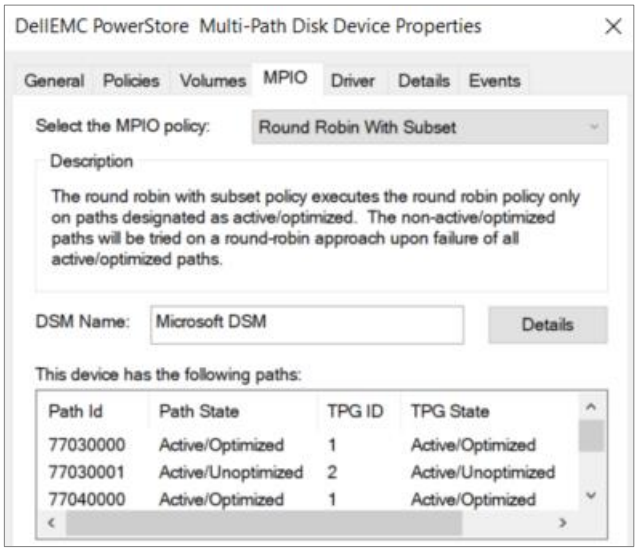


Figure 56. A Windows host defaults to Round Robin with Subset when a disk has Active Optimized and Active Unoptimized paths

See the E-Lab Host Connectivity Guide for Microsoft Windows for the recommended MPIO timeout values.

The PowerStore Manager **Host Connectivity Option** that you chose for a host server determines the type of paths presented to the host.

Host Connectivity Options

In PowerStore Manager, choose one of the following options for each host server.

- Local Connectivity
- Metro Connectivity - host is co-located with this system (the local system)
- Metro Connectivity - host is co-located with the remote system
- Metro Connectivity - host is co-located with both systems

Choose the option in Figure 57 when a host server is mapped to only one PowerStore system that is local to that host. The host will default to Round Robin with Subset and detect AO and AU paths to the PowerStore system. Use this option for hosts that are not uniformly mapped to both PowerStore systems in a Metro Volume configuration.

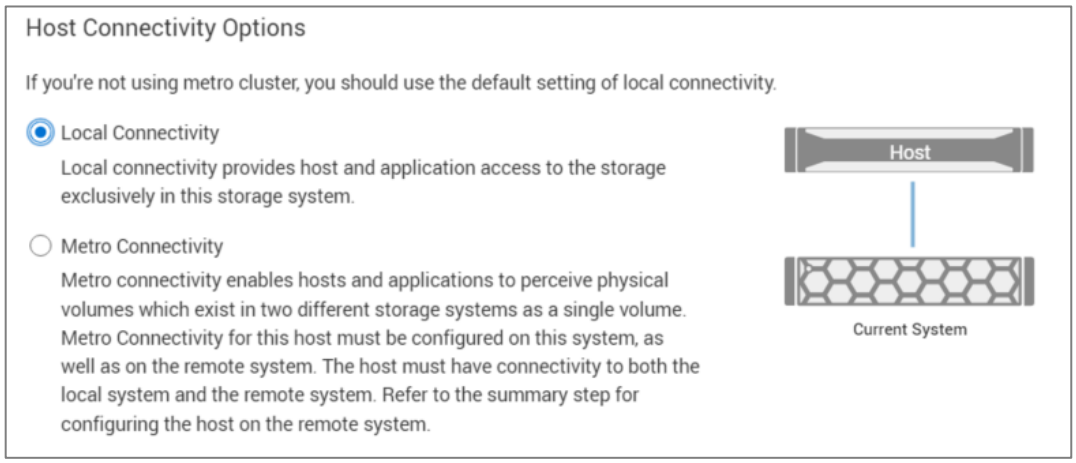


Figure 57. Local Connectivity option

Choose the option in Figure 58 when a local host is uniformly mapped to two PowerStore systems separated by metro distance. Metro distance injects latency between the current (local) and remote PowerStore system.

- The local host server paths to the local (current) PowerStore system on the left (solid blue line) consist of AO and AU paths.
- The local host server paths to the remote PowerStore system on the right (dashed blue line) are higher latency due to metro distance separation. PowerStore presents these paths to the local host as AU paths.

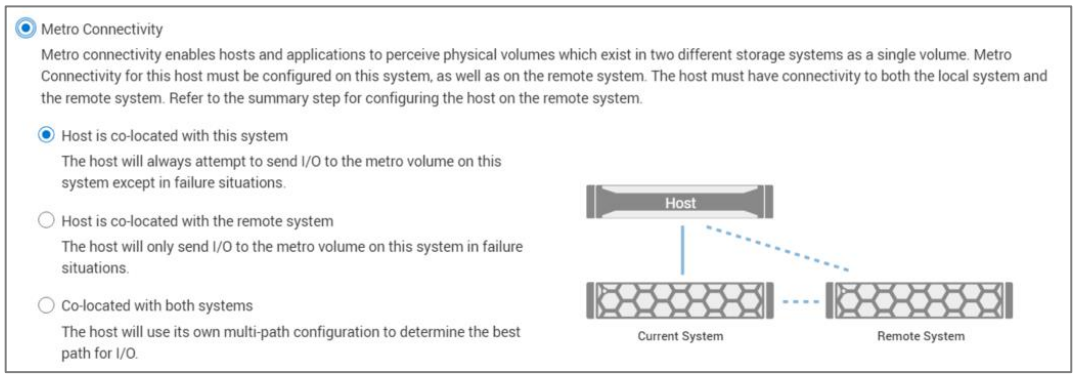


Figure 58. Metro Connectivity – host is co-located with the local PowerStore system

Choose the option in Figure 59 when a remote host is uniformly mapped to two PowerStore systems over metro distance. Greater distance injects more round-trip latency between the current and remote PowerStore system.

- The remote host server paths to the remote PowerStore system on the right are “local” to the remote host (represented by the solid blue line). The paths are AO and AU paths.
- The remote host server paths to the “current” PowerStore system on the left (dashed blue line) are higher latency due to metro distance separation. PowerStore presents these paths to the remote host as AU paths.

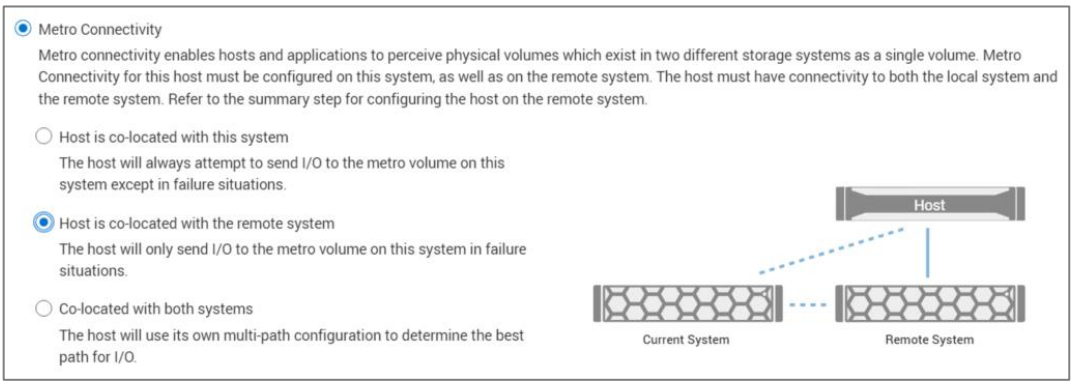


Figure 59. Metro Connectivity – host is co-located with the remote PowerStore system

Choose the option in Figure 60 when a host is uniformly mapped to two co-located PowerStore systems within the same building or data center with equidistant paths. This configuration provides similar latency and bandwidth for all data paths.

- The host server paths to the current (left) and remote (right) PowerStore system consist of AO and AU paths (represented by the solid blue lines).

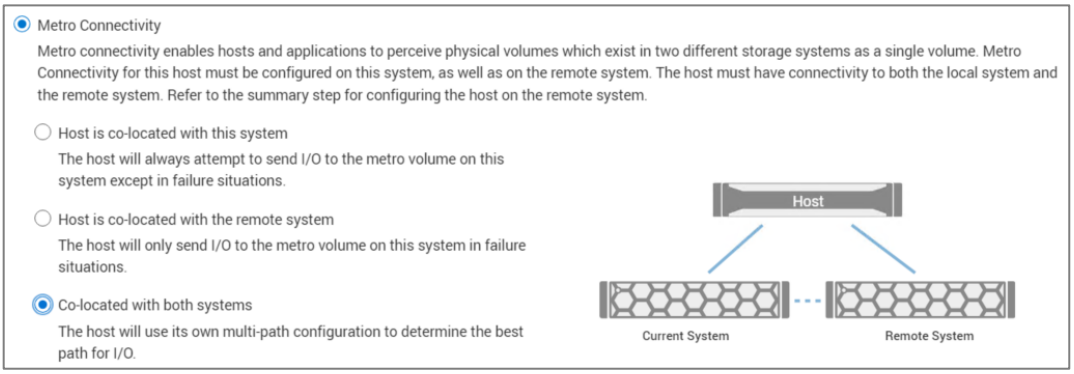


Figure 60. Metro Connectivity – host is co-located with both PowerStore systems

When a Windows server detects optimal and nonoptimal paths to a data volume, the server will default to the Round Robin with Subset MPIO policy.

A Windows server will route I/O traffic over all AO paths with Round Robin. AU paths are not used unless all AO paths become unavailable. When no AO paths are available, the server will route I/O traffic over all available AU paths with Round Robin. The server will resume using AO paths once one or more AO paths become available.

Other MPIO failover policies

Allow Windows hosts to automatically detect the type of PowerStore data paths (and default to Round Robin with Subset) as a best practice.

You can override the default MPIO behavior if your workload or situation requires it. Use the Microsoft `mpclaim` commandline utility to configure a host to use a specific MPIO policy if Round Robin with Subset is not wanted.

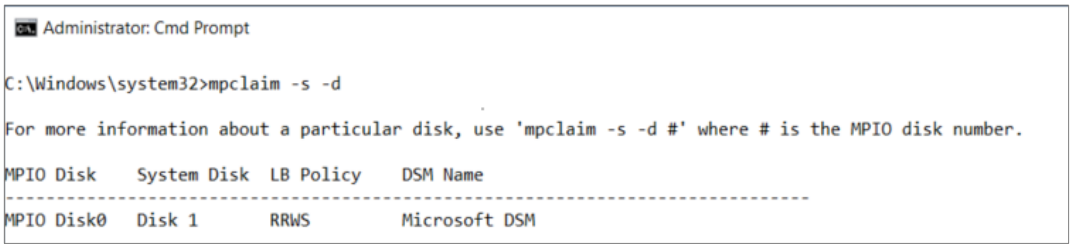


Figure 61. Use Microsoft mpclaim to alter default MPIO policy settings

Verify the behavior of altered MPIO policies or time-out settings in a test or development environment before doing so in production.

For more information see the Dell PowerStore: Microsoft Hyper-V Best Practices white paper on the Dell PowerStore Info Hub.

Quorum witness

Physical and virtual Windows Server failover clusters and Hyper-V clusters support a Microsoft quorum witness. A witness is a voting member that helps determine quorum for surviving server nodes when some nodes in a failover cluster or Hyper-V cluster become isolated or go offline.

Note: Metro Volume witness and Microsoft quorum witness function independently. They have no awareness of each other.

When to configure a Microsoft quorum witness

It is a best practice to configure a Microsoft quorum witness whenever you need a tie-breaking vote for surviving server nodes to achieve quorum if there is an outage. If your failover cluster or Hyper-V cluster (local or stretched) has an even number of server nodes, configure a quorum witness to optimize resiliency as a best practice. However, a Microsoft quorum witness is not required.

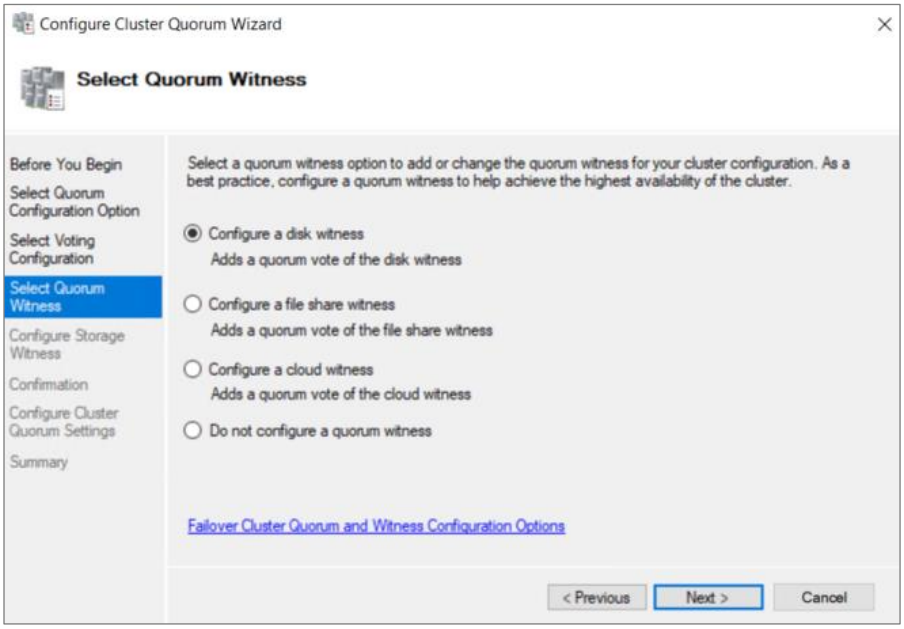


Figure 62. Microsoft Quorum Witness options

In a Metro Volume configuration, a Microsoft quorum witness is an important component that optimizes the resiliency of your clustered Microsoft environment. Microsoft provides several witness options.

File share witness

The optimal quorum configuration for clustered Microsoft environments on Metro Volume is a file share witness or cloud-based witness. The witness must have a reliable low-latency connection to all nodes in the cluster.

If you have a stretched cluster, place the file share witness at a third site (or configure a cloud-based witness) to avoid site bias. Site bias can cause an unintended outage if a file share witness is not placed at a third site.

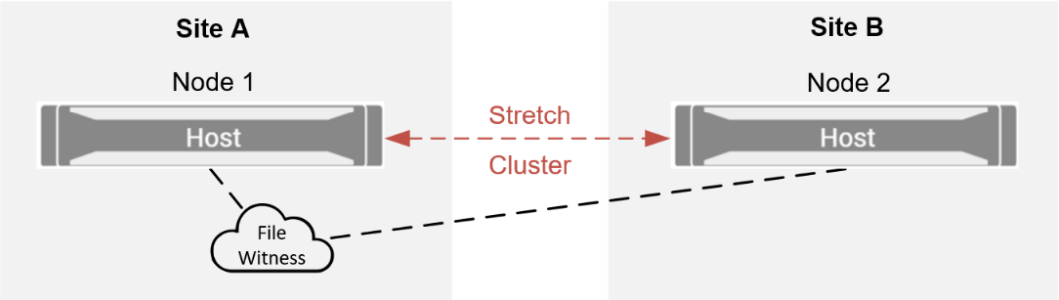


Figure 63. Placing a file share witness at Site A (or at Site B) creates site bias

For example:

1. Site A has a Windows failover cluster node and a file share witness (not an optimal configuration).
2. Site B also has a Windows failover cluster node.
3. A stretched cluster consisting of the two nodes is configured to use Metro Volume (with a Metro Volume witness) between Site A and Site B over metro distance.
4. Site A suffers an outage. The cluster node and the file share witness at Site A go offline.
5. Site B cannot access the file share witness because Site A is down.
6. Site A and Site B have one server node each (one vote each).
7. The server node at Site B requires a vote from the file share witness to achieve quorum (two votes).
    a. The server node at Site B goes offline because quorum cannot be achieved.
    b. Two votes are required in this scenario to achieve a quorum.
8. The result is a complete (unintended) outage for the two-node cluster.

Remedy: Place the Microsoft file share witness at a third site (Site C). Given the same scenario, the surviving node at Site B (one vote) and the file share witness (one vote) can achieve quorum (two votes) and stay online.

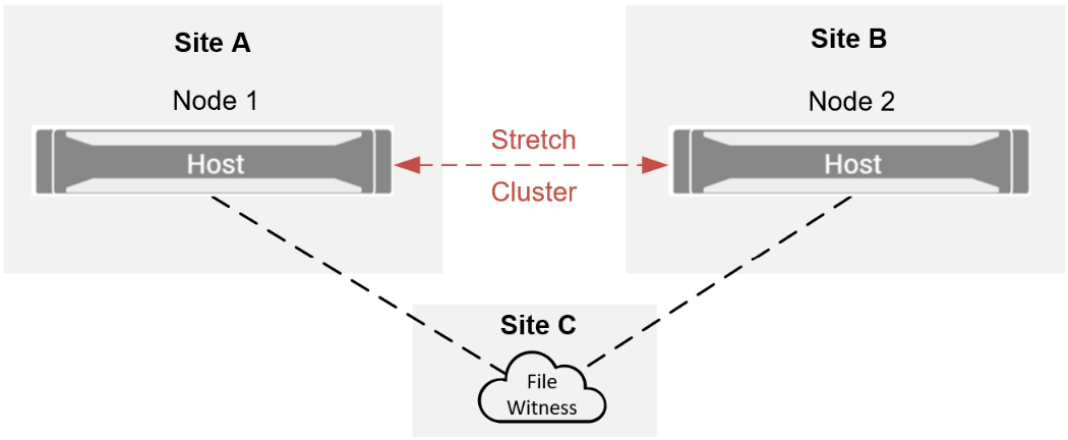


Figure 64. Place the file share witness at a third site to avoid site bias

If a third site is not available for a file share witness, place the witness at the primary (preferred) site.

Disk witness

You can also configure a disk as a quorum witness. A disk witness is less resilient than a file share witness if you have a stretch-cluster configuration over metro distance. It is not possible to place a disk witness at a third site.

- Create a witness disk. The witness disk should be a Metro Volume.
- Maintain awareness of which site and which Windows Server node owns a witness disk if you chose this method to achieve quorum.
  - If you reboot a Windows Server or Hyper-V node that owns a witness disk, the ownership is moved automatically to another server node in the cluster.
  - Ensure that a Windows or Hyper-V node at Site A (the preferred site) maintains ownership of the witness disk.
  - Manually reassign witness disk ownership when necessary to keep it owned by a server node at the preferred site.
  - You optimize the ability of the primary site to maintain quorum when a server node at the preferred site owns the witness disk.
- If a server node at Site A owns the witness disk, and Site A goes down unexpectedly:
  - Ownership of the witness disk may not transition automatically or cleanly to a surviving server node at Site B.
  - The surviving nodes at Site B may fail to achieve quorum if a vote from the witness disk is needed.
  - Site B will also go offline if the surviving nodes cannot achieve quorum.
- Microsoft has no awareness of Metro Volume functionality or behavior in the background. Windows Server node and cluster resilience depends on two critical factors.
  - Availability of data paths to shared volumes.

- The ability of surviving resources in a cluster to maintain quorum when there is an outage.

Note: Hyper-V guest VMs configured to use HA with Metro Volume should use shared virtual hard disks on a CSV. Use of in-guest iSCSI or virtual Fibre Channel (vFC) to achieve HA is not recommended. Use of pass-through disks is also not recommended except for temporary or specific use cases. To learn more about direct-attached and pass-through disks, see the Dell PowerStore: Microsoft Hyper-V Best Practices white paper on the Dell PowerStore Info Hub.

SCVMM/SCCM and PRO

Microsoft System Center Virtual Machine Manager (SCVMM) with System Center Configuration Manager (SCCM) supports intelligent placement and automatic migration of virtual machines. VM migration triggers include the level of node utilization across the server nodes within a server cluster. Disable Performance and Resource Optimization (PRO) when there are latency and bandwidth limitations for nonoptimized data paths to a remote site. Use the Manual action for VM placement instead.

Situational awareness is important when you manage VMs or a workload on a stretch cluster over metro distance. Where you place a workload impacts performance. Place a VM or workload at the preferred site with the preferred Metro Volume to optimize performance. Tools or automation might load-balance (move) VMs solely based on node utilization without regard to the performance of the data paths. Avoid a configuration where a VM might inadvertently get moved to a less used server node at Site B that has higher latency or lower bandwidth data paths.

Cluster Shared Volumes

A Microsoft cluster shared volume (CSV) is a shared data volume that is initially formatted as NTFS or ReFS. When you add the volume to a failover cluster as a CSV, the file system is changed from NTFS or ReFS to Cluster Shared Volume File System (CSVFS).

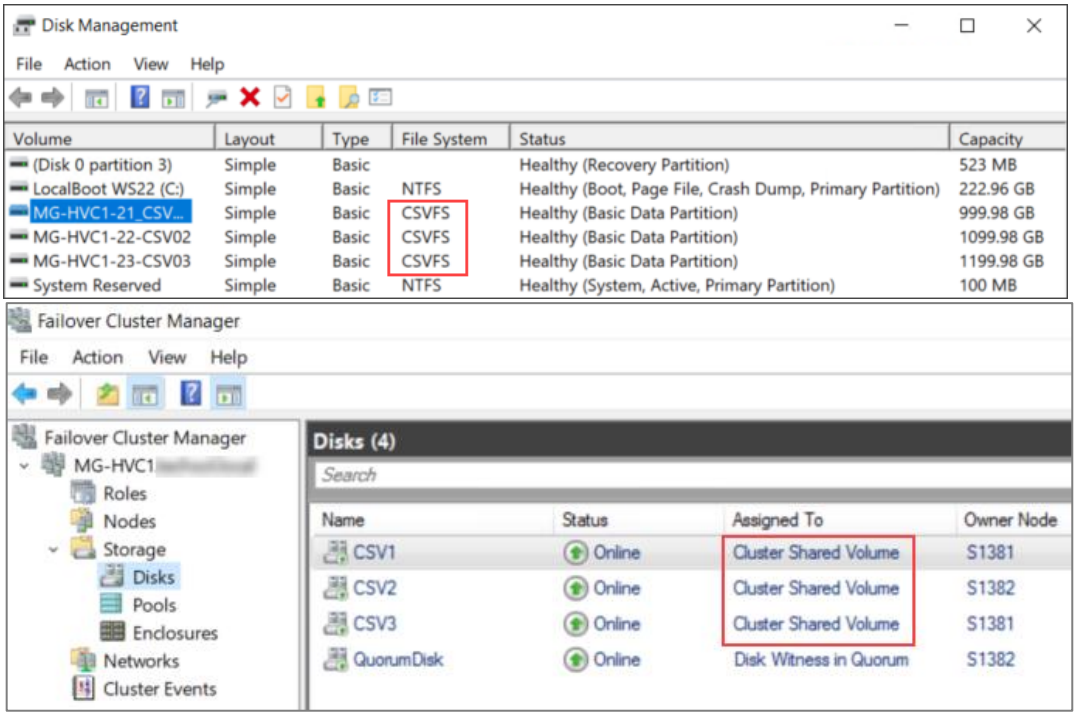


Figure 65. Cluster Shared Volumes use the CSVFS file system

A CSV enables concurrent read and write access from multiple server nodes regardless of which server node in the cluster owns the CSV.

Network redirection

Network redirection is an integrated resiliency feature that CSVs use to improve fault tolerance. The node that owns a CSV can redirect data I/O over the network to other server nodes if the other server nodes lose data path I/O access to the CSV. However, network redirection injects latency that will degrade performance. Network redirection is meant to be a temporary means of allowing I/O to continue. Resolve the issue that is causing network redirection as soon as possible.

Node ownership

Optimize CSV resiliency when using nonuniform server mappings. Make sure that a server node at the preferred Metro Volume site (Site A) owns the CSV. In this way, if the CSV goes into Network Redirected mode, the CSV owner is already at the preferred site.

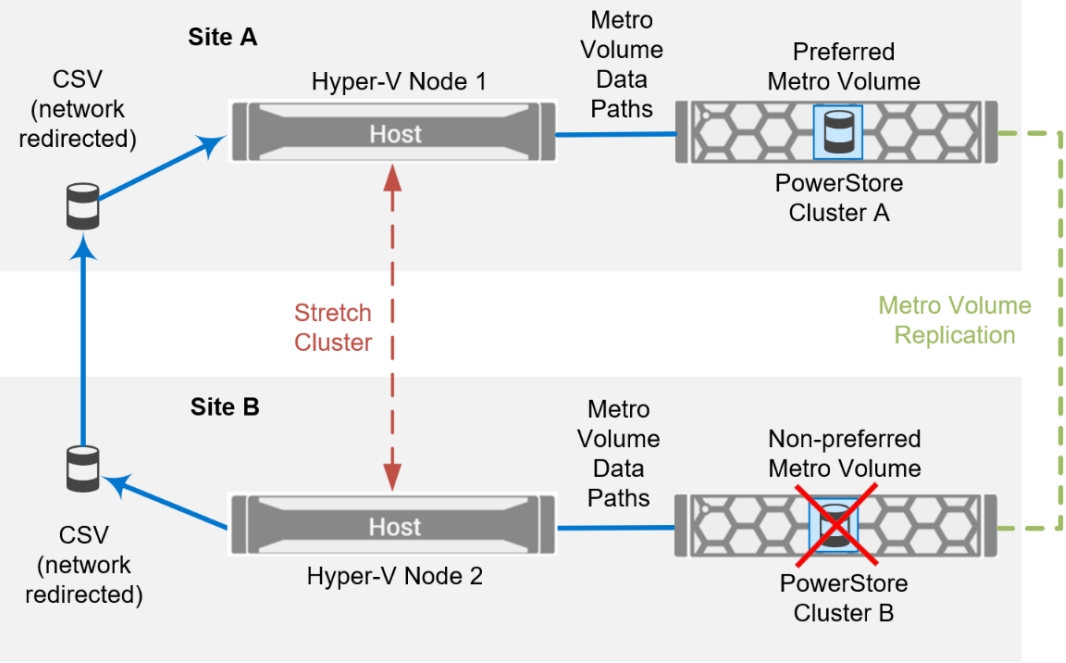


Figure 66. CSV resiliency with Metro Volume and Network Redirection

Configuration options and use cases

Metro Volume supports a diverse range of configuration options. Modify or expand on the examples and use cases in this section to suit your needs.

Single Site

A simple Metro Volume configuration at a single site might consist of a single Windows server host or Hyper-V host, or a failover cluster or Hyper-V cluster.

Use Cases:

- Move a workload from PowerStore A to PowerStore B as part of an upgrade.
- Move a workload to PowerStore B temporarily to perform offline maintenance on PowerStore A.

- Load-balance between PowerStore A and PowerStore B.

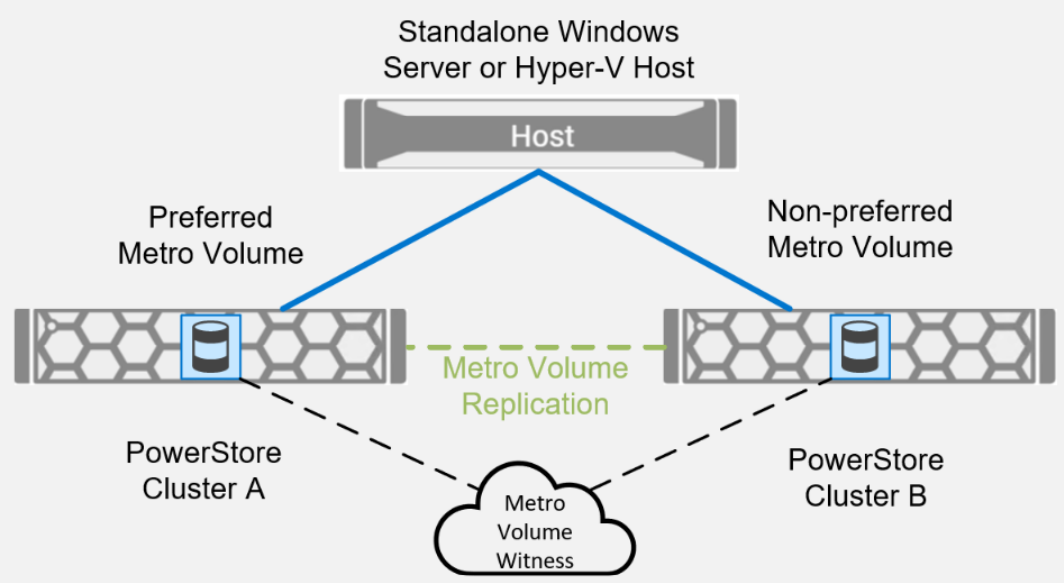


Figure 67. Metro Volume with a single Windows host at one site

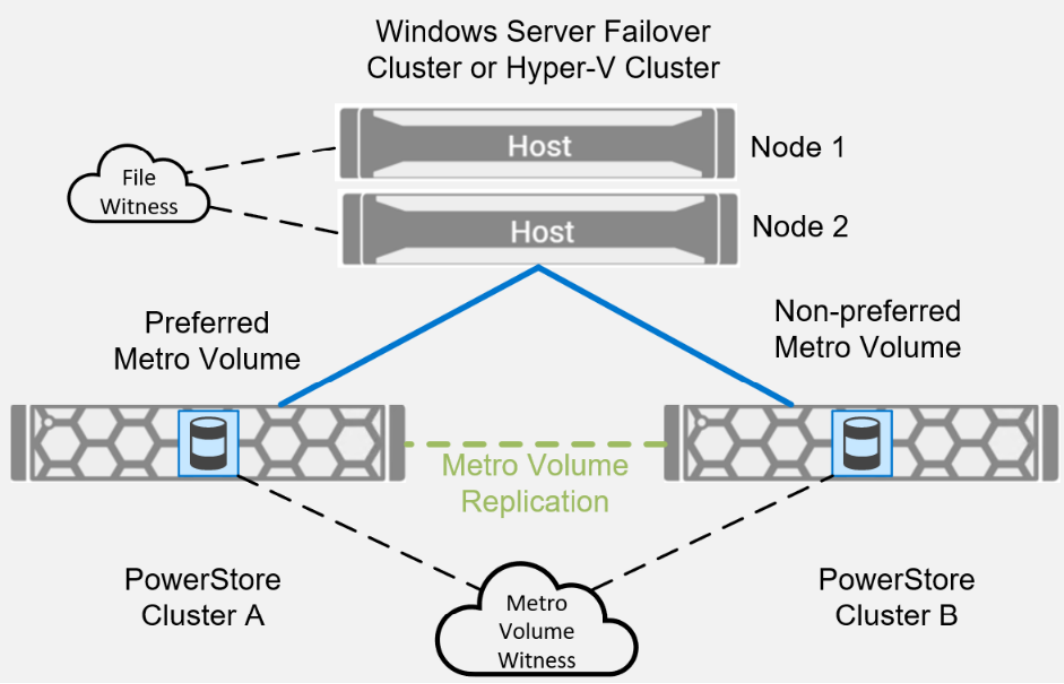


Figure 68. Metro Volume with a Windows Server failover cluster/Hyper-V cluster at one site with a file witness

Configure a witness disk as a quorum disk as an alternate way for your cluster to achieve quorum with failure scenarios.

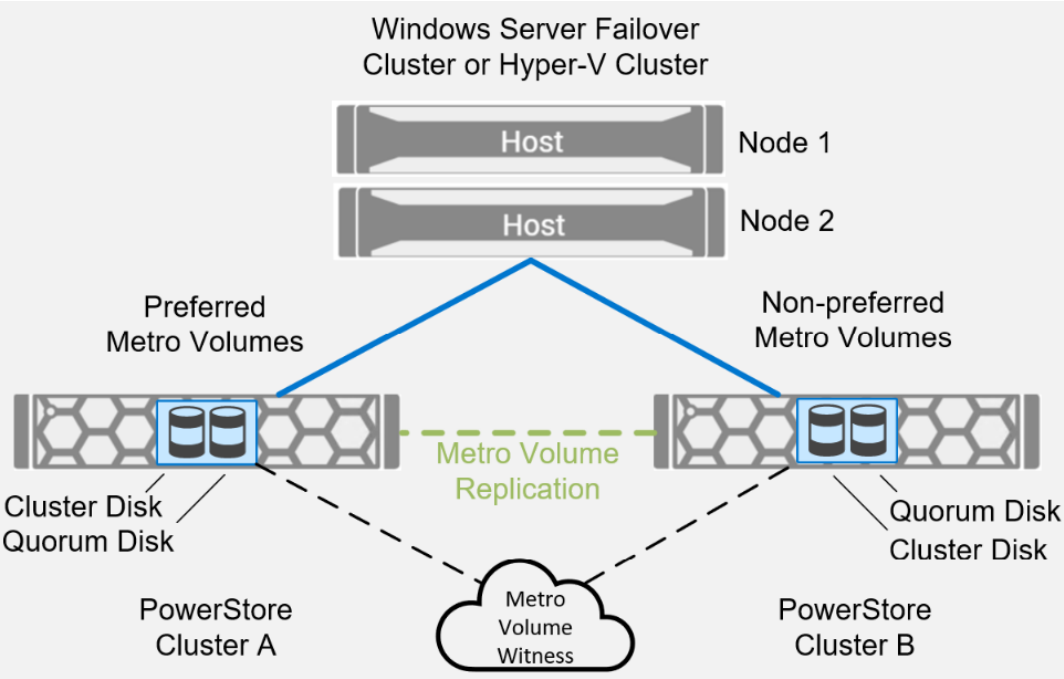


Figure 69. Metro Volume with a Windows Server failover cluster/Hyper-V cluster at a single site with a quorum disk witness

Configuration notes:

- All components are in the same building or data center.
- A Microsoft file share witness or quorum disk witness is not needed if you have a single-server host.
- A Microsoft witness is recommended if you have an even number of server cluster nodes (two, four, six, and so on).
  - Configure a Microsoft file share witness as a best practice.
  - A quorum disk will also work well as a witness when all components are local (in the same data center).
- Server hosts/nodes are uniformly mapped to both PowerStore clusters.
- A Metro Volume witness increases resiliency but is optional. Place it locally, or remotely.
- All Metro Volume data paths are local, so they are equidistant and experience similar latency and bandwidth.
  - Choose the PowerStore Host Connectivity Option: **Metro Connectivity – Co-located with both systems**.
- A single site offers no protection against a site failure such as an unexpected disaster or power outage.

Multisite with a stretched failover cluster/Hyper-V cluster

Configure a second site within metro distance to take advantage of additional Metro Volume functionality, such as site-specific disaster avoidance. When you configure a second site, careful planning is required because it introduces complexity. If you configure a second site and use Metro Volume, use witnesses to increase resiliency.

- A third site is recommended for the Microsoft file share witness (optional) and the Metro Volume witness (optional) to avoid site bias.
- Place witnesses at the preferred site if a third site is not available.

Windows failover clustering and Hyper-V are not aware of Metro Volume and the abstraction that masks the underlying architecture. Failover clustering and Hyper-V behavior are a function of the availability and state of the data storage paths, and the ability of surviving nodes to achieve quorum given an outage. Careful planning is required to ensure wanted behavior with failure scenarios. Carefully review the impact of where you place critical components such as cluster disks and the quorum witness. A continual focus on maintaining situational awareness is required to avoid unexpected results.

Use cases:

- Move a workload from PowerStore A to PowerStore B as part of an upgrade.
- Move a workload to the PowerStore at Site B temporarily to perform offline maintenance of the PowerStore at site A.
- Load-balance between PowerStore A and PowerStore B.
- Disaster Avoidance.
  - Disaster avoidance requires that you have enough time to invoke your contingency plan.
    - When Site A will suffer an outage due to a planned event such as a power outage.
    - When Site A is at risk of a service-impacting natural event such as an approaching hurricane.
- Disaster Recovery:
  - Recover at Site B if site A suffers an unexpected and extended outage.

Note: Test your disaster recovery and disaster avoidance plans regularly. Make sure they will work when they are needed.

Figure 70 shows a stretched cluster with uniform server mappings, a Microsoft file share witness at a third site (optional), and a Metro Volume witness at a third site (optional).

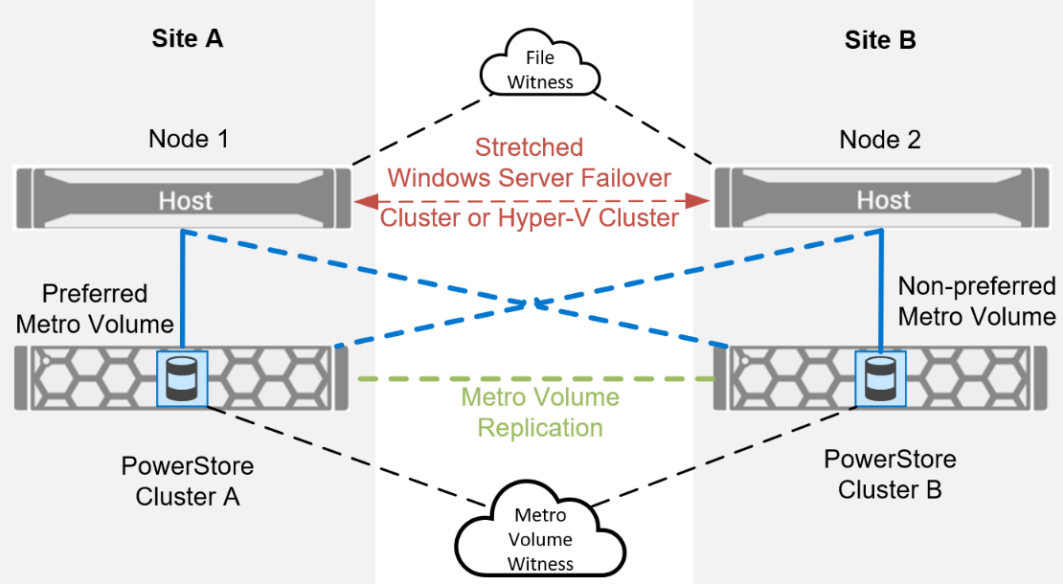


Figure 70. Metro Volume with a stretched cluster over metro distance with uniform server mappings.

Figure 71 is the same as Figure 70, except that it is configured to use nonuniform server mappings.

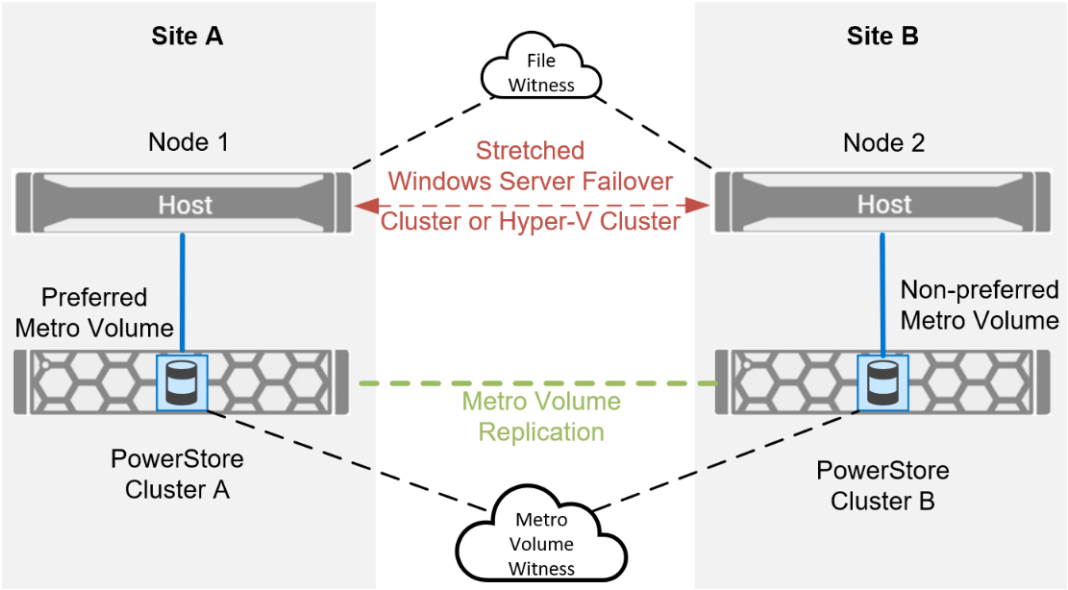


Figure 71. Metro Volume with a stretched cluster with nonuniform server mappings

Figure 72 and Figure 73 show the impact of site bias that favors Site A when a third site is not available.

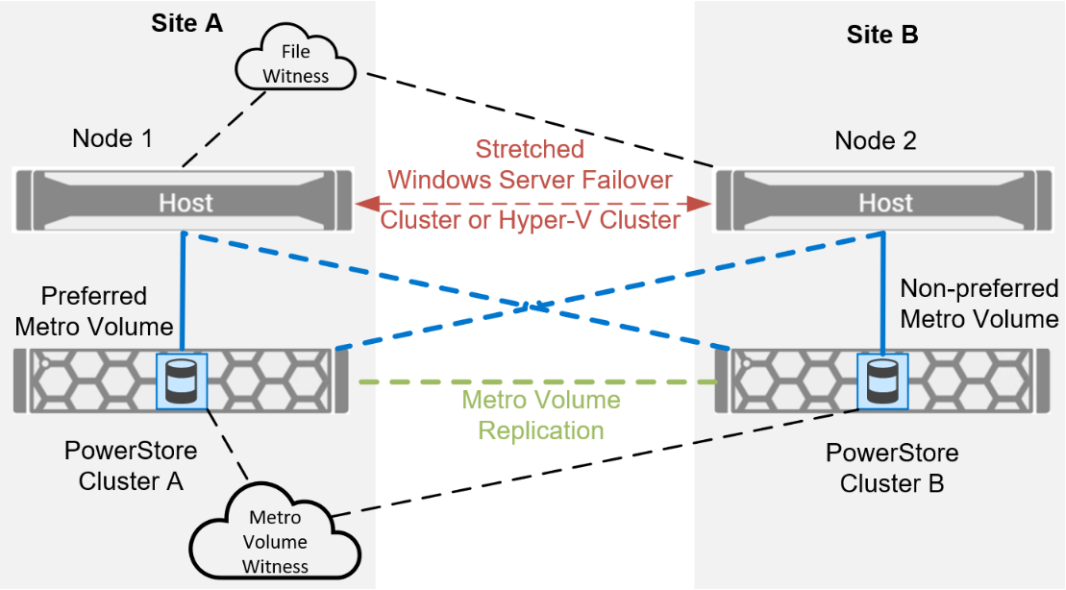


Figure 72. Metro Volume with uniform mappings and site bias for Site A.

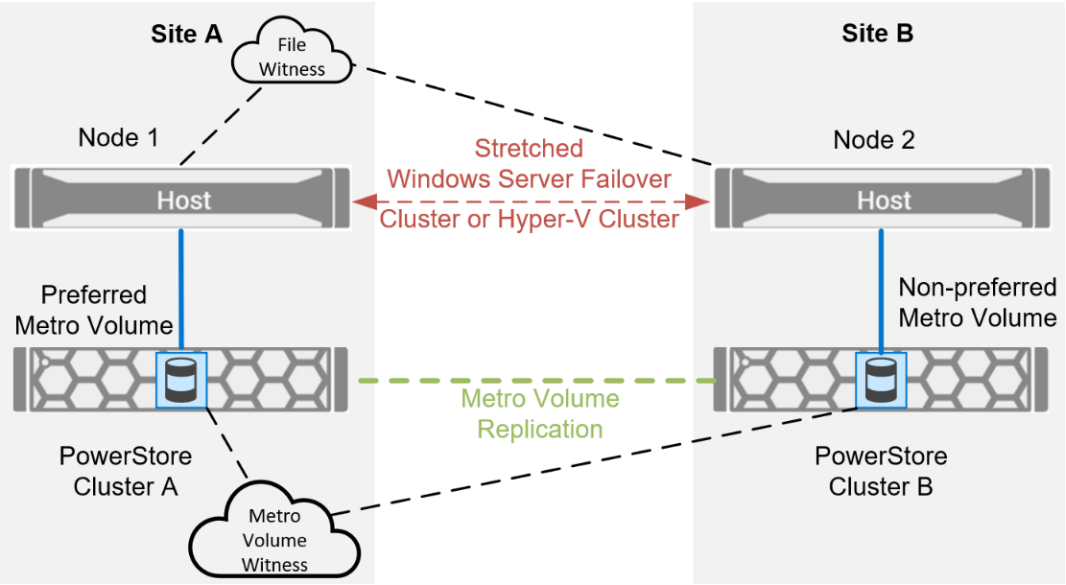


Figure 73. Metro Volume with nonuniform mappings and site bias for Site A

Figure 74 and Figure 75 show configuration examples with a Microsoft quorum disk witness (optional) instead of a file share witness, along with a Metro Volume witness (optional) at a third site.

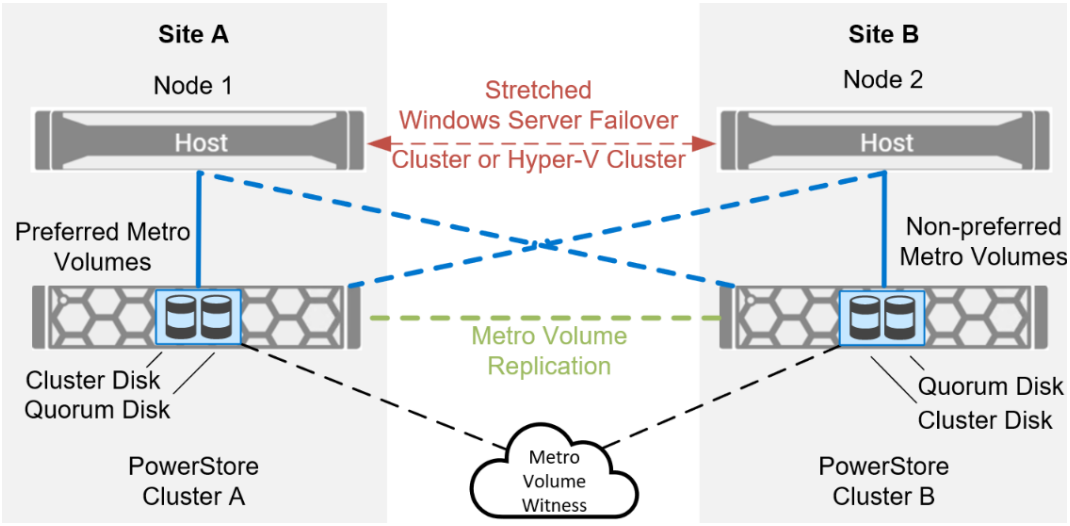


Figure 74. Uniform mappings with a quorum disk and Metro Volume witness

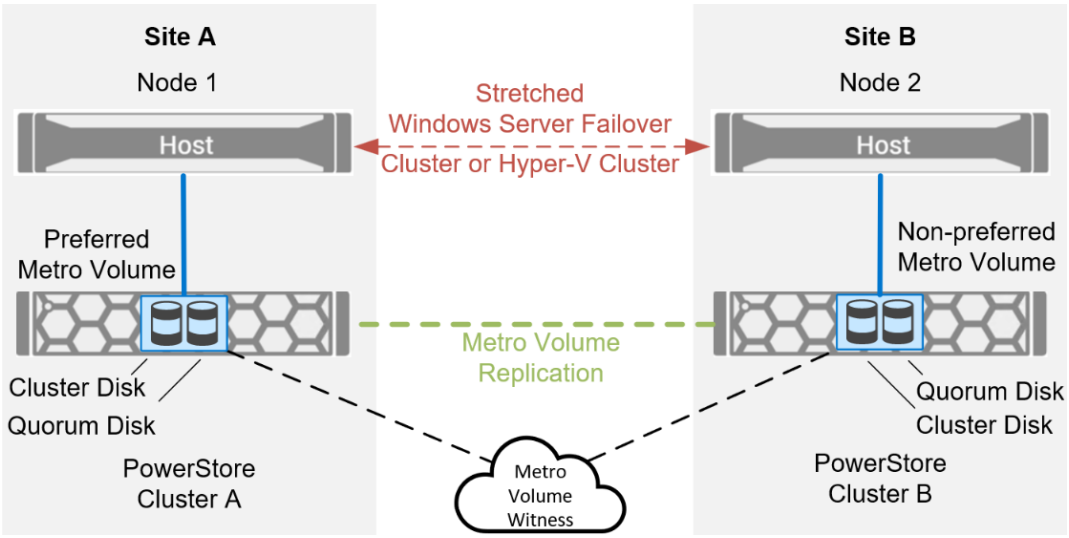


Figure 75. Nonuniform mappings with a quorum disk and Metro Volume witness

Figure 76 and Figure 77 show configuration examples with a Microsoft quorum disk witness (optional) instead of file witness, along with a Metro Volume witness (optional) at Site A (site bias).

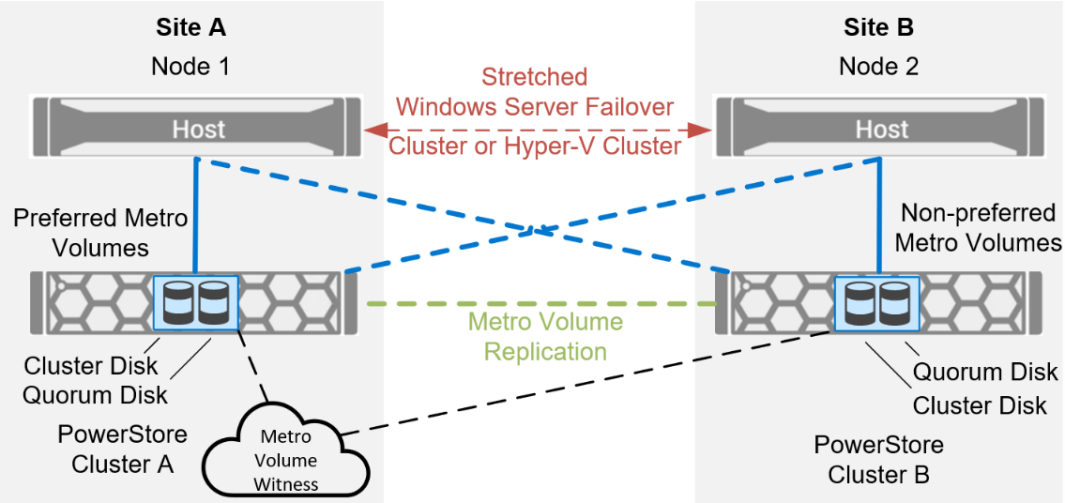


Figure 76. Site bias that favors Site A with uniform server mappings and a quorum disk

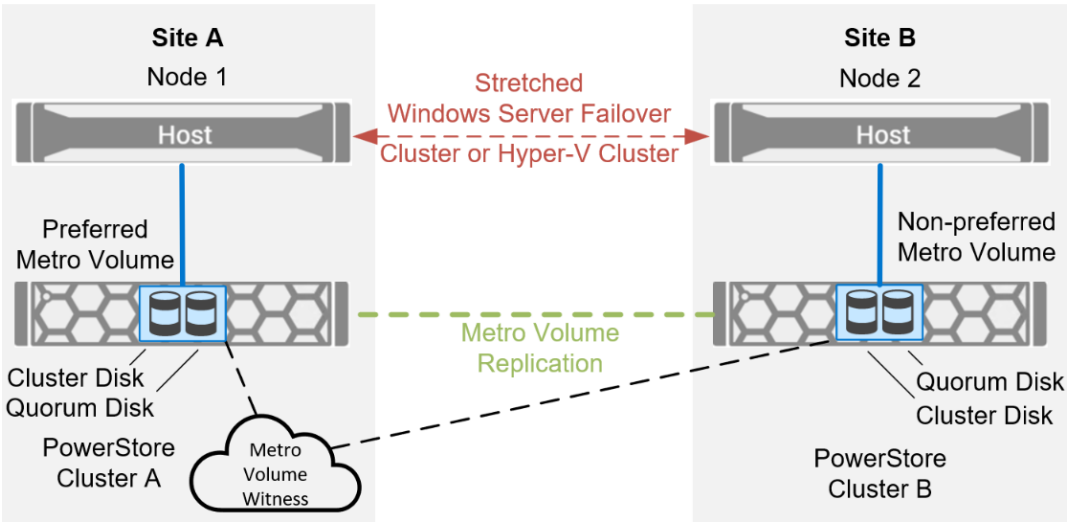


Figure 77. Site bias that favors Site A with nonuniform mappings and a quorum disk

Data path examples

Several examples are shown in this section to help you understand MPIO path behavior given a Microsoft failover cluster or Hyper-V cluster configured locally, or over metro distance.

The examples show Fibre Channel (FC) connectivity with two fabrics for redundancy. Use of iSCSI is also supported. The architecture and MPIO path behavior is similar regardless of the transport.

Single site with co-located PowerStore systems

This example shows a path configuration example when you choose **Metro Connectivity - Co-located with both systems**.

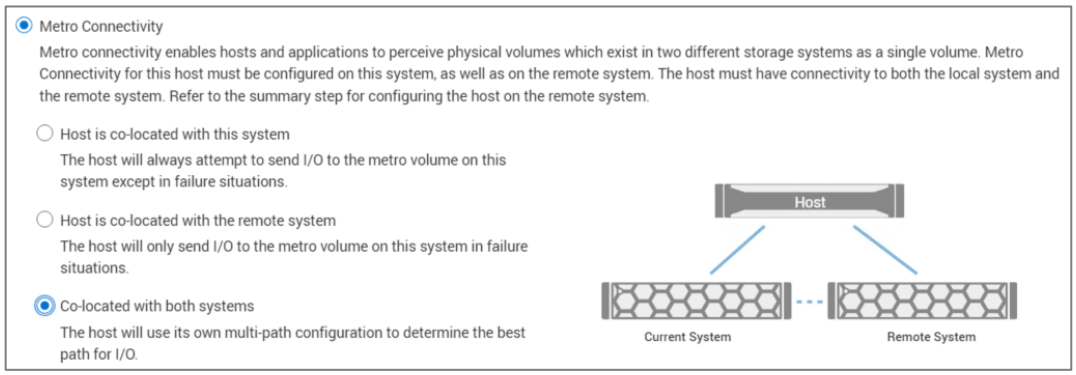


Figure 78. PowerStore Manager Metro Connectivity – Co-located with both systems option

See Figure 79 as you review the following configuration example:

- Node A on each PowerStore system owns a Metro Volume and presents AO paths to each cluster server node.
- Node B on each PowerStore system presents AU paths.
- Dual storage fabrics (or networks) increases resilience.
- Two Windows failover cluster nodes or Hyper-V cluster nodes are uniformly mapped as co-located hosts.
  - Each server node has a two-port FC HBA with one port connected to each fabric.
- Four FC ports are connected from each PowerStore system: two from Node A, and two from Node B.
- The configuration presents eight paths to each host. Four paths are AO, and four paths are AU.

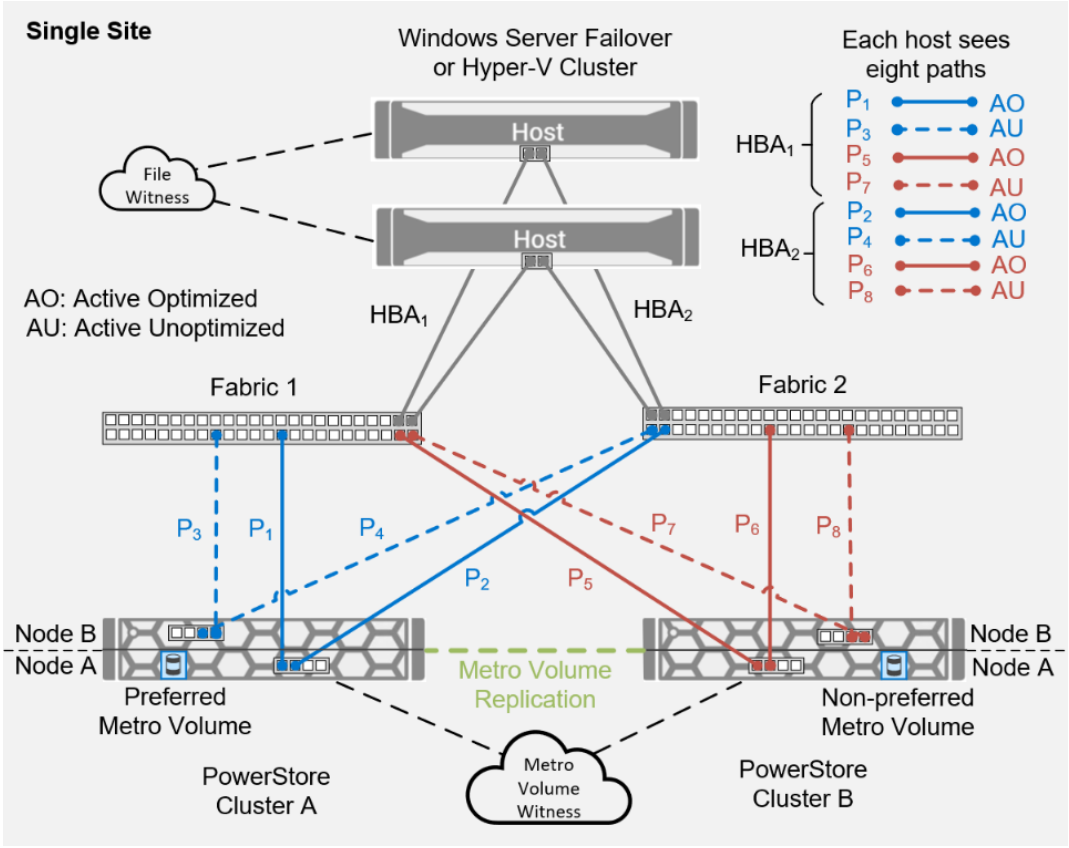


Figure 79. AO and AU paths in a co-located Metro Volume configuration

Multisite with non-uniform server mappings

This example shows a path configuration when you have a stretch cluster that is not uniformly mapped to both PowerStore systems. Choose the **Local Connectivity** option in PowerStore Manager for this configuration.

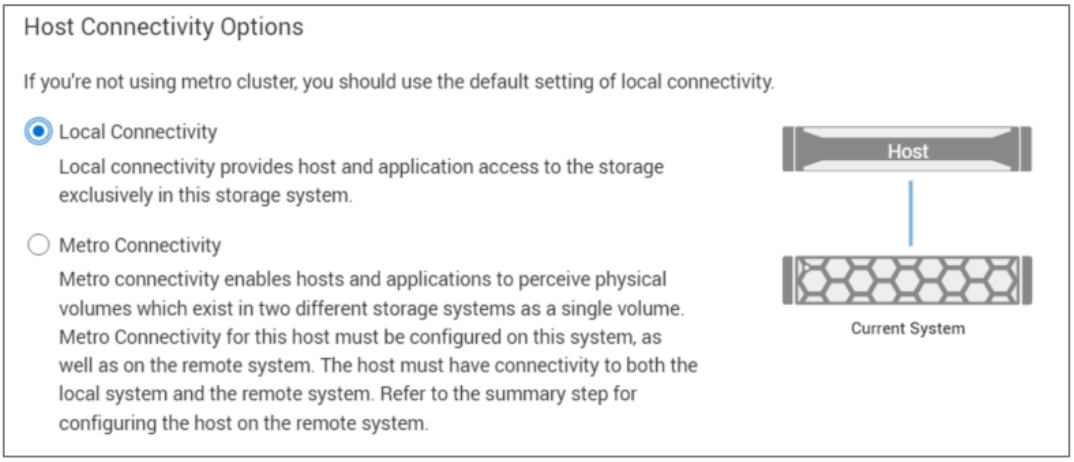


Figure 80. Use the PowerStore Manager Local Connectivity option for non-uniform mappings

See Figure 81 as you review the following configuration example:

- Node A on each PowerStore system owns a Metro Volume and presents AO paths to each local cluster server node.
- Node B on each PowerStore presents AU paths.
- Dual fabrics (or networks) increases resilience.
- Two Windows failover cluster or Hyper-V cluster nodes are non-uniformly mapped as local hosts.
  - Each server node has a two-port FC HBA with one port connected to each fabric.
- Four FC ports are connected from each PowerStore system: two from Node A, and two from Node B.
- The configuration presents four paths to each host. Two paths are AO, and two paths are AU.

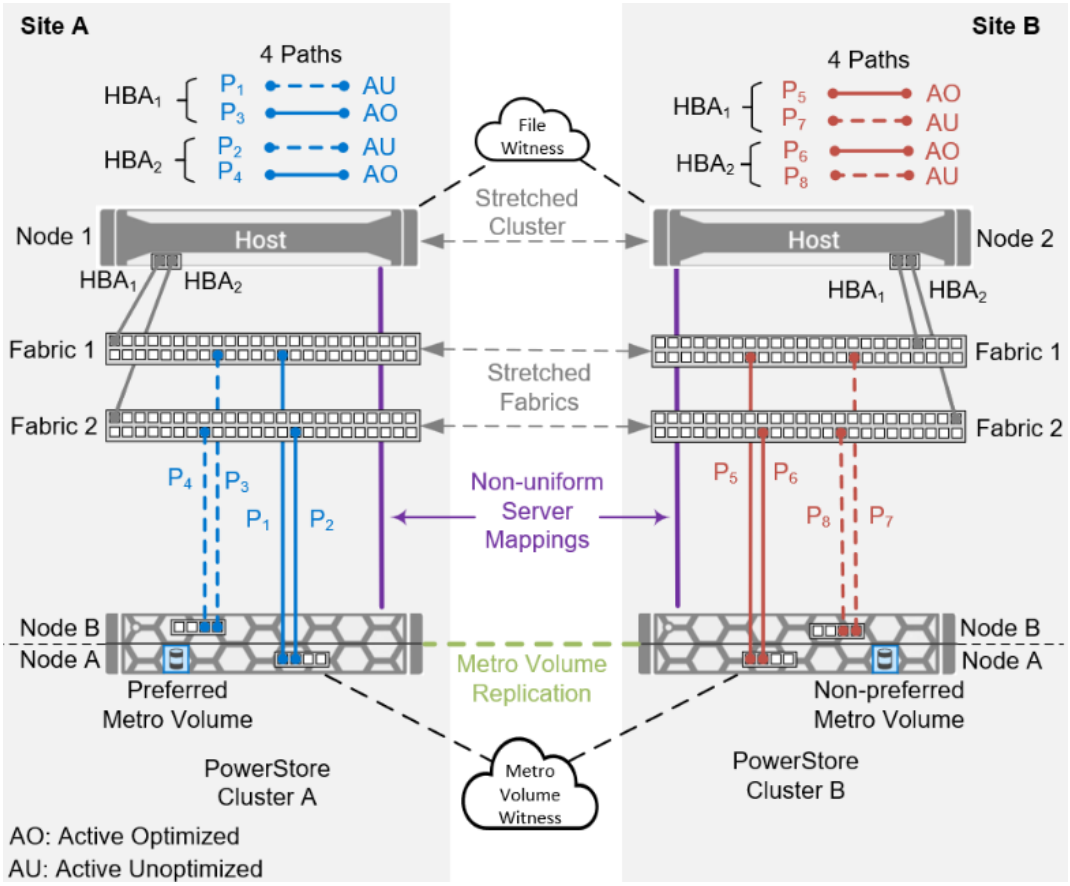


Figure 81. Non-uniform server mappings path example

Multisite with uniform server mappings

This example shows a path configuration when you have a stretch cluster that is uniformly mapped to both PowerStore systems. Choose the correct **Host Connectivity** option in PowerStore Manager based on the location of the host server.

- **Host is co-located with this system** (with the “local” PowerStore system)

- **Host is co-located with the remote system** (with the “remote” PowerStore system)

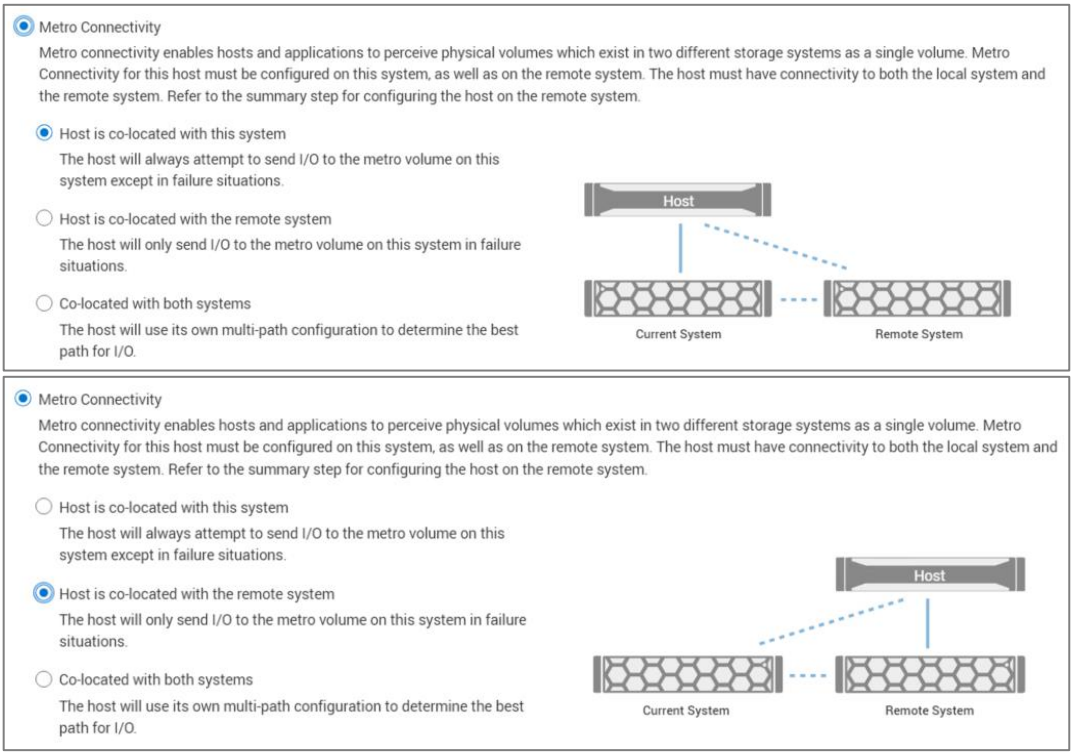


Figure 82. Choose the correct Metro Connectivity option: co-located with the local system, or co-located with the remote system

See Figure 83 as you review the following configuration example:

- Node A on each PowerStore system owns a Metro Volume and presents AO paths to each fabric.
  - The **Host Configuration** option selected in Figure 82 determines how the server nodes detect the paths.
- Node B on each PowerStore presents AU paths.
- Dual fabrics (or networks) increases resilience.
- Two Windows failover cluster or Hyper-V cluster nodes are uniformly mapped.
  - Each server node has a two-port FC HBA with one port connected to each fabric.
- Four FC ports are connected from each PowerStore system: two from Node A, and two from Node B.
- The configuration presents eight paths to each server node. Two paths are AO, and six paths are AU.

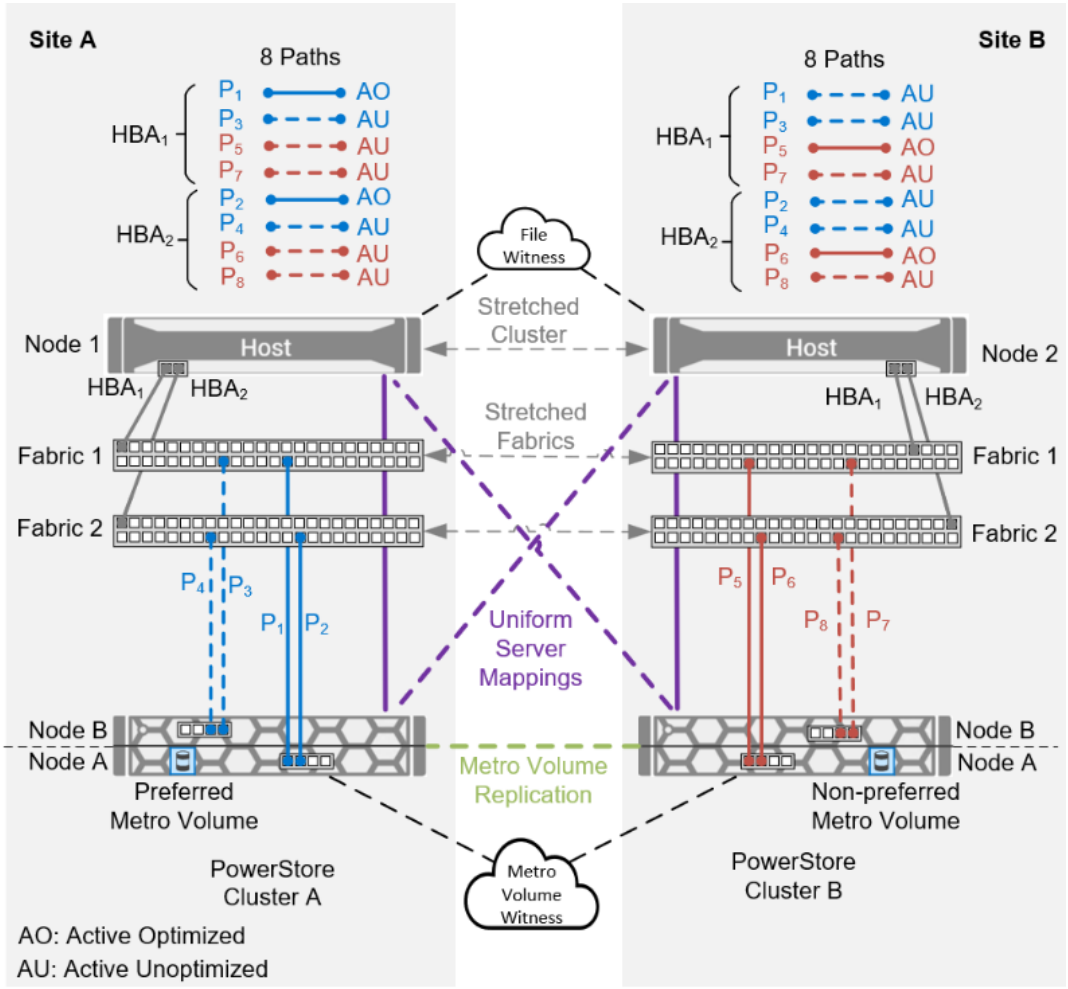


Figure 83. Uniform server mappings path example

Disk Rescans

Maintain situational awareness when managing stand-alone or clustered Windows or Hyper-V servers with PowerStore Metro Volume. Administrators need to understand and anticipate the behavior of Windows and Hyper-V when there are changes to the disks or data I/O paths.

For example, a stand-alone or clustered server might sit indefinitely without detecting the presence of a new volume or a new (or restored) data path. A manual rescan is often required to force the server to detect changes. A failure to understand this behavior might cause unintended consequences, including service interruptions, as you manage your environment.

Perform a manual disk rescan whenever a server does not automatically detect changes to disks or data I/O paths.

- Windows Disk Management UI > Action > Rescan Disks.
- PowerShell command `Update-HostStorageCache`.
- Reboot the server. This action is not practical in a production environment.

When a disk rescan is needed in a failover cluster or Hyper-V cluster environment, perform the action on all server nodes in the cluster to make sure they are consistent. You can use PowerShell to automate a disk rescan on all nodes in the cluster to save time.

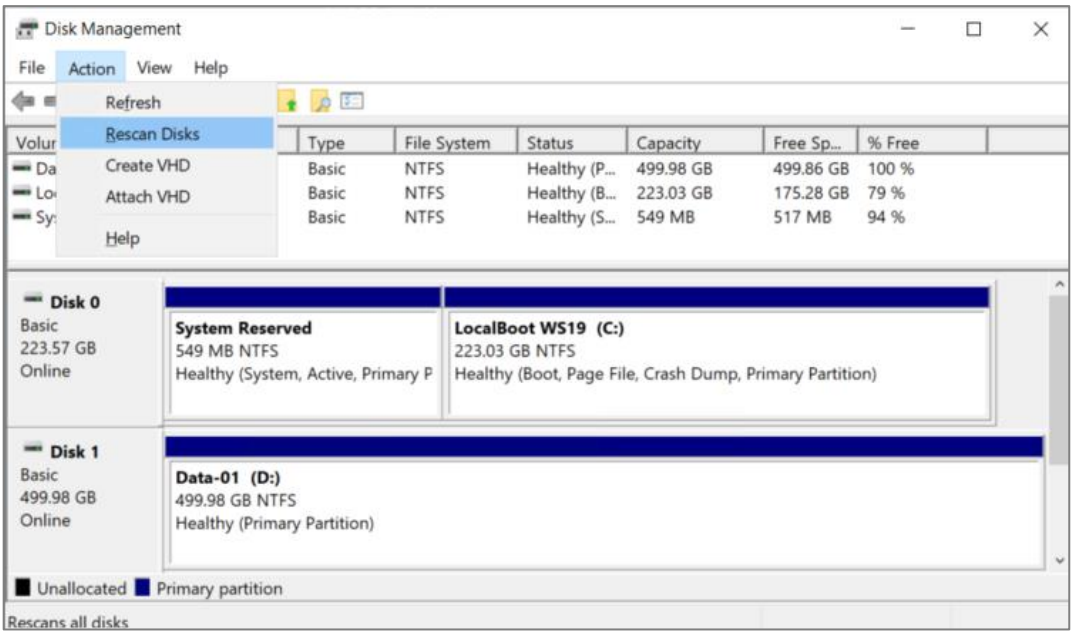


Figure 84. Run Rescan Disks from the Disk Management UI

Administrators might need to run manual disk rescans at critical points for many operations.

- Configuration changes, such as adding or removing a data volume or data path
- Data migrations
- Planned or unexpected failure scenarios
- Disaster avoidance
- Disaster recovery

For a Hyper-V cluster, place your VMs on CSVs as a best practice. Placing VMs on a regular cluster disk is supported but not recommended. CSVs are more resilient and require fewer manual disk rescans in failure scenarios.

Test your maintenance and business continuity plans to verify Windows behavior so you can work around disk and path discovery limitations. Document the critical points where manual disk rescans are needed before completing the next steps. Where possible, automate processes such as contingency plans to reduce the risk of missing steps or making mistakes due to human error.

Metro Volume support for Linux

Introduction

Metro volume was initially introduced in PowerStoreOS 3.0 for the VMware environment. With PowerStoreOS 4.0, its capabilities have been expanded to include various Linux operating system versions and configuration. PowerStoreOS 4.0 adds the support of SCSI-3 reservations and registrations for Metro Volume, enabling it to be used in a

standalone Linux host and clustered Linux hosts configuration. PowerStore manages the synchronization of data and SCSI reservations across PowerStore clusters. In addition, PowerStoreOS 4.0 allows grouping several volumes into a single Metro Volume Group, allowing them to be managed in a logical unit.

The previous sections, Metro Volume and Witness, lay the groundwork for understanding Metro Volume, detailing its functionality, operation, and various failure protection scenarios. It is recommended to familiar yourself with this information. This section focuses on the interactions and ALUA states of Metro Volumes on Linux.

Metro Volume is a storage-level feature that requires no modifications to the Linux hosts or applications. It is a stretched volume spanning across two PowerStore clusters. Each Metro Volume consists of two volumes, one on each cluster, synchronized bidirectionally over a replication network.

Requirements for Metro Volume for Linux

To use this feature, the following requirements must be met:

- Two PowerStore clusters connected by a replication network.
- Configure the two PowerStore Clusters as Remote Systems to enable replication.
- Linux hosts should have SCSI connectivity to one of the PowerStore clusters in a non-uniform configuration.
- Linux hosts must have SCSI connectivity to both PowerStore clusters in a uniform configuration.
- Metro Volume only supports FC and iSCSI protocols for host access.
- Although optional, Metro Witness is highly recommended for additional resiliency. It acts as a tie-breaker in split-brain situations and further mitigates risks of unwanted downtime and interruption. For more information about the witness, see the Witness section in this document.

Metro Volume for Linux adds support of Red Hat Enterprise Linux (RHEL), SUSE Linux Enterprise Server (SLES, and Oracle Linux (OL) for standalone hosts, and RHEL and SLES for clustered hosts.

The supported list of operating systems changes over time. For the most current information, refer to the Dell PowerStore Simple Support Matrix on the Dell E-Lab Navigator portal.

ALUA optimization

PowerStore supports ALUA in implicit mode. This means that it sets the device's Target Port Group states and controls the ALUA states, rather than the host. For standard volumes, ALUA optimizes the paths between the nodes within a PowerStore cluster. For Metro Volumes, ALUA optimizes the paths between the nodes across two PowerStore clusters. ALUA optimization depends on the **host connectivity** configuration and the **Node Affinity** of the volume in PowerStore clusters.

Each PowerStore appliance features two controller nodes. Node Affinity is a feature that enables PowerStore to define the node providing optimized I/O for a volume.

The sections Host connectivity and Host configuration detail the various options and configuration. Users can configure the host connectivity in the PowerStore Manager UI with one of the following options:

- **Host is co-located with local system**: Use this option when the Linux host and this PowerStore cluster are co-located (low latency) at the same location.
- **Host is co-located with the remote system**: Use this option when the Linux host and the remote PowerStore cluster are co-located at the same location.
- **Host is co-located with both systems**: Use this option when the Linux host is equidistant from both PowerStore clusters.

To view Host connectivity information in PowerStore Manager UI, go to **Compute** > **Host Information**. Use the column selection dropdown to add the Host Connectivity column to the table.

To view Node Affinity information of the volumes in PowerStore Manager UI, go to **Storage** > **Volumes**. Use the column selection dropdown to add the Node Affinity column to the table.

Linux MPIO

For Metro Volume to be effective, deploying Multipath I/O software on the connected Linux hosts is required. PowerStore supports the native Linux multipath I/O software, Device Mapper Multipath (DM-Multipath). Red Hat based Linux includes the multipath software in the **device-mapper-multipath** package. SUSE Linux Enterprise Server includes it in the **multipath-tools** package. PowerStore also supports Dell PowerPath. However, this paper does not address its configuration. For details about configuring PowerPath, see the E-Lab Host Connectivity Guides.

ALUA optimization and DM-Multipath work together to prioritize I/Os to the Active/Optimized paths for optimal performance. When a Linux host loses access to these paths, Active/Non-optimized paths become Active/Optimized paths, allowing I/Os to continue without disruption.

Native Linux multipath configuration

The E-Lab Host Connectivity Guide provides a recommended configuration for DM-Multipath, which Dell has tested to ensure the effective operation of multipath I/O software with PowerStore. The following is the multipath.conf configuration from the E-Lab Host Connectivity Guide.

```
defaults {
    polling_interval        5
    checker_timeout         15
    disable_changed_wwids   yes
    find_multipaths         no
}

devices {
    device {
        vendor                  DellEMC
        product                 PowerStore
        detect_prio             "yes"
```

```
        path_selector           "queue-length 0"
        path_grouping_policy    "group_by_prio"
        path_checker            tur
        failback                immediate
        fast_io_fail_tmo        5
        no_path_retry           3
        rr_min_io_rq            1
        max_sectors_kb          1024
        dev_loss_tmo            10
        hardware_handler        "1 alua"  ##Only for Oracle Linux
    }
    device {
        vendor                  .*
        product                 dellemc-powerstore
        uid_attribute           ID_WWN
        prio                    ana
        failback                immediate
        path_grouping_policy    "group_by_prio"
        path_checker            "none"
        path_selector           "queue-length 0"
        detect_prio             "yes"
        fast_io_fail_tmo        5
        no_path_retry           3
        rr_min_io_rq            1
        max_sectors_kb          1024
        dev_loss_tmo            10
    }
}
```

Device alias

To ensure consistent device names on Linux hosts in a cluster, and enhance reporting clarity, replace the default multipath device names with aliases.

To create a multipath device alias, add the following to the main configuration file at /etc/multipath.conf or a separate configuration file in the /etc/multipath/conf.d directory. The wwid is the world wide name (WWN) of the Metro Volume. You can find this information in the PowerStore Manager UI by going to **Storage** > **Volumes**. Use the column selection dropdown to add the WWN column to the table.

```
# cat /etc/multipath/conf.d/ps-vol-alias.conf
multipaths {
      multipath {
            wwid "368ccf0980076c2881eb573719c382287"
            alias metro-vol-001
      }
}
```

Uniform and non-uniform storage presentation

PowerStore supports uniform and non-uniform storage presentation. Uniform storage presentation provides the highest level of protection, allowing the Linux host to access the

Metro Volume from both PowerStore clusters. In a non-uniform storage presentation, the host has access to a single PowerStore cluster.

In a uniform storage presentation, both sides of a Metro Volume are mapped to a Linux host. The Metro Volume appears to the Linux operating system as a regular volume (LUN) with paths connected to both PowerStore clusters, because the two volumes behind the Metro Volume share the same WWID. Metro Volume allows multiple hosts to simultaneously read from and write to both sides of the volume, while maintaining the write-order and data integrity between the volumes.

Examining path priority and ALUA state

The multipath -ll command displays the paths and their priorities of a volume. The following example groups the paths according to their priority, as defined by the path_grouping_policy parameter in the multipath configuration file.

Multipath I/O software does not distinguish between a regular volume and a Metro Volume. However, when a Metro Volume is mapped to the host with a uniform storage presentation, the multipath command displays it as having more paths, and these paths have different priorities.

The paths with higher priority (50) are Active/Optimized paths, while those with lower priority (10) are Active/Non-optimized paths.

The example below shows a Metro Volume in a uniform storage presentation where the volume has paths connected to Array 1 (local array) and Array 2 (remote array).

```
metro-vol-001 (368ccf0980076c2881eb573719c382287) dm-7 DellEMC,PowerStore
size=200G features='1 queue_if_no_path' hwhandler='1 alua' wp=rw
|-+- policy='queue-length 0' prio=50 status=active
| |- 11:0:6:7 sdat 66:208 active ready running #Array 1,Node B,I/O module 0,Port 1
| `- 12:0:6:7 sdan 66:112 active ready running #Array 1,Node B I/O module 0,Port 0
`-+- policy='queue-length 0' prio=10 status=enabled
  |- 11:0:4:7 sdl  8:176  active ready running #Array 2,Node B,I/O module 0,Port 1
  |- 12:0:3:7 sdal 66:80  active ready running #Array 2,Node B,I/O module 0,Port 0
  |- 11:0:3:7 sdb  8:16   active ready running #Array 2,Node A,I/O module 0,Port 1
  |- 12:0:2:7 sdab 65:176 active ready running #Array 2,Node A,I/O module 0,Port 0
  |- 12:0:5:7 sdq  65:0   active ready running #Array 1,Node A,I/O module 0,Port 0
  `- 11:0:5:7 sdn  8:208  active ready running #Array 1,Node A,I/O module 0,Port 1
```

When the Metro Volume is in non-uniform storage presentation, multipath -ll shows that the volume has only paths connected to the local array.

```
metro-vol-003 (368ccf09800582e7fcfdaa4de979ccb9c) dm-9 DellEMC,PowerStore
size=80G features='1 queue_if_no_path' hwhandler='1 alua' wp=rw
|-+- policy='queue-length 0' prio=50 status=active
| |- 11:0:6:9 sdba 67:64  active ready running #Array 1,Node B,I/O module 0,Port 1
| `- 12:0:6:9 sdau 66:224 active ready running #Array 1,Node B,I/O module 0,Port 0
`-+- policy='queue-length 0' prio=10 status=enabled
  |- 11:0:5:9 sdw  65:96  active ready running #Array 1,Node A,I/O module 0,Port 1
  `- 12:0:5:9 sdaa 65:160 active ready running #Array 1,Node A,I/O module 0,Port 0
```

Additionally, the Node Affinity of the volume designates Node B of Array 1 to provide optimized I/O for host access. Therefore, paths connected to Array 1 Node B are in an Active/Optimized ALUA state, while all other paths are in Active/Non-optimized ALUA state.

Querying path and examining ALUA state using SCSI commands

We can also query each path directly using the scsi-inq and sg_rtpg commands. First retrieve the Target Port Group of a path, then examine the asymmetric access state of that Target Port Group. For example, the target port group of /dev/sdag is 0x2 and the asymmetric access state for the target port group 0x2 is active/optimized.

```
#  sg_inq -p 0x83 /dev/sdaq|egrep -Ei "Target port group:"
      Target port group: 0x2

# sg_rtpg -d /dev/sdaq|grep -EiA3 "target port group.*0x2"
  target port group id : 0x2 , Pref=0, Rtpg_fmt=0
    target port group asymmetric access state : 0x00 (active/optimized)
    T_SUP : 0, O_SUP : 0,LBD_SUP : 0, U_SUP : 1, S_SUP : 0, AN_SUP : 1, AO_SUP : 1
    status code : 0x02 (target port asym. state changed by implicit lu behaviour)
```

Mapping volume paths to hardware ports on PowerStore appliances

To determine where a path is connected on PowerStore, you can query the target port of each path using the **multipathd** command.

The following multipathd CLI example shows the **target WWPN** and **target WWNN** for each path. The target WWPN corresponds to the FC SCSI WWPN on the PowerStore appliance, and the target WWNN to the iSCSI IQN on the Ethernet port of the PowerStore. To locate the WWPN and WWNN of PowerStore ports in PowerStore Manager UI, go to **Hardware**, select the **Appliance**, and go to the **Ports** subtab.

```
# multipathd show paths format "%m %d %t %p %R %r %P %n %w"
multipath            dev  dm_st  pri host WWPN          target WWPN
protocol    target WWNN        uuid
metro-vol-001    sdan active 50  0x2001000e1efd0a6b 0x58ccf0984a600feb scsi:fcp
0x58ccf090ca600feb 368ccf0980076c2881eb573719c382287
metro-vol-001    sdat active 50  0x2001000e1efd0a6a 0x58ccf0984a610feb scsi:fcp
0x58ccf090ca600feb 368ccf0980076c2881eb573719c382287
metro-vol-001    sdab active 10  0x2001000e1efd0a6b 0x58ccf0904a600faa scsi:fcp
0x58ccf090ca600faa 368ccf0980076c2881eb573719c382287
metro-vol-001    sdal active 10  0x2001000e1efd0a6b 0x58ccf0984a600faa scsi:fcp
0x58ccf090ca600faa 368ccf0980076c2881eb573719c382287
metro-vol-001    sdb  active 10  0x2001000e1efd0a6a 0x58ccf0904a610faa scsi:fcp
0x58ccf090ca600faa 368ccf0980076c2881eb573719c382287
metro-vol-001    sdl  active 10  0x2001000e1efd0a6a 0x58ccf0984a610faa scsi:fcp
0x58ccf090ca600faa 368ccf0980076c2881eb573719c382287
metro-vol-001    sdn  active 10  0x2001000e1efd0a6a 0x58ccf0904a610feb scsi:fcp
0x58ccf090ca600feb 368ccf0980076c2881eb573719c382287
metro-vol-001    sdq  active 10  0x2001000e1efd0a6b 0x58ccf0904a600feb scsi:fcp
0x58ccf090ca600feb 368ccf0980076c2881eb573719c382287
```

To assist with monitoring and validating the ALUA state of the paths, a script, map-paths.sh, is available (in Appendix A) that consolidates and displays path information from the specified commands. Here is a sample output from the script showing the ALUA path states and the target port connections on PowerStore appliances. The script assigns friendly names to the target WWNN and target WWPN, to enhance the clarity of the report. For example, path sdan and sdat, which connect to Array 1 Node B, have an Active/Optimized ALUA state.

```
# ./map-paths.sh
metro-vol-001 sdan active 50 active/optimized  array1-nb-iom0-p0
metro-vol-001 sdat active 50 active/optimized  array1-nb-iom0-p1
metro-vol-001 sdab active 10 active/non optimized  array2-na-iom0-p0
metro-vol-001 sdal active 10 active/non optimized  array2-nb-iom0-p0
metro-vol-001 sdb active 10 active/non optimized  array2-na-iom0-p1
metro-vol-001 sdl active 10 active/non optimized  array2-nb-iom0-p1
metro-vol-001 sdn active 10 active/non optimized  array1-na-iom0-p1
metro-vol-001 sdq active 10 active/non optimized  array1-na-iom0-p0
```

Linux applications with Metro Volume

Metro Volume can be used with these popular Linux applications to meet the storage needs of other applications:

- Native Linux MPIO
- Regular Linux Logical Volume Manager on standalone host
- ext4 and xfs file systems on a standalone host
- Storage pool for KVM
- Clustered Linux Logical Volume Manager
- GFS2 cluster file system
- Red Hat High Availability Cluster
- SUSE Linux Enterprise High Availability

A sample use case for Metro Volume and Red Hat High-Availability Cluster

Metro Volume effectively protects hosts and applications in various failure scenarios. To illustrate its use in a cluster environment, the following example demonstrates Metro Volume's protection capabilities with a Red Hat High-Availability Cluster.

Red Hat High-Availability Cluster

Red Hat High-Availability Cluster is an Add-On that creates and configures high availability clusters. The cluster comprises various components, including Pacemaker, Corosync, Cluster Logical Volume Manager (CLVM), and GFS2 cluster file system.

This paper does not provide the procedure to install and configure the cluster. For information about Red Hat High-Availability Cluster, see the document Configuring and managing high availability clusters on the Red Hat documentation portal.

Figure 85 depicts the use of Metro Volume in a Red Hat High-Availability Cluster environment.

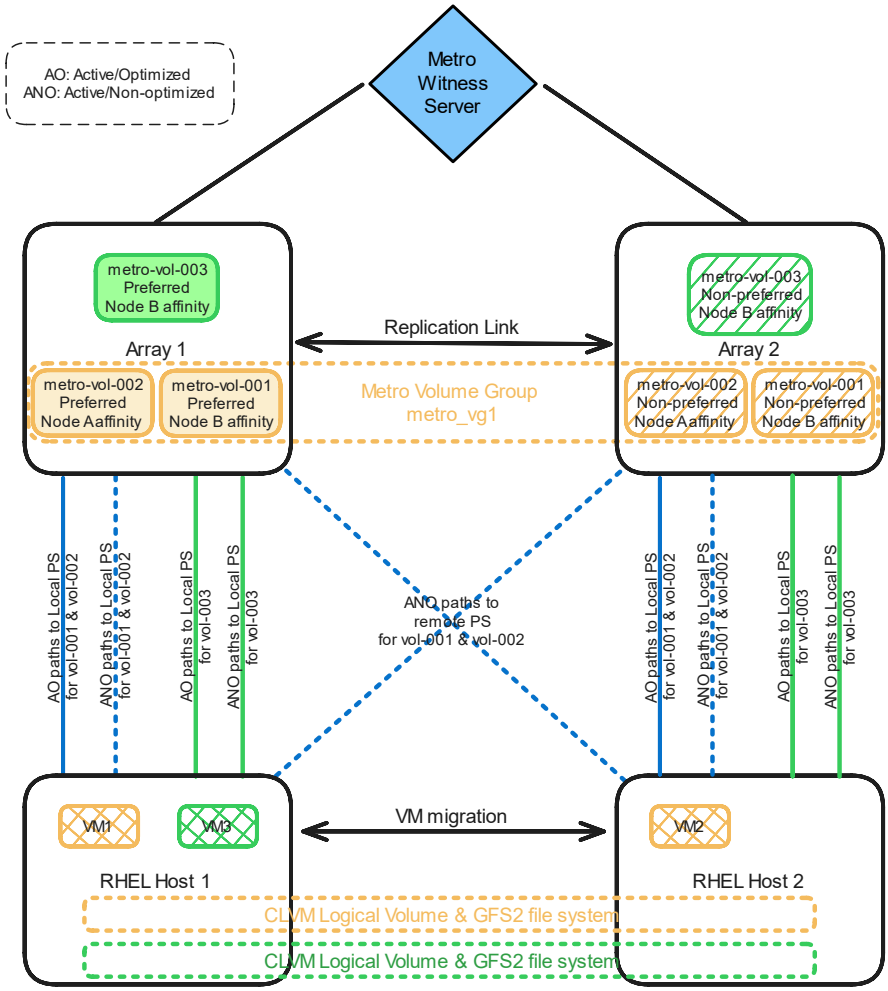


Figure 85. Metro Volumes with Red Hat High-Availability Cluster

Host connectivity of Host 1 and Host 2

Two PowerStore systems are configured to replicate to each other on a replication network. Array 1 and Host 1 are in the same location, while Array 2 and Host 2 are in a different location. Therefore, the hosts are configured with the following host connectivity.

Table 10. Linux Host connectivity configuration

PowerStore	Linux hosts	Host connectivity
Array 1	Host 1	Host is co-located with this system
Array 1	Host 2	Host is co-located with the remote system
Array 2	Host 2	Host is co-located with this system
Array 2	Host1	Host is co-located with the remote system

Metro volumes and volume group for Host 1 and Host 2

Three volumes are created on Array 1. metro-vol-001 and metro-vol-002 are members of a PowerStore volume group, metro_vg1. The Metro protection is then enabled for the volume group and the individual volume metro-vol-003. This creates the Metro replication

session for each volume and volume group, and mirror copies of the volume and volume group on Array 2. For more information about how to configure Metro Volume, see the Metro Volume operations section in this document.

Table 11. Preferred and Non-preferred role of Metro Volume

Metro volume/volume group	Array	Map to Linux host	Metro Role
metro_vg1 (metro-vol-001 and metro-vol-002)	Array 1	Host 1	Preferred
metro_vg1 (metro-vol-001 and metro-vol-002)	Array 2	Host 1	Non-Preferred
metro-vol-003	Array 1	Host 1	Preferred
metro-vol-003	Array 2	Host 2	Non-Preferred

Metro roles

Each side of a Metro Volume is designated as either Preferred or Non-preferred. The preferred volume is the one used to initiate the metro sync session. Preferred volumes are not restricted to the same array but can be distributed across arrays. Furthermore, Metro roles can be modified after their initial designation.

To view the Metro role and Metro Remote System for a volume in PowerStore Manager UI, go to **Storage** > **Volumes**. Use the column selection dropdown to add the Metro Role and Metro Remote System columns to the table.

In our example, for simplicity, because all metro sessions are initiated on Array 1, the volumes on Array 1 are designated as Preferred.

The role plays a crucial part in handling failure in split-brain situations through the Polarization mechanism. Without a witness, the Polarization mechanism always keeps the preferred volume online, while the non-preferred volume is taken offline. With a witness, the decision-making process gains intelligence, allowing it to handle a broader range of failure scenarios.

For more information about Polarization and Witness, see the Polarization section and the Witness section in this document.

Provision Metro volumes to Host 1 and Host 2

Metro Volume in uniform storage presentation

The Metro volume group metro_vg1 consists of metro-vol-001 and metro-vol-002, and is mapped to Host 1 from both Array 1 and Array 2. This forms a uniform storage presentation where Host 1 has access to both sides of the Metro Volume, shown as blue lines in Figure 85.

The Metro volume group metro_vg1 is also mapped to Host 2 from both Array 1 and Array 2, shown as blue lines in Figure 85.

The solid lines in the figure represent Active/Optimized paths to the arrays, while the dashed lines represent Active/Non-optimized paths.

Metro Volume in non-uniform storage presentation

The volume metro-vol-003 from Array 1 is mapped to Host 1, while metro-vol-003 from Array 2 is mapped to Host 2. This forms a non-uniform storage presentation where each

Linux host can only access one side of the Metro Volume, shown as green lines in Figure 85.

Configure consistent device names on clustered hosts

To ensure consistent device names across cluster hosts, a multipath alias is created for each Metro Volume based on its UUID. See the Device alias section in this document.

Verify paths configuration and ALUA states

After performing a SCSI scan on each clustered host, examine the paths and their ALUA states using the CLI or the script provided in the Examining path priority and ALUA state section in this document.

SCSI scan command: rescan-scsi-bus.sh -a

Run the map-paths.sh script on both Linux hosts to display the path states of each volume as follows.

On Host 1:

- Active/Optimized paths connect to Array 1.
- Active/Non-optimized paths connect to Array 1 and Array 2.

```
# Uniform presentation
metro-vol-001 sdan active 50 active/optimized  array1-nb-iom0-p0
metro-vol-001 sdat active 50 active/optimized  array1-nb-iom0-p1
metro-vol-001 sdab active 10 active/non optimized  array2-na-iom0-p0
metro-vol-001 sdal active 10 active/non optimized  array2-nb-iom0-p0
metro-vol-001 sdb active 10 active/non optimized  array2-na-iom0-p1
metro-vol-001 sdl active 10 active/non optimized  array2-nb-iom0-p1
metro-vol-001 sdn active 10 active/non optimized  array1-na-iom0-p1
metro-vol-001 sdq active 10 active/non optimized  array1-na-iom0-p0
# Uniform presentation
metro-vol-002 sdu active 50 active/optimized  array1-na-iom0-p1
metro-vol-002 sdz active 50 active/optimized  array1-na-iom0-p0
metro-vol-002 sdac active 10 active/non optimized  array2-na-iom0-p0
metro-vol-002 sdam active 10 active/non optimized  array2-nb-iom0-p0
metro-vol-002 sdas active 10 active/non optimized  array1-nb-iom0-p0
metro-vol-002 sdaz active 10 active/non optimized  array1-nb-iom0-p1
metro-vol-002 sdc active 10 active/non optimized  array2-na-iom0-p1
metro-vol-002 sdm active 10 active/non optimized  array2-nb-iom0-p1
# Non-uniform presentation
metro-vol-003 sdau active 50 active/optimized  array1-nb-iom0-p0
metro-vol-003 sdba active 50 active/optimized  array1-nb-iom0-p1
metro-vol-003 sdaa active 10 active/non optimized  array1-na-iom0-p0
metro-vol-003 sdw active 10 active/non optimized  array1-na-iom0-p1
```

On Host 2:

- Active/Optimized paths connect to Array 2.
- Active/Non-optimized paths connect to Array 2 and Array 1.

```
# Uniform presentation
metro-vol-001 sdan active 50 active/optimized  array2-nb-iom0-p0
```

```
metro-vol-001 sdn active 50 active/optimized  array2-nb-iom0-p1
metro-vol-001 sdab active 10 active/non optimized  array2-na-iom0-p0
metro-vol-001 sdap active 10 active/non optimized  array1-nb-iom0-p0
metro-vol-001 sdav active 10 active/non optimized  array1-nb-iom0-p1
metro-vol-001 sdb active 10 active/non optimized  array2-na-iom0-p1
metro-vol-001 sdp active 10 active/non optimized  array1-na-iom0-p1
metro-vol-001 sdr active 10 active/non optimized  array1-na-iom0-p0
# Uniform presentation
metro-vol-002 sdac active 50 active/optimized  array2-na-iom0-p0
metro-vol-002 sdc active 50 active/optimized  array2-na-iom0-p1
metro-vol-002 sdaa active 10 active/non optimized  array1-na-iom0-p0
metro-vol-002 sdam active 10 active/non optimized  array2-nb-iom0-p0
metro-vol-002 sdau active 10 active/non optimized  array1-nb-iom0-p0
metro-vol-002 sdba active 10 active/non optimized  array1-nb-iom0-p1
metro-vol-002 sdm active 10 active/non optimized  array2-nb-iom0-p1
metro-vol-002 sdy active 10 active/non optimized  array1-na-iom0-p1
# Non-uniform presentation
metro-vol-003 sdao active 50 active/optimized  array2-nb-iom0-p0
metro-vol-003 sdo active 50 active/optimized  array2-nb-iom0-p1
metro-vol-003 sdad active 10 active/non optimized  array2-na-iom0-p0
metro-vol-003 sdd active 10 active/non optimized  array2-na-iom0-p1
```

Clustered LVM Logical Volumes and GFS2 on Metro Volumes

Clustered LVM (CLVM) is an extension of the standard Logical Volume Manager. It allows multiple Linux hosts to manage a shared storage pool simultaneously. In our example, two CLVM logical volumes are created to manage the Metro Volumes on the clustered hosts.

- CLVM logical volume 1 (CLV1) is created on the metro_vg1, which consists of metro-vol-001 and metro-vol-002.
- CLVM logical volume 2 (CLV2) is created on metro-vol-003.

GFS2 is a Linux cluster file system for RHEL. It allows multiple hosts in a cluster to mount the file system simultaneously, enabling concurrent file access. GFS2 is commonly used with CLVM in a HA environment. A GFS2 file system is created on each CLVM logical volume and mounted on both Host 1 and Host 2.

Distribute application workloads

You can host application workloads on these GFS2 file systems. In our example, we distribute the data files of three KVM VMs across these file systems:

- VM1 is started on Host 1 and accesses its data files on CLV1.
- VM2 is started on Host 2 and accesses its data files on CLV1.
- VM3 is started on Host 1 and accesses its data files on CLV2.

Failure Scenario – Array failure

In this scenario, Array 1 experienced a complete failure.

Handling failure with a witness

Each metro session on Array 2 sends a fracture request to the witness. Because Array 1 is down, the requests from Array 2 are declared winners, allowing the volumes on Array 2 to stay online.

On Host1

- VM1 may experience a temporary I/O pause when the Active/Optimized paths of metro-vol-001 and metro-vol-002 connected to Array 1 fail. Multipath then promotes Active/Non-optimized paths to Active/Optimized, allowing VM1 I/Os to continue to Array 2.

For example, path sdal and sdl of metro-vol-001, connected to Array 2, change from Active/Non-optimized to Active/Optimized, and path sdan, sdat, sdn, sdq connected to Array 1 are removed. Similar changes happen to metro-vol-002.

```
metro-vol-001 sdal active 10 active/optimized  array2-nb-iom0-p0
metro-vol-001 sdl active 10 active/optimized  array2-nb-iom0-p1
metro-vol-001 sdab active 10 active/non optimized  array2-na-iom0-p0
metro-vol-001 sdb active 10 active/non optimized  array2-na-iom0-p0
```

- VM3 experiences irrecoverable I/O errors when all paths of metro-vol-003 fail. To continue operations, migrate VM3 to Host 2 where the data on metro-vol-003 is synchronized on Array 2, and is readily accessible.

On Host 2

- VM2 is not impacted because the Active/Optimized paths of metro-vol-001 and metro-vol-002 connected to Array 2 remain operational. Paths connected to Array 1 are removed, resulting in fewer paths in the multipath CLI display.
- For example, the path status of metro-vol-001 appears as follows. Similar changes happen to metro-vol-002.

```
metro-vol-001 sdan active 50 active/optimized  array2-nb-iom0-p0
metro-vol-001 sdn active 50 active/optimized  array2-nb-iom0-p1
metro-vol-001 sdab active 10 active/non optimized  array2-na-iom0-p0
metro-vol-001 sdb active 10 active/non optimized  array2-na-iom0-p1
```

- There is no change in the path status for metro-vol-003, because all paths connect to Array 2. When VM3 is restarted on Host 2, it will be able to access its data files on the volume.

Metro and Witness status on the surviving PowerStore appliance Array 2

- When Array 2 cannot communicate with Array 1, the Metro Status of Metro sessions becomes fractured.
- The Witness State of Metro sessions becomes Disengaged.

Recovery of Array 1

- The Witness State of Metro sessions revert to Engaged from Disengaged.
- The Metro sessions re-establish the connections and start the recovery of the metro volumes. The Metro Status displays Synchronizing, and changes to

Operating Normally (Active-Active) when the session is fully synchronized and healthy.

- During the synchronization, the paths return on the hosts. However, they remain in a failed and unavailable state. When they are fully synchronized, the paths become active and the ALUA states revert to the original state.

Handling failure without a witness

In the absence of a witness, the Polarization mechanism would have taken the non-preferred volumes on Array 2 offline, resulting in I/O errors and data unavailable for VM1 and VM2. It is possible to promote the non-preferred volumes and bring them online manually, but it requires an operator to intervene and execute the process manually, which prolongs the downtime of the applications.

Failure Scenario – Replication link failure

In this scenario, all replication links between the two arrays have failed. However, both arrays remain online and can communicate with the witness.

Handling failure with a witness

Each metro session on Array 1 and Array 2 sends a fracture request to the witness. By design, the fracture requests from the preferred volumes reach the witness before those from the non-preferred volume. Consequently, the preferred volumes win and stay online, while the non-preferred volumes are taken offline. For more information about this witness communication flow design, see Communication flow – fracture in the Witness section.

On Host 1

- VM1 is not impacted because the Active/Optimized paths of metro-vol-001 and metro-vol-002 connected to Array 1 remain online. Paths connected to Array 2 become failed and unavailable.
- For example, the path status of metro-vol-001 appears as follows. Similar changes happen to metro-vol-002.

```
metro-vol-001 sdaz active 50 active/optimized  array1-nb-iom0-p1
metro-vol-001 sdbd active 50 active/optimized  array1-nb-iom0-p0
metro-vol-001 sdad failed 10 unavailable  array2-na-iom0-p0
metro-vol-001 sdbf failed 10 unavailable  array2-nb-iom0-p1
metro-vol-001 sdbl failed 10 unavailable  array2-nb-iom0-p0
metro-vol-001 sdj failed 10 unavailable  array2-na-iom0-p1
metro-vol-001 sdc active 10 active/non optimized  array1-na-iom0-p0
metro-vol-001 sdh active 10 active/non optimized  array1-na-iom0-p1
```

- VM3 is not impacted because there is no path change for metro-vol-003, and all paths connected to Array 1 remain online.

On Host 2

- VM2 may experience a temporary I/O pause when the Active/Optimized paths of metro-vol-001 and metro-vol-002 connected to Array 2 failed. Multipath then promotes Active/Non-optimized paths to Active/Optimized, allowing VM2 I/Os to continue to Array 1.
- For example, the path status of metro-vol-001 appears as follows. Similar changes happen to metro-vol-002.

```
metro-vol-001 sdaw failed 50 unavailable  array2-nb-iom0-p0
metro-vol-001 sdm failed 50 unavailable  array2-nb-iom0-p1
metro-vol-001 sdaa failed 10 unavailable  array2-na-iom0-p0
metro-vol-001 sdt failed 10 unavailable  array2-na-iom0-p1
metro-vol-001 sdaq active 10 active/optimized  array1-nb-iom0-p0
metro-vol-001 sds active 10 active/optimized  array1-nb-iom0-p1
metro-vol-001 sdb active 10 active/non optimized  array1-na-iom0-p0
metro-vol-001 sdd active 10 active/non optimized  array1-na-iom0-p1
```

Metro and Witness status on Array 1 and Array 2

- When Array 1 and Array 2 cannot communicate with each other, the Metro Status of Metro sessions becomes fractured.
- The Witness State of Metro sessions becomes Disengaged.
- The Witness connection between Array 1 and the witness remains online (OK status).
- The Witness connection between Array 2 and the witness remains online (OK status).

Recovery of replication links

- The Metro sessions re-establish the connections and start the recovery of the metro volumes. The Metro Status changes from Fractured to Synchronizing, and eventually changes to Operating Normally (Active-Active) when the session is fully synchronized and healthy.
- The Witness State of Metro sessions revert to Engaged from Disengaged.
- During the synchronization, the paths return on the hosts. However, they remain in a failed and unavailable state. When they are fully synchronized, the paths become active and the ALUA states revert to the original state.

Handling failure without a witness

In the absence of a witness, the Polarization mechanism would have taken the non-preferred volumes on Array 2 offline. However, because all VMs run on the hosts that retain access to the preferred volumes, they continue to operate without interruption.

Other failure scenarios

Metro Volume and witness enhance the capability to manage a broad range of single and compound resource failure scenarios. The Failure scenarios section in this document is a valuable resource for understanding these possible scenarios and expected outcomes.

Conclusion

Summary

Dell PowerStore Metro Volume provides seamless data mobility for various proactive and reactive use cases. Metro Volume is easy to deploy and manage. However, you should perform careful planning to ensure that critical workloads that require high availability are polarized to the preferred Metro Volume. Also, take necessary steps to ensure that the preferred Metro Volume remains highly available to storage hosts with uniform storage presentation and vSphere Availability configured and enabled where possible.

Appendix A

map-paths.sh script

This sample script is provided as-is and requires minor modification to match your PowerStore appliances.

Update the target_arrays section in the script with your PowerStore appliance port WWPN and WWNN, and assign a descriptive label for each port.

```
#!/bin/bash

# Initialize variables to store the option values
long_format=false
search_string=""
extra_info=""

# Use getopts to process the command-line options
# -l display additional fields
# -s multipath device pattern
while getopts ":ls:" opt; do
  case $opt in
    l) long_format=true ;;
    s) search_string=$OPTARG ;;
    \?) echo "Invalid option: -$OPTARG" >&2; exit 1 ;;
    :) echo "Option -$OPTARG requires an argument." >&2; exit 1 ;;
  esac
done

declare -A target_arrays=(
  ["0x58ccf0904a600faa"]="array2-na-iom0-p0"
  ["0x58ccf0904a610faa"]="array2-na-iom0-p1"
  ["0x58ccf0984a600faa"]="array2-nb-iom0-p0"
  ["0x58ccf0984a610faa"]="array2-nb-iom0-p1"
  ["0x58ccf0904a600feb"]="array1-na-iom0-p0"
  ["0x58ccf0904a610feb"]="array1-na-iom0-p1"
  ["0x58ccf0984a600feb"]="array1-nb-iom0-p0"
  ["0x58ccf0984a610feb"]="array1-nb-iom0-p1"
  ["iqn.2015-10.com.dell:dellemc-powerstore-fnm72141021010-a-499c2a02"]="array2-na-bond0"
  ["iqn.2015-10.com.dell:dellemc-powerstore-fnm72141021010-b-771ac1b0"]="array2-nb-bond0"
  ["iqn.2015-10.com.dell:dellemc-powerstore-fnm72141021011-a-200105b5"]="array1-na-bond0"
  ["iqn.2015-10.com.dell:dellemc-powerstore-fnm72141021011-b-79d782c0"]="array1-nb-bond0"
)

while read multipath_name device status prio host_wwpn target_wwpn protocol iscsi_iqn
lun_wwn
do

    # Includes only relevant information for the protocol
    case "$protocol" in
        "scsi:fcp")
    array_port_wwn="$host_wwpn $target_wwpn"
```

```
        target_array=${target_arrays[$target_wwpn]:-"Unknown array"} ;;
        "scsi:iscsi")
    array_port_wwn="$iscsi_iqn"
        target_array=${target_arrays[$iscsi_iqn]:-"Unknown array"} ;;
        "scsi:unspec") break ;;
    esac

    # Query the target port group and the ALUA state of the path
    tpgid=$(sg_inq -p 0x83 /dev/$device |awk '/Target port group: / {print $NF}')
    aluastate=$(sg_rtpg -d /dev/$device |grep -EiA1 "target port group id.*$tpgid" |awk -
F"(" '/asymmetric access state/ {print $NF}' | tr -d '()' )


    if [[ "$long_format" == "true" ]];then
    extra_info="$protocol $array_port_wwn $lun_wwn"
    fi

    # Format color highlights for AO and Unavailable ALUA status.
    if [[ "$aluastate" =~ "active/optimized" ]]; then
    echo "$multipath_name $(tput setaf 2)$device $status $prio $aluastate $extra_info
$target_array$(tput sgr0)"
    elif [[ "$aluastate" =~ "unavailable" ]];then
    echo "$multipath_name $(tput setaf 1)$device $status $prio $aluastate $extra_info
$target_array$(tput sgr0)"
    else
    echo "$multipath_name $device $status $prio $aluastate $extra_info $target_array"
    fi

done < <(multipathd show paths format "%m %d %t %p %R %r %P %n %w" | grep -Eiv
"uuid|orphan") | grep -Ei "$search_string" | sort -k1,1 -k4,4nr -k2,2
```

References

Dell Technologies documentation

The following Dell Technologies documentation provides other information related to this document. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

- PowerStore Info Hub
- PowerStore Product Documentation and Videos
- PowerStore: Host Configuration Guide
- PowerStore: VMware vSphere Best Practices
- PowerStore: Protecting Your Data
- PowerStore: Replication Technologies
- PowerStore: Snapshots and Thin Clones
- PowerStore: Microsoft SQL Server Best Practices

VMware documentation

See the following links for related VMware resources:

- VMware Documentation
- VMware vSphere Metro Storage Cluster Recommended Practices

Red Hat documentation

The following Red Hat documentation provides information related to Red Hat Enterprise Linux and Red Hat High-Availability Cluster

- Red Hat Enterprise Linux 8 Documentation
- Configuring and managing high availability clusters