vSphere Metro Storage Cluster with ONTAP

Enterprise applications

NetApp
April 25, 2024

This PDF was generated from https://docs.netapp.com/us-en/ontap-apps-dbs/vmware/vmware_vmsc_overview.html on April 25, 2024. Always check docs.netapp.com for the latest.

Table of Contents

vSphere Metro Storage Cluster with ONTAP

vSphere Metro Storage Cluster with ONTAP

VMware's industry-leading vSphere hypervisor can be deployed as a stretched cluster referred to as a vSphere Metro Storage Cluster (vMSC).

vMSC solutions are supported with both NetApp® MetroCluster™ and SnapMirror active sync (formerly known as SnapMirror Business Continuity, or SMBC) and provide advanced business continuity if one or more failure domains suffer a total outage. The resilience to different modes of failure depends on which configuration options you choose.

Continuous Availability Solutions for vSphere Environments

ONTAP architecture is a flexible and scalable storage platform that provides SAN (FCP, iSCSI, and NVMe-oF) and NAS (NFS v3 and v4.1) services for datastores. The NetApp AFF, ASA, and FAS storage systems use the ONTAP operating system to offer additional protocols for guest storage access like S3 and SMB/CIFS.

NetApp MetroCluster uses NetApp's HA (controller failover or CFO) function to protect against controller failures. It also includes local SyncMirror technology, cluster failover on disaster (controller failover on demand or CFOD), hardware redundancy, and geographical separation to achieve high levels of availability. SyncMirror synchronously mirrors data across the two halves of the MetroCluster configuration by writing data to two plexes: the local plex (on the local shelf) actively serving data and the remote plex (on the remote shelf) normally not serving data. Hardware redundancy is put in place for all MetroCluster components such as controllers, storage, cables, switches (used with fabric MetroCluster), and adapters.

NetApp SnapMirror active sync provides datastore-granular protection with FCP and iSCSI SAN protocols, allowing you to selectively protect only high-priority workloads. It offers active-active access to both local and remote sites, unlike NetApp MetroCluster which is an active-standby solution. At present, active sync is an asymmetric solution where one side is preferred over the other, providing better performance. This is achieved using ALUA (Asymmetric Logical Unit Access) functionality which automatically informs the ESXi host which controllers to prefer. However, NetApp has announced that active sync will soon enable fully symmetric access.

To create a VMware HA/DRS cluster across two sites, ESXi hosts are used and managed by a vCenter Server Appliance (VCSA). The vSphere management, vMotion®, and virtual machine networks are connected through a redundant network between the two sites. The vCenter Server managing the HA/DRS cluster can connect to the ESXi hosts at both sites and should be configured using vCenter HA.

Refer to How Do You Create and Configure Clusters in the vSphere Client to configure vCenter HA.

You should also refer to VMware vSphere Metro Storage Cluster Recommended Practices.

What is vSphere Metro Storage Cluster?

vSphere Metro Storage Cluster (vMSC) is a certified configuration that protects virtual machines (VMs) and containers against failures. This is achieved by using stretched storage concepts along with clusters of ESXi hosts, which are distributed across different failure domains such as racks, buildings, campuses, or even cities. The NetApp MetroCluster and SnapMirror active sync storage technologies are used to provide RPO=0 or near RPO=0 protection respectively to the host clusters. The vMSC configuration is designed to ensure that data is always available even if a complete physical or logical "site" fails. A storage device that is part of the vMSC configuration must be certified after undergoing a successful vMSC certification process. All the supported

storage devices can be found in the VMware Storage Compatibility Guide.

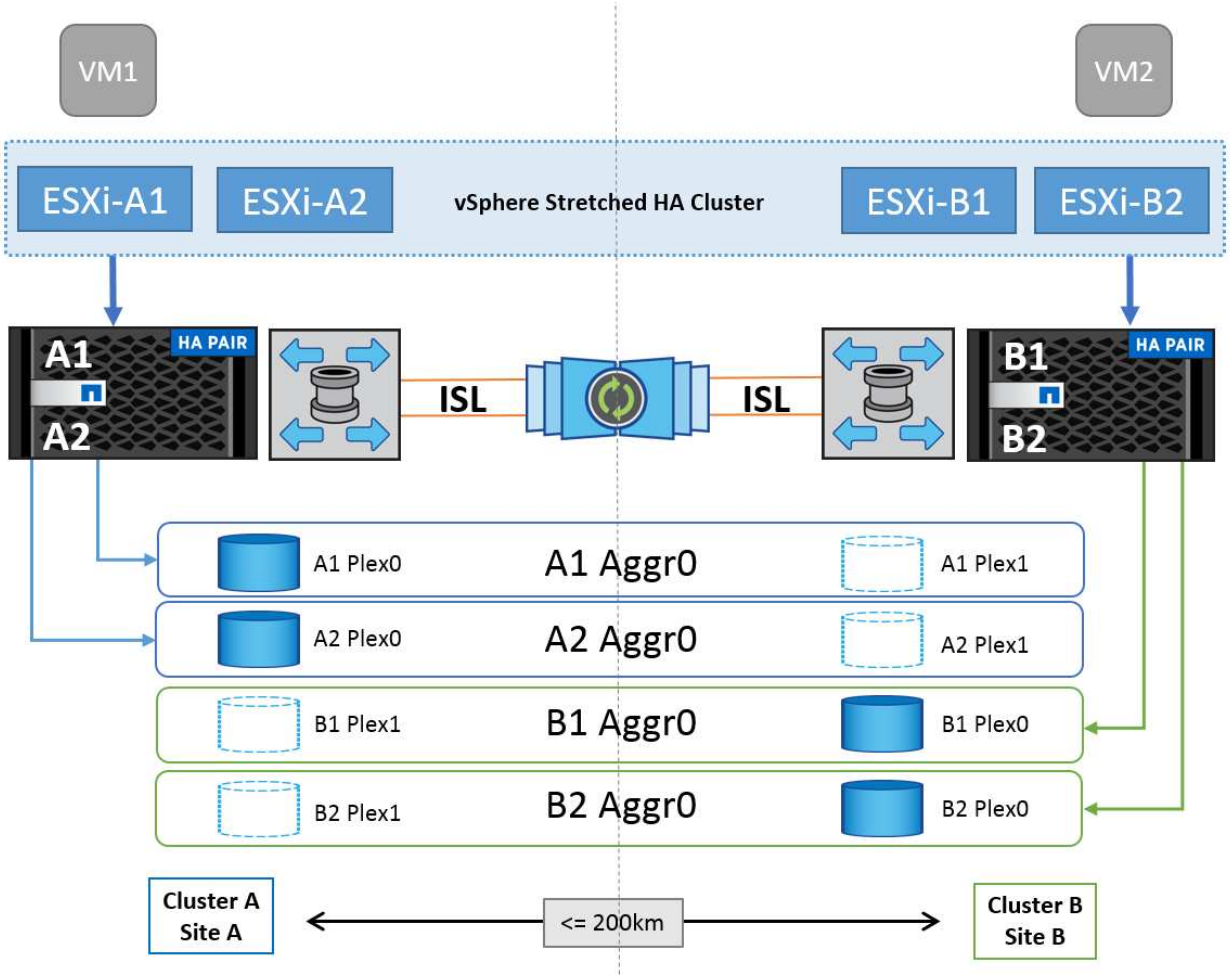
If you want more information about the design guidelines for vSphere Metro Storage Cluster, you can refer to the following documentation:

- VMware vSphere support with NetApp MetroCluster
- VMware vSphere support with NetApp SnapMirror Business Continuity (now known as SnapMirror active sync)

Depending on the latency considerations, NetApp MetroCluster can be deployed in two different configurations for use with vSphere:

- Stretch MetroCluster
- Fabric MetroCluster

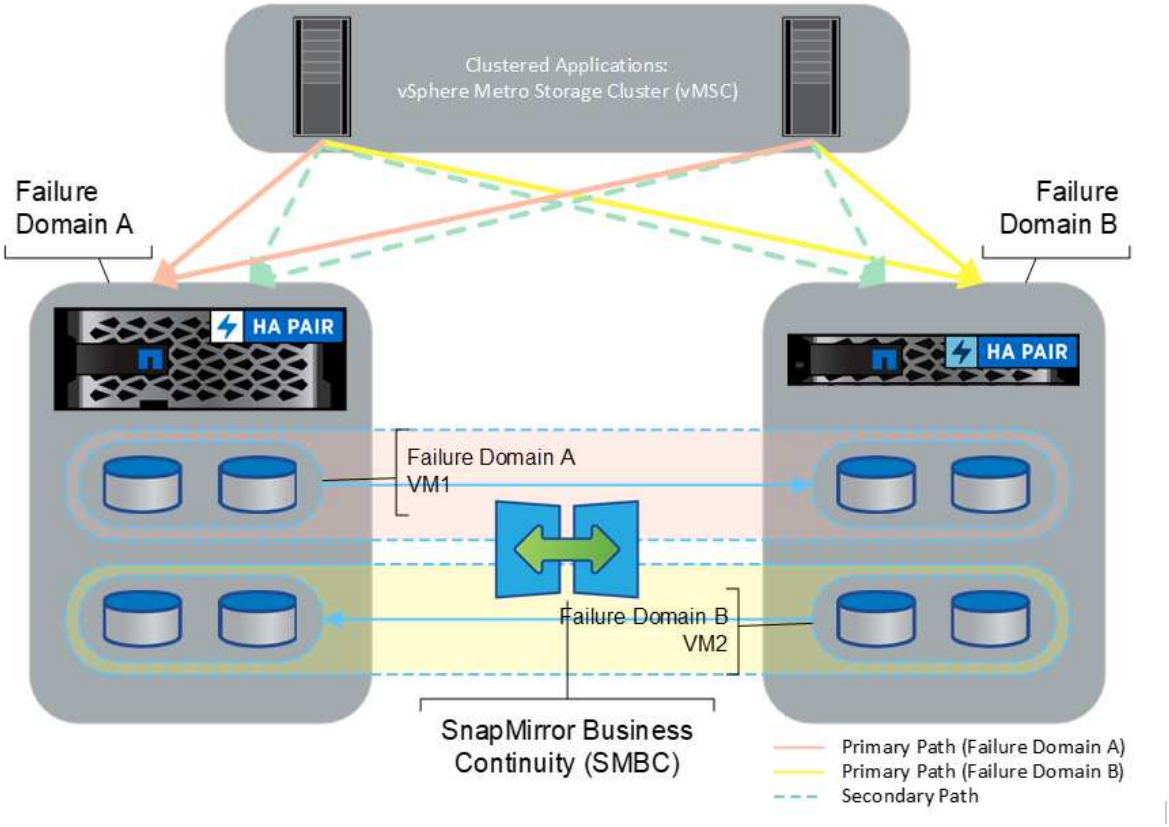
The following illustrates a high-level topology diagram of stretch MetroCluster.



Refer to MetroCluster documentation for specific design and deployment information for MetroCluster.

SnapMirror active sync can also be deployed in two different ways.

- Asymmetric
- Symmetric (private preview in ONTAP 9.14.1)



Refer to NetApp Docs for specific design and deployment information for SnapMirror active sync.

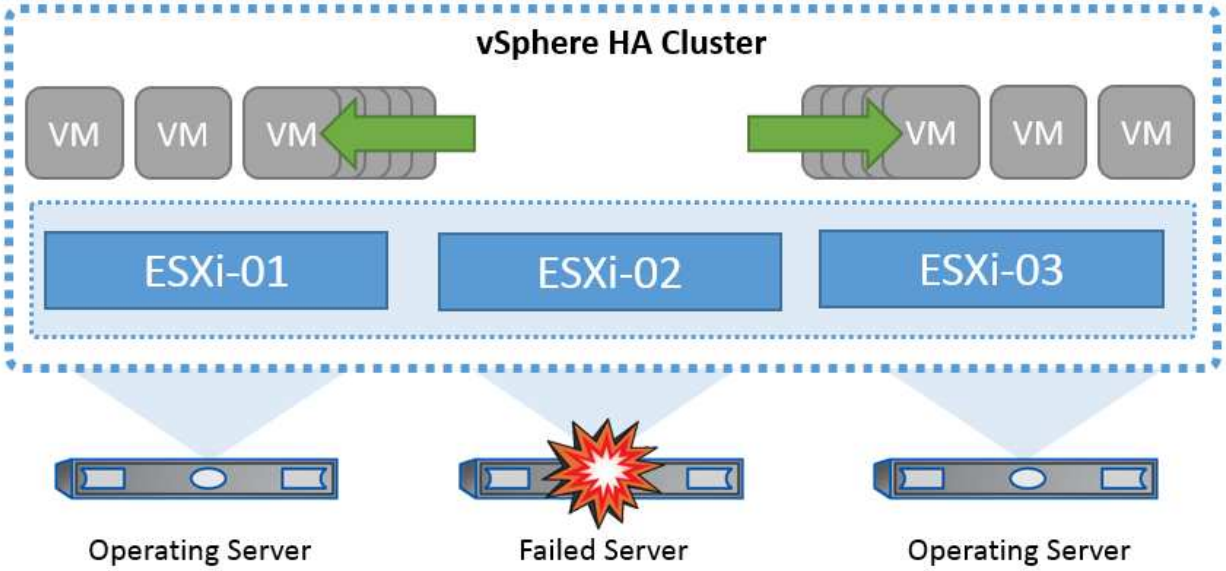
VMware vSphere Solution Overview

The vCenter Server Appliance (VCSA) is the powerful centralized management system and single pane of glass for vSphere that enables administrators to effectively operate ESXi clusters. It facilitates key functions such as VM provisioning, vMotion operation, High Availability (HA), Distributed Resource Scheduler (DRS), Tanzu Kubernetes Grid, and more. It is an essential component in VMware cloud environments and should be designed with service availability in mind.

vSphere High Availability

VMware's cluster technology groups ESXi servers into pools of shared resources for virtual machines and provides vSphere High Availability (HA). vSphere HA provides easy-to-use, high availability for applications running in virtual machines. When the HA feature is enabled on the cluster, each ESXi server maintains communication with other hosts so that if any ESXi host becomes unresponsive or isolated, the HA cluster can negotiate the recovery of the virtual machines that were running on that ESXi host among surviving hosts in the cluster. In the event of a guest operating system failure, vSphere HA restarts the affected virtual machine on the same physical server. vSphere HA makes it possible to reduce planned downtime, prevent unplanned downtime, and rapidly recover from outages.

vSphere HA cluster recovering VMs from failed server.



It's important to understand that VMware vSphere has no knowledge of NetApp MetroCluster or SnapMirror active sync and sees all ESXi hosts in the vSphere cluster as eligible hosts for HA cluster operations depending on host and VM group affinity configurations.

Host Failure Detection

As soon as the HA cluster is created, all hosts in the cluster participate in election, and one of the hosts becomes a master. Each slave performs network heartbeat to the master, and the master in turn performs network heartbeat on all slave hosts. The master host of a vSphere HA cluster is responsible for detecting the failure of slave hosts.

Depending on the type of failure detected, the virtual machines running on the hosts might need to be failed over.

In a vSphere HA cluster, three types of host failure are detected:

- Failure - A host stops functioning.
- Isolation - A host becomes network isolated.
- Partition - A host loses network connectivity with the master host.

The master host monitors the slave hosts in the cluster. This communication is done through the exchange of network heartbeats every second. When the master host stops receiving these heartbeats from a slave host, it checks for host liveness before declaring the host to have failed. The liveness check that the master host performs is to determine whether the slave host is exchanging heartbeats with one of the datastores. Also, the master host checks whether the host responds to ICMP pings sent to its management IP addresses to detect whether it is merely isolated from its master node or completely isolated from the network. It does this by pinging the default gateway. One or more isolation addresses can be specified manually to enhance the reliability of isolation validation.

Best Practice

NetApp recommends specifying a minimum of two additional isolation addresses, and that each of these addresses be site-local. This will enhance the reliability of isolation validation.

Host Isolation Response

Isolation Response is a setting in vSphere HA that determines the action triggered on Virtual Machines when a host in a vSphere HA cluster loses its management network connections but continues to run. There are three options for this setting, “Disabled”, “Shut Down and Restart VMs,” and “Power Off and Restart VMs.”

“Shut Down” is better than “Power Off”, which does not flush most recent changes to disk or commit transactions. If virtual machines have not shut down in 300 seconds they are powered off. To change the wait time, use the advanced option das.isolationshutdowntimeout.

Before HA initiates the isolation response, it first checks to see if the vSphere HA master agent owns the datastore that contains the VM config files. If not, then the host will not trigger the isolation response, because there is no master to restart the VMs. The host will periodically check the datastore state to determine if it is claimed by a vSphere HA agent that holds the master role.

Best Practice

NetApp recommends setting the “Host Isolation Response” to Disabled.

A split-brain condition can occur if a host becomes isolated or partitioned from the vSphere HA master host and the master is unable to communicate via heartbeat datastores or by ping. The master declares the isolated host dead and restarts the VMs on other hosts in the cluster. A split-brain condition now exists because there are two instances of the virtual machine running, only one of which can read or write the virtual disks. Split-brain conditions can now be avoided by configuring VM Component Protection (VMCP).

VM Component Protection (VMCP)

One of the feature enhancements in vSphere 6, relevant to HA, is VMCP. VMCP provides enhanced protection from All Paths Down (APD) and Permanent Device Loss (PDL) conditions for block (FC, iSCSI, FCoE) and file storage (NFS).

Permanent Device Loss (PDL)

PDL is a condition that occurs when a storage device permanently fails or is administratively removed and is not expected to return. The NetApp storage array issues a SCSI Sense code to ESXi declaring that the device is permanently lost. In the Failure Conditions and VM Response section of vSphere HA, you can configure what the response should be after a PDL condition is detected.

Best Practice

NetApp recommends setting the “Response for Datastore with PDL” to “**Power off and restart VMs**”. When this condition is detected a VM will be restarted instantly on a healthy host within the vSphere HA cluster.

All Paths Down (APD)

APD is a condition that occurs when a storage device becomes inaccessible to the host and no paths to the array are available. ESXi considers this a temporary problem with the device and is expecting it to become available again.

When an APD condition is detected, a timer is started. After 140 seconds, the APD condition is officially declared, and the device is marked as APD time out. When the 140 seconds have passed, HA will start counting the number of minutes specified in the Delay for VM Failover APD. When the specified time has passed, HA will restart the impacted virtual machines. You can configure VMCP to respond differently if desired (Disabled, Issue Events, or Power Off and Restart VMs).

Best Practice

NetApp recommends configuring the “Response for Datastore with APD” to “**Power off and restart VMs (conservative)**”.

Conservative refers to the likelihood of HA being able to restart VMs. When set to Conservative, HA will only restart the VM that is impacted by the APD if it knows another host can restart it. In the case of Aggressive, HA will try to restart the VM even if it doesn’t know the state of the other hosts. This can result in VMs not being restarted if there is no host with access to the datastore it is located on.

If the APD status is resolved and access to the storage is restored before the time-out has passed, HA will not unnecessarily restart the virtual machine unless you explicitly configure it to do so. If a response is desired even when the environment has recovered from the APD condition, then Response for APD Recovery After APD Timeout should be configured to Reset VMs.

Best Practice

NetApp recommends configuring Response for APD Recovery After APD Timeout to Disabled.

VMware DRS Implementation for NetApp MetroCluster

VMware DRS is a feature that aggregates the host resources in a cluster and is primarily used to load balance within a cluster in a virtual infrastructure. VMware DRS primarily calculates the CPU and memory resources to perform load balancing in a cluster. Because vSphere is unaware of stretched clustering, it considers all hosts in both sites when load balancing. To avoid cross-site traffic, NetApp recommends configuring DRS affinity rules to manage a logical separation of VMs. This will ensure that unless there is a complete site failure, HA and DRS will only use local hosts.

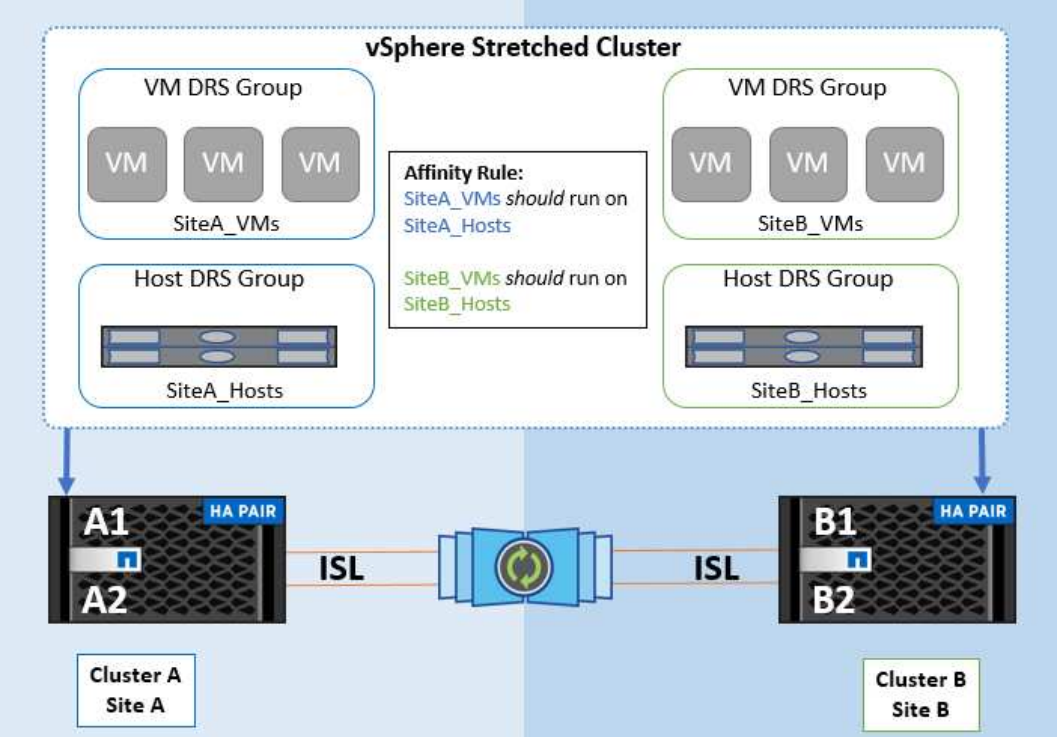
If you create a DRS affinity rule for your cluster, you can specify how vSphere applies that rule during a virtual machine failover.

There are two types of rules you can specify vSphere HA failover behavior:

- VM anti-affinity rules force specified virtual machines to remain apart during failover actions.
- VM host affinity rules place specified virtual machines on a particular host or a member of a defined group of hosts during failover actions.

Using VM host affinity rules in VMware DRS, one can have a logical separation between site A and site B so that the VM runs on the host at the same site as the array that is configured as the primary read/write controller for a given datastore. Also, VM host affinity rules enable virtual machines to stay local to the storage, which in turn ascertains the virtual machine connection in case of network failures between the sites.

The following is an example of VM host groups and affinity rules.



Best Practice

NetApp recommends implementing “should” rules instead of “must” rules because they are violated by vSphere HA in the case of a failure. Using “must” rules could potentially lead to service outages.

Availability of services should always prevail over performance. In the scenario where a full data center fails, “must” rules must choose hosts from the VM host affinity group, and when the data center is unavailable, the virtual machines will not restart.

VMware Storage DRS Implementation with NetApp MetroCluster

The VMware Storage DRS feature enables the aggregation of datastores into a single unit and balances virtual machine disks when storage I/O control thresholds are exceeded.

Storage I/O control is enabled by default on Storage DRS–enabled DRS clusters. Storage I/O control allows an administrator to control the amount of storage I/O that is allocated to virtual machines during periods of I/O congestion, which enables more important virtual machines to have preference over less important virtual machines for I/O resource allocation.

Storage DRS uses Storage vMotion to migrate the virtual machines to different datastores within a datastore cluster. In a NetApp MetroCluster environment, a virtual machine migration needs to be controlled within the datastores of that site. For example, virtual machine A, running on a host at site A, should ideally migrate within the datastores of the SVM at site A. If it fails to do so, the virtual machine will continue to operate but with degraded performance, since the virtual disk read/write will be from site B through inter-site links.

Best Practice

NetApp recommends creating datastore clusters with respect to storage site affinity; that is, datastores with site affinity for site A should not be mixed with datastore clusters with datastores with site affinity for site B.

Whenever a virtual machine is newly provisioned or migrated using Storage vMotion, NetApp recommends that all the VMware DRS rules specific to those virtual machines be manually updated, accordingly. This will

ascertain the virtual machine affinity at the site level for both host and datastore and thus reduce the network and storage overhead.

vMSC Design and Implementation Guidelines

This document outlines the design and implementation guidlines for vMSC with ONTAP storage systems.

NetApp Storage Configuration

Setup instructions for NetApp MetroCluster (referred to as an MCC configuration) are available at MetroCluster Documentation. Instructions for SnapMirror active sync are also available at SnapMirror Business Continuity overview.

Once you have configured MetroCluster, administering it is like managing a traditional ONTAP environment. You can set up Storage Virtual Machines (SVMs) using various tools like the Command Line Interface (CLI), System Manager, or Ansible. Once the SVMs are configured, create Logical Interfaces (LIFs), volumes, and Logical Unit Numbers (LUNs) on the cluster that will be used for normal operations. These objects will automatically be replicated to the other cluster using the cluster peering network.

If not using MetroCluster, you can use SnapMirror active sync which provides datastore-granular protection and active-active access across multiple ONTAP clusters in different failure domains. SnapMirror active sync uses consistency groups to ensure write-order consistency among one or more datastores and you can create multiple consistency groups depending on your application and datastore requirements. Consistency groups are especially useful for applications that require data synchronization between multiple datastores. SnapMirror active sync also supports Raw Device Mappings (RDMs) and guest-connected storage with in-guest iSCSI initiators. You can learn more about consistency groups at Consistency groups overview.

There is some difference in managing a vMSC configuration with SnapMirror active sync when compared to a MetroCluster. First, this is a SAN-only configuration, no NFS datastores can be protected with SnapMirror active sync. Second, you must map both copies of the LUNs to your ESXi hosts for them to access the replicated datastores in both failure domains.

VMware vSphere HA

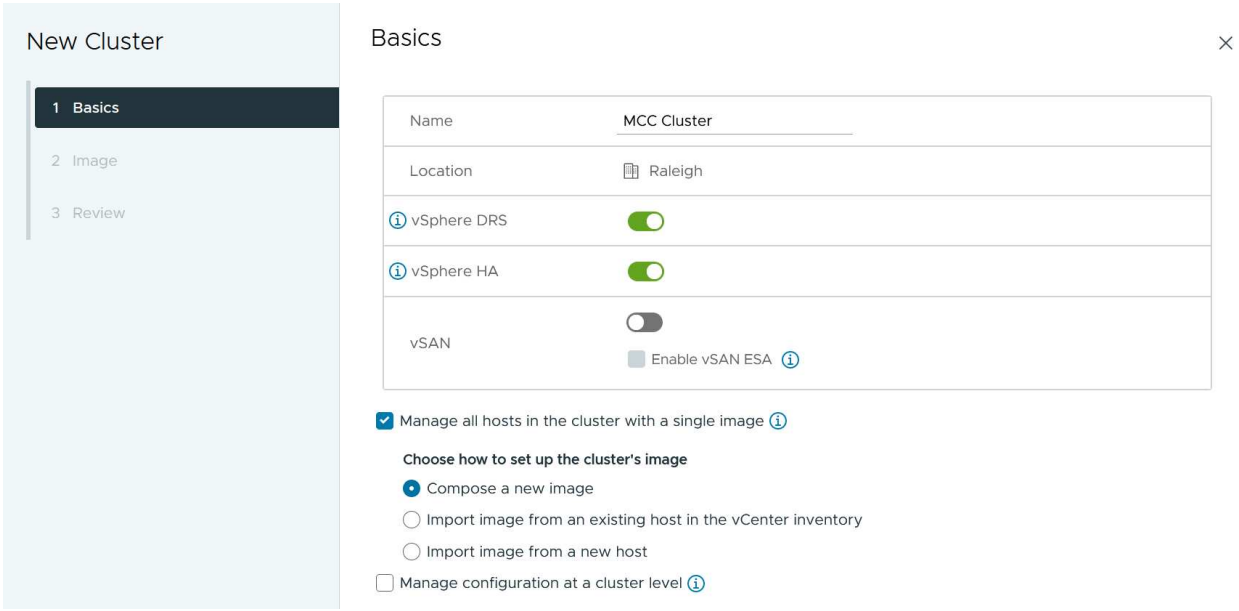
Create a vSphere HA Cluster

Creating a vSphere HA cluster is a multi-step process that is fully documented at How Do You Create and Configure Clusters in the vSphere Client on docs.vmware.com. In short, you must first create an empty cluster, and then, using vCenter, you must add hosts and specify the cluster's vSphere HA and other settings.

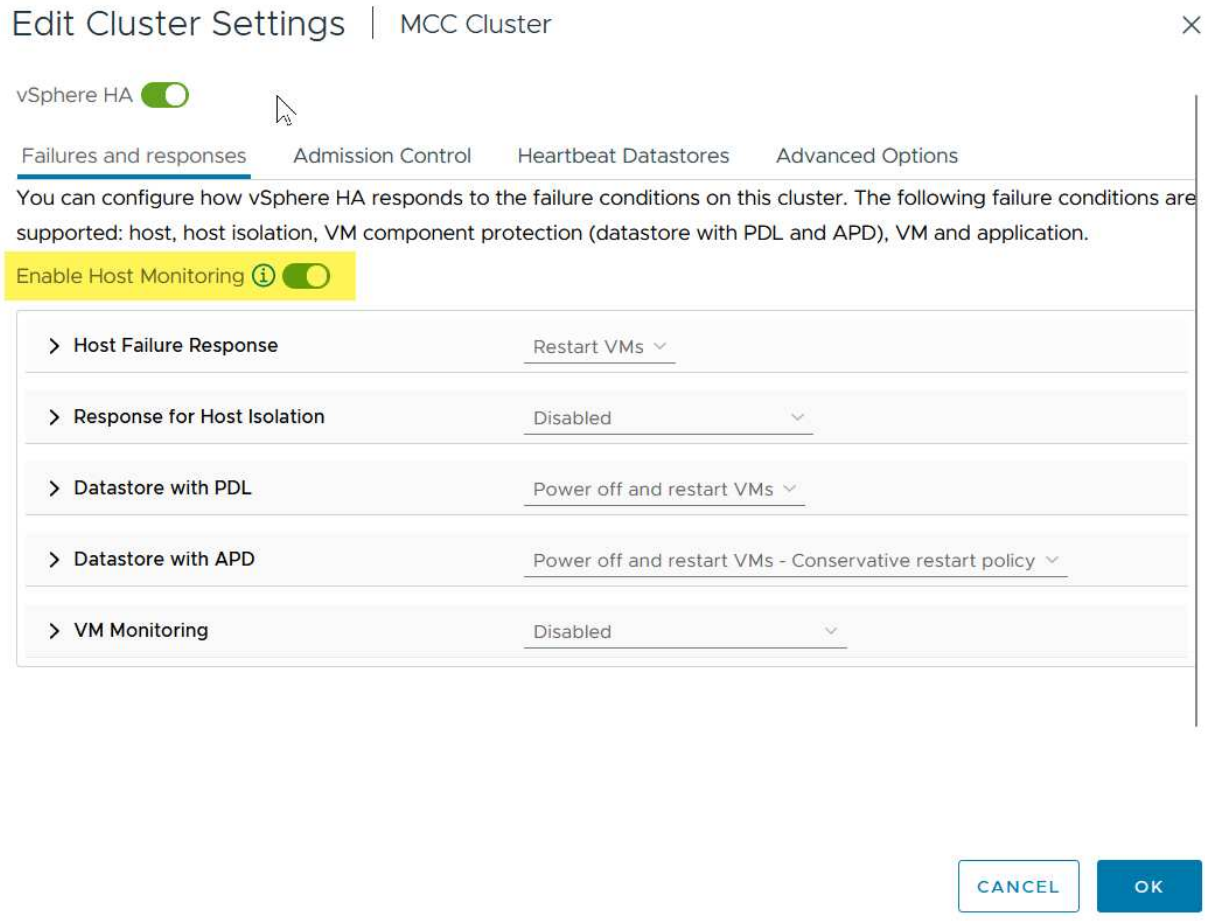
Note: Nothing in this document supersedes VMware vSphere Metro Storage Cluster Recommended Practices

To configure an HA cluster, complete the following steps:

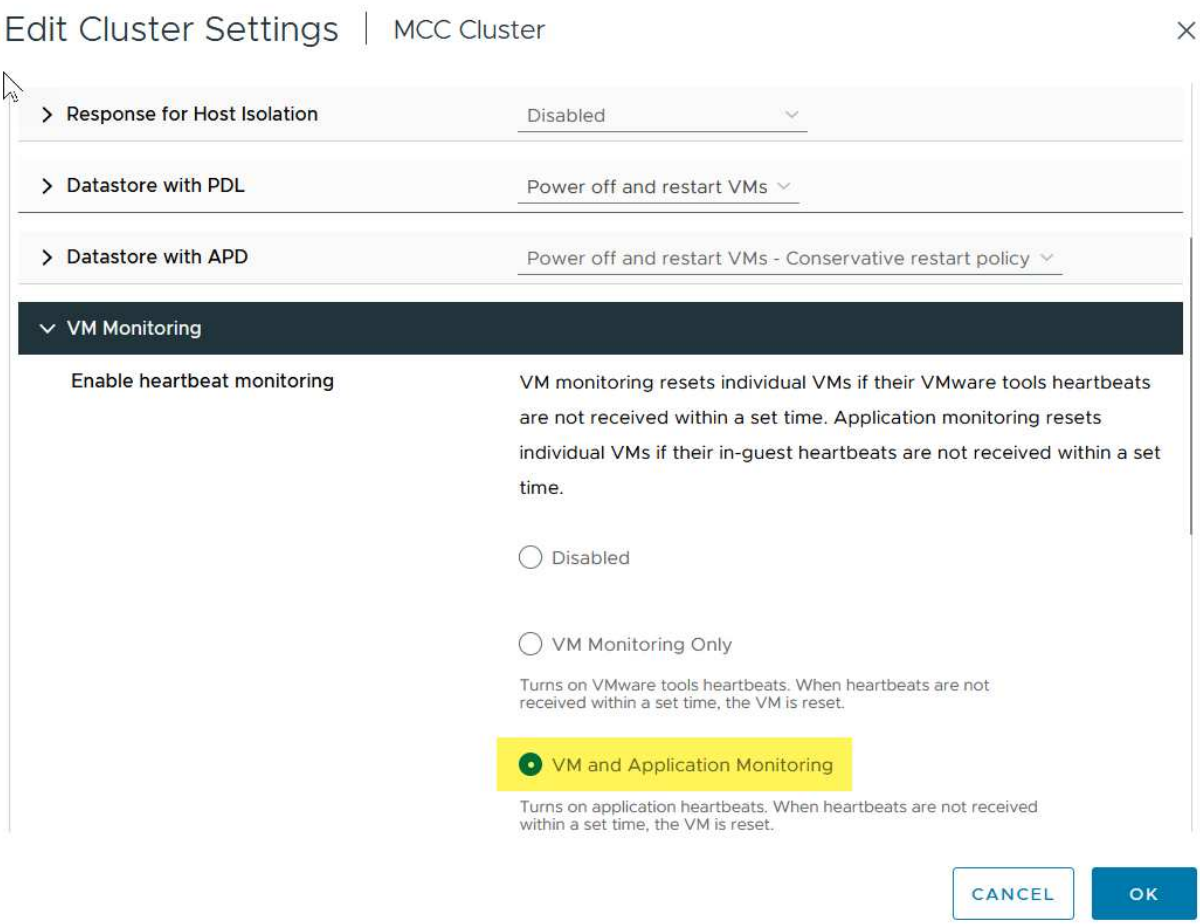
1. Connect to the vCenter UI.
2. In Hosts and Clusters, browse to the data center where you want to create your HA cluster.
3. Right-click the data center object and select New Cluster. Under basics ensure you have enabled vSphere DRS and vSphere HA. Complete the wizard.



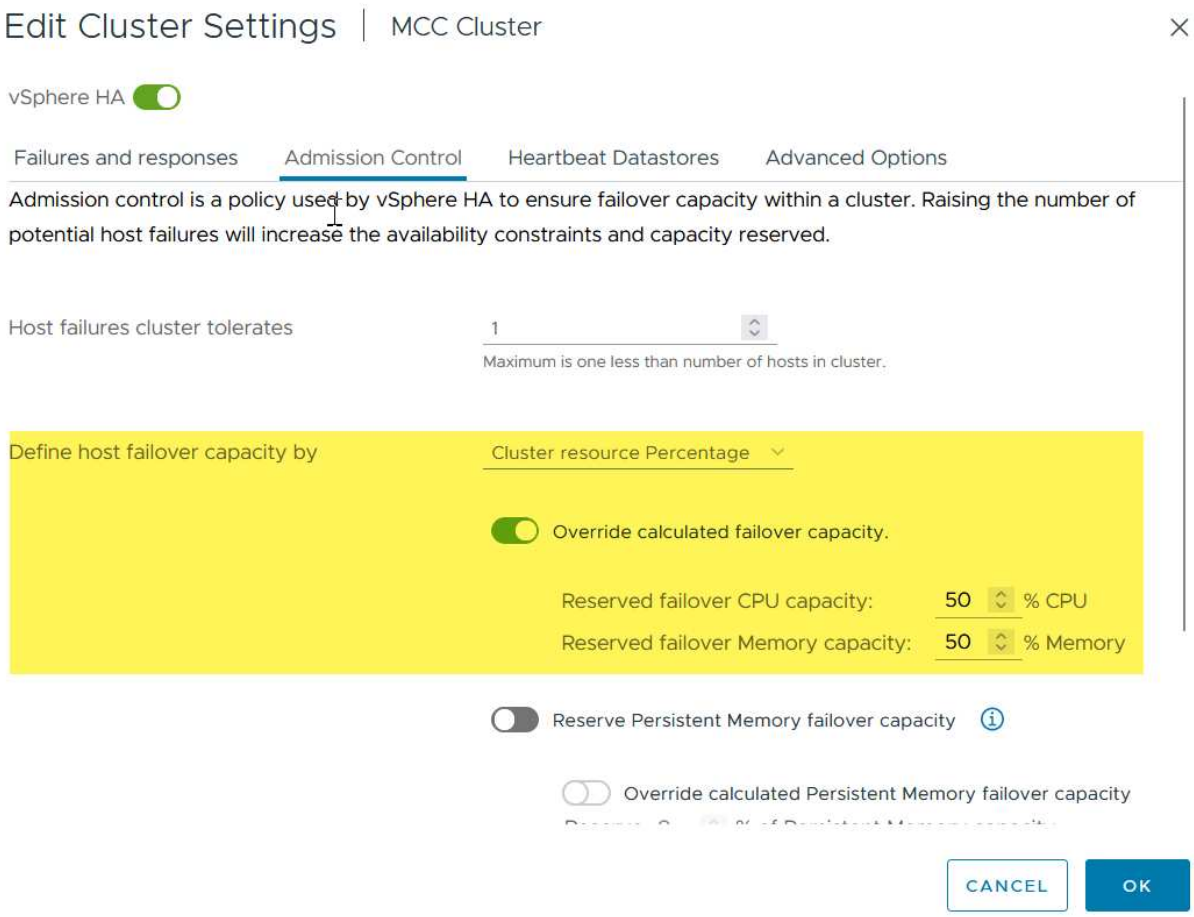
4. Select the cluster and go to the configure tab. Select vSphere HA and click edit.
5. Under Host Monitoring, select the Enable Host Monitoring option.



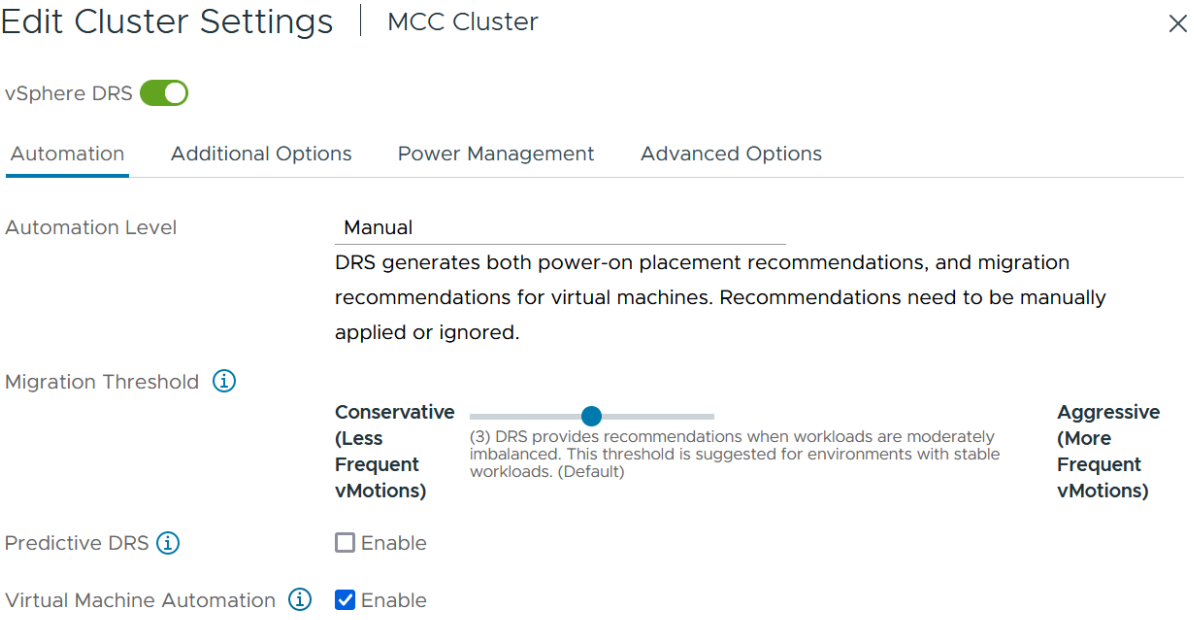
6. While still on the Failures and Responses tab, Under VM Monitoring, select the VM Monitoring Only option or VM and Application Monitoring option.



7. Under Admission Control, set the HA admission control option to cluster resource reserve; use 50% CPU/MEM.



8. Click “OK”.
9. Select DRS and click EDIT.
10. Set the automation level to manual unless required by your applications.



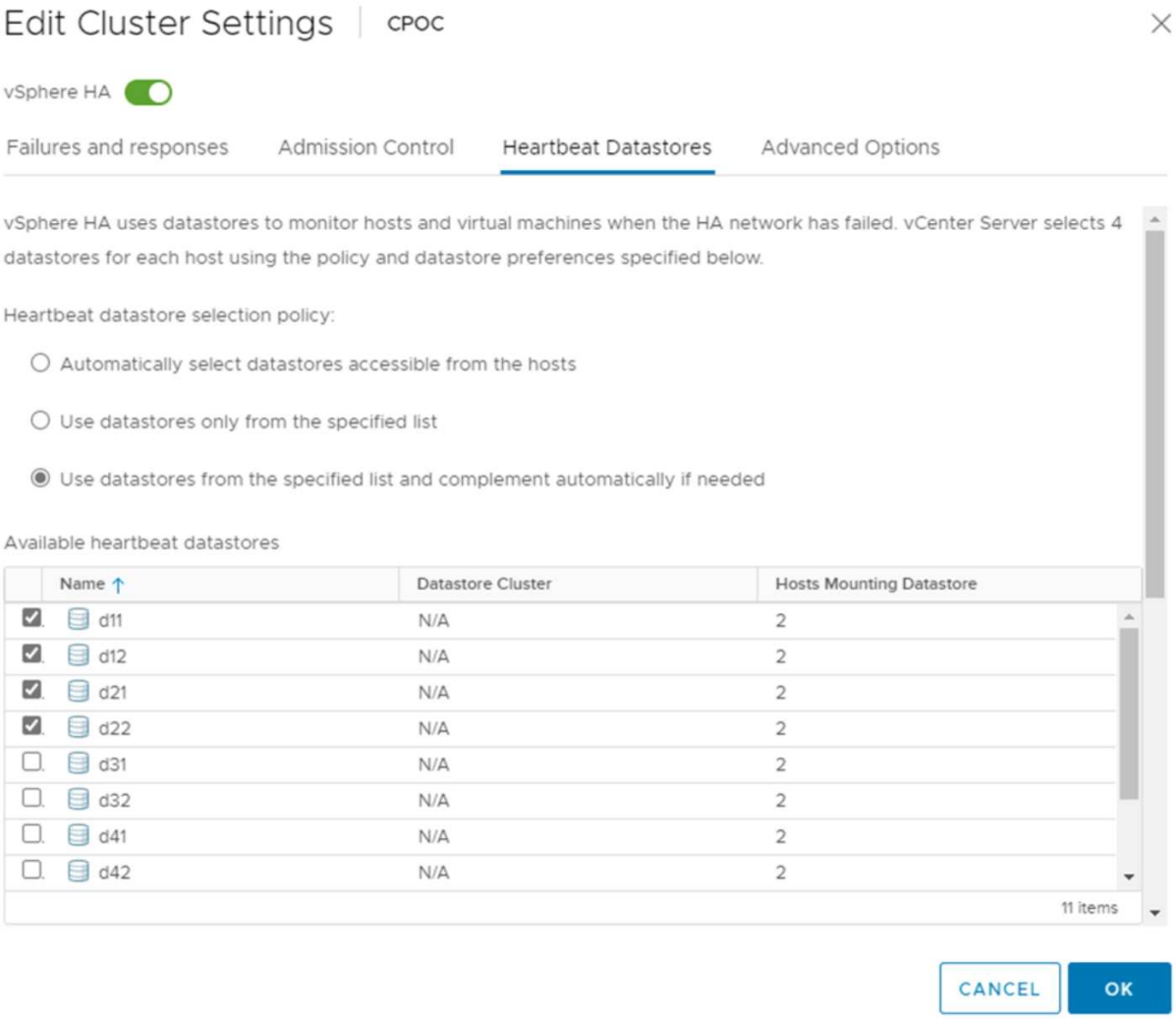
11. Enable VM Component Protection, refer to docs.vmware.com.
12. The following additional vSphere HA settings are recommended for vMSC with MCC:

Failure	Response
Host failure	Restart VMs
Host isolation	Disabled
Datastore with Permanent Device Loss (PDL)	Power off and restart VMs
Datastore with All paths Down (APD)	Power off and restart VMs
Guest not heartbeating	Reset VMs
VM restart policy	Determined by the importance of the VM
Response for host isolation	Shut down and restart VMs
Response for datastore with PDL	Power off and restart VMs
Response for datastore with APD	Power off and restart VMs (conservative)
Delay for VM failover for APD	3 minutes
Response for APD recovery with APD timeout	Disabled
VM monitoring sensitivity	Preset high

Configure Datastores for Heartbeating

vSphere HA uses datastores to monitor hosts and virtual machines when the management network has failed. You can configure how vCenter selects heartbeat datastores. To configure datastores for heartbeating, complete the following steps:

1. In the Datastore Heartbeating section, select Use Datastores from the Specified List and Compliment Automatically if Needed.
2. Select the datastores you want vCenter to use from both sites and press OK.

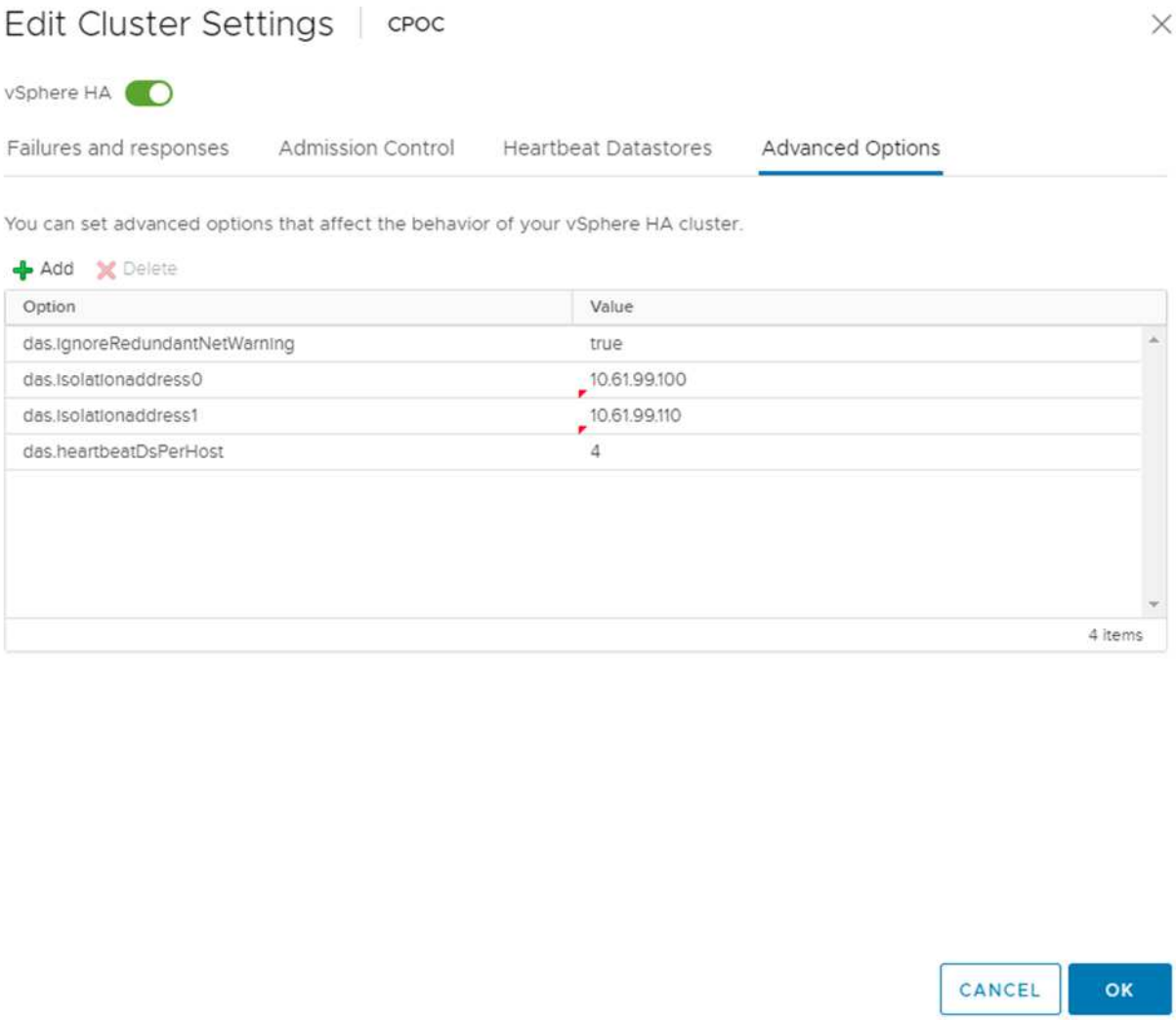


Configure Advanced Options

Host Failure Detection

Isolation events occur when hosts within an HA cluster lose connectivity to either the network or other hosts in the cluster. By default, vSphere HA will use the default gateway for its management network as the default isolation address. However, you can specify additional isolation addresses for the host to ping to determine whether an isolation response should be triggered. Add two isolation IPs that can ping, one per site. Do not use the gateway IP. The vSphere HA advanced setting used is das.isolationaddress. You can use ONTAP or Mediator IP addresses for this purpose.

Refer to core.vmware.com for more information.



Adding an advanced setting called das.heartbeatDsPerHost can increase the number of heartbeat datastores. Use four heartbeat datastores (HB DSs)—two per site. Use the “Select from List but Compliment” option. This is needed because if one site fails, you still need two HB DSs. However, those don’t have to be protected with MCC or SnapMirror active sync.

Refer to core.vmware.com for more information.

VMware DRS Affinity for NetApp MetroCluster

In this section, we create DRS groups for VMs and hosts for each site\cluster in the MetroCluster environment. Then we configure VM\Host rules to align VM host affinity with local storage resources. For example, site A VMs belong to VM group sitea_vms and site A hosts belong to host group sitea_hosts. Next, in VM\Host Rules, we state that sitea_vms should run on hosts in sitea_hosts.

Best Practice

- NetApp highly recommends the specification **Should Run on Hosts in Group** rather than the specification **Must Run on Hosts in Group**. In the event of a site A host failure, the VMs of site A need to be restarted on hosts at site B through vSphere HA, but the latter specification does not allow HA to restart VMs on site B because it’s a hard rule. The former specification is a soft rule and will be violated in the event of HA, thus enabling availability rather than performance.

Note: You can create an event-based alarm that is triggered when a virtual machine violates a VM-Host affinity rule. In the vSphere Client, add a new alarm for the virtual machine and select “VM is violating VM-Host Affinity Rule” as the event trigger. For more information about creating and editing alarms, refer to vSphere Monitoring and Performance documentation.

Create DRS Host Groups

To create DRS host groups specific to site A and site B, complete the following steps:

1. In the vSphere web client, right-click the cluster in the inventory and select Settings.
2. Click VM\Host Groups.
3. Click Add.
4. Type the name for the group (for instance, sitea_hosts).
5. From the Type menu, select Host Group.
6. Click Add and select the desired hosts from site A and click OK.
7. Repeat these steps to add another host group for site B.
8. Click OK.

Create DRS VM Groups

To create DRS VM groups specific to site A and site B, complete the following steps:

1. In the vSphere web client, right-click the cluster in the inventory and select Settings.
9. Click VM\Host Groups.
10. Click Add.
11. Type the name for the group (for instance, sitea_vms).
12. From the Type menu, select VM Group.
13. Click Add and select the desired VMs from site A and click OK.
14. Repeat these steps to add another host group for site B.
15. Click OK.

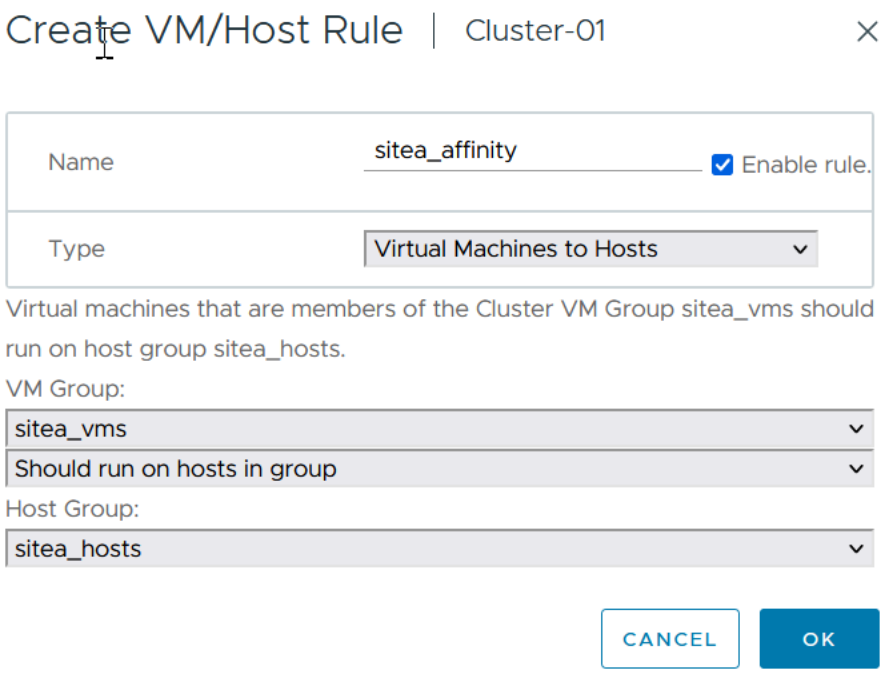
Create VM Host Rules

To create DRS affinity rules specific to site A and site B, complete the following steps:

1. In the vSphere web client, right-click the cluster in the inventory and select Settings.

1. Click VM\Host Rules.
2. Click Add.
3. Type the name for the rule (for instance, sitea_affinity).
4. Verify the Enable Rule option is checked.
5. From the Type menu, select Virtual Machines to Hosts.
6. Select the VM group (for instance, sitea_vms).
7. Select the Host group (for instance, sitea_hosts).

8. Repeat these steps to add another VM\Host Rule for site B.
9. Click OK.



VMWare vSphere Storage DRS for NetApp MetroCluster

Create Datastore Clusters

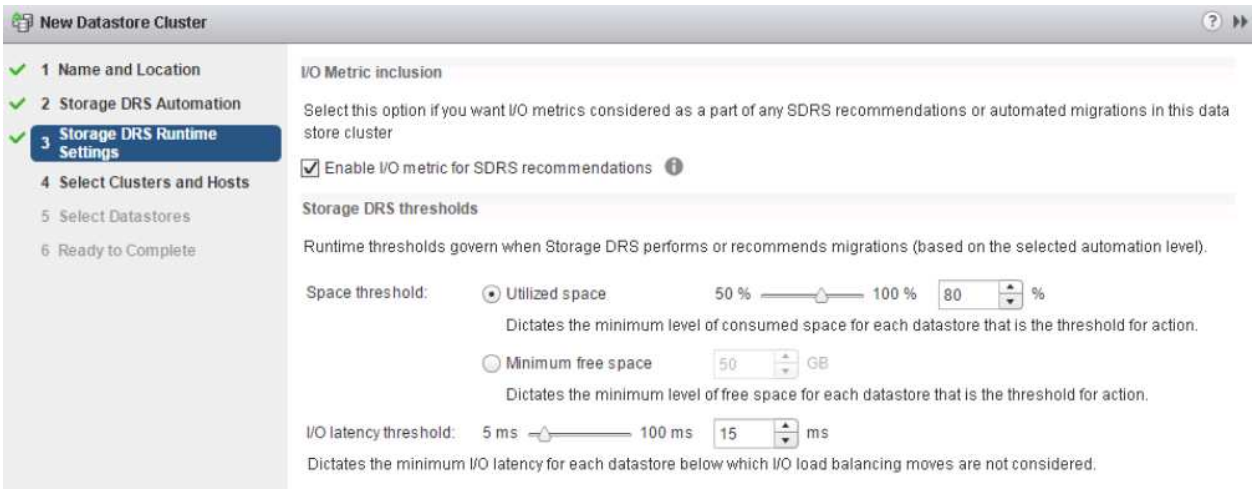
To configure a datastore cluster for each site, complete the following steps:

1. Using the vSphere web client, browse to the data center where the HA cluster resides under Storage.
2. Right-click the data center object and select Storage > New Datastore Cluster.
3. Select the Turn ON Storage DRS option and click Next.
4. Set all options to No Automation (Manual Mode) and click Next.

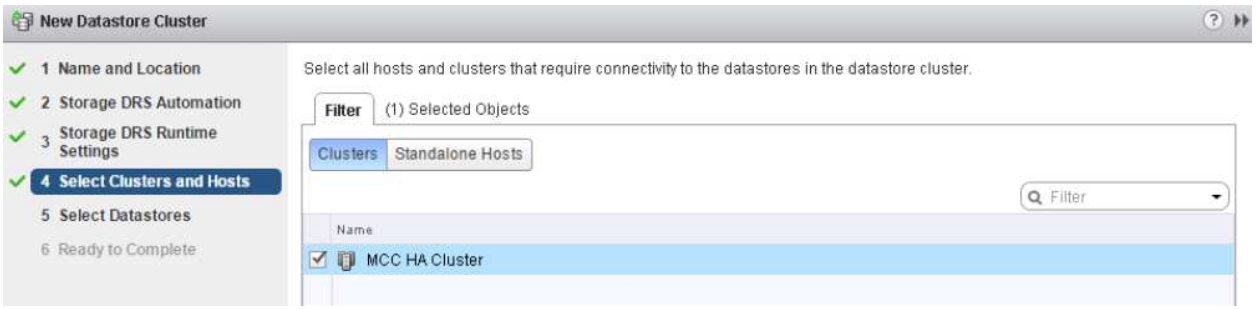
Best Practice

- NetApp recommends that Storage DRS be configured in manual mode, so that the administrator gets to decide and control when migrations need to happen.

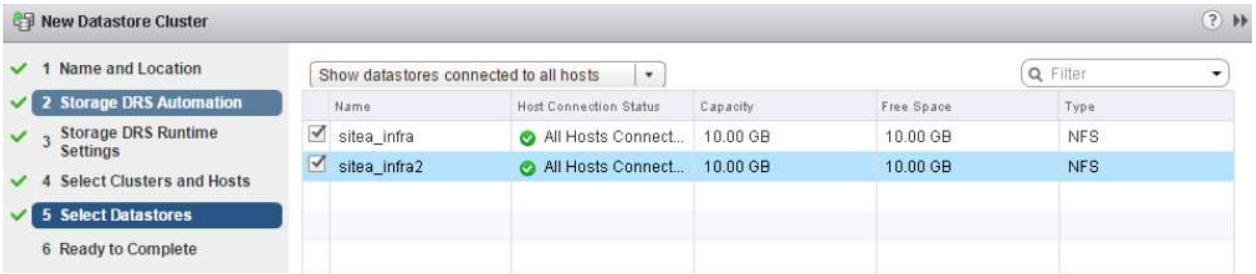
5. Verify that the Enable I/O Metric for SDRS Recommendations checkbox is checked; metric settings can be left with default values.



6. Select the HA cluster and click Next.



7. Select the datastores belonging to site A and click Next.



8. Review options and click Finish.
9. Repeat these steps to create the site B datastore cluster and verify that only datastores of site B are selected.

vCenter Server Availability

Your vCenter Server Appliances (VCSAs) should be protected with vCenter HA. vCenter HA allows you to deploy two VCSAs in an active-passive HA pair. One in each failure domain. You can read more about vCenter HA on docs.vmware.com.

Resiliency for Planned and Unplanned Events

NetApp MetroCluster and SnapMirror active sync are powerful tools that enhance the high availability and non-disruptive operations of NetApp hardware and ONTAP® software.

These tools provide site-wide protection for the entire storage environment, ensuring that your data is always available. Whether you are using standalone servers, high-availability server clusters, Docker containers, or virtualized servers, NetApp technology seamlessly maintains storage availability in the event of a total outage due to loss of power, cooling, or network connectivity, storage array shutdown, or operational error.

MetroCluster and SnapMirror active sync provide three basic methods for data continuity in the event of planned or unplanned events:

- Redundant components for protection against single-component failure
- Local HA takeover for events affecting a single controller
- Complete site protection – rapid resumption of service by moving storage and client access from the source cluster to the destination cluster

This means operations continue seamlessly in case of a single component failure and return automatically to redundant operation when the failed component is replaced.

All ONTAP clusters, except single-node clusters (typically software-defined versions, such as ONTAP Select for example), have built-in HA features called takeover and giveback. Each controller in the cluster is paired with another controller, forming an HA pair. These pairs ensure that each node is locally connected to the storage.

Takeover is an automated process where one node takes over the other's storage to maintain data services. Giveback is the reverse process that restores normal operation. Takeover can be planned, such as when performing hardware maintenance or ONTAP upgrades, or unplanned, resulting from a node panic or hardware failure.

During a takeover, Network Attached Storage Logical Interfaces (NAS LIFs) in MetroCluster configurations automatically failover. However, Storage Area Network LIFs (SAN LIFs) do not fail over; they will continue to use the direct path to the Logical Unit Numbers (LUNs).

For more information on HA takeover and giveback, please refer to the HA pair management overview. It's worth noting that this functionality is not specific to MetroCluster or SnapMirror active sync.

Site switchover with MetroCluster occurs when one site is offline or as a planned activity for site-wide maintenance. The remaining site assumes ownership of the storage resources (disks and aggregates) of the offline cluster, and the SVMs on the failed site are brought online and restarted on the disaster site, preserving their full identity for client and host access.

With SnapMirror active sync, since both copies are actively used simultaneously, your existing hosts will continue to operate. The NetApp Mediator is required to ensure site failover occurs correctly.

Failure Scenarios for vMSC with MCC

The following sections outline the expected results from various failure scenarios with vMSC and NetApp MetroCluster systems.

Single Storage Path Failure

In this scenario, if components such as the HBA port, the network port, the front-end data switch port, or an FC or Ethernet cable fails, that particular path to the storage device is marked as dead by the ESXi host. If several paths are configured for the storage device by providing resiliency at the HBA/network/switch port, ESXi ideally performs a path switchover. During this period, virtual machines remain running without getting affected, because availability to the storage is taken care of by providing multiple paths to the storage device.

Note: There is no change in MetroCluster behavior in this scenario, and all the datastores continue to be intact from their respective sites.

Best Practice

In environments in which NFS/iSCSI volumes are used, NetApp recommends having at least two network uplinks configured for the NFS vmkernel port in the standard vSwitch and the same at the port group where the NFS vmkernel interface is mapped for the distributed vSwitch. NIC teaming can be configured in either active-active or active-standby.

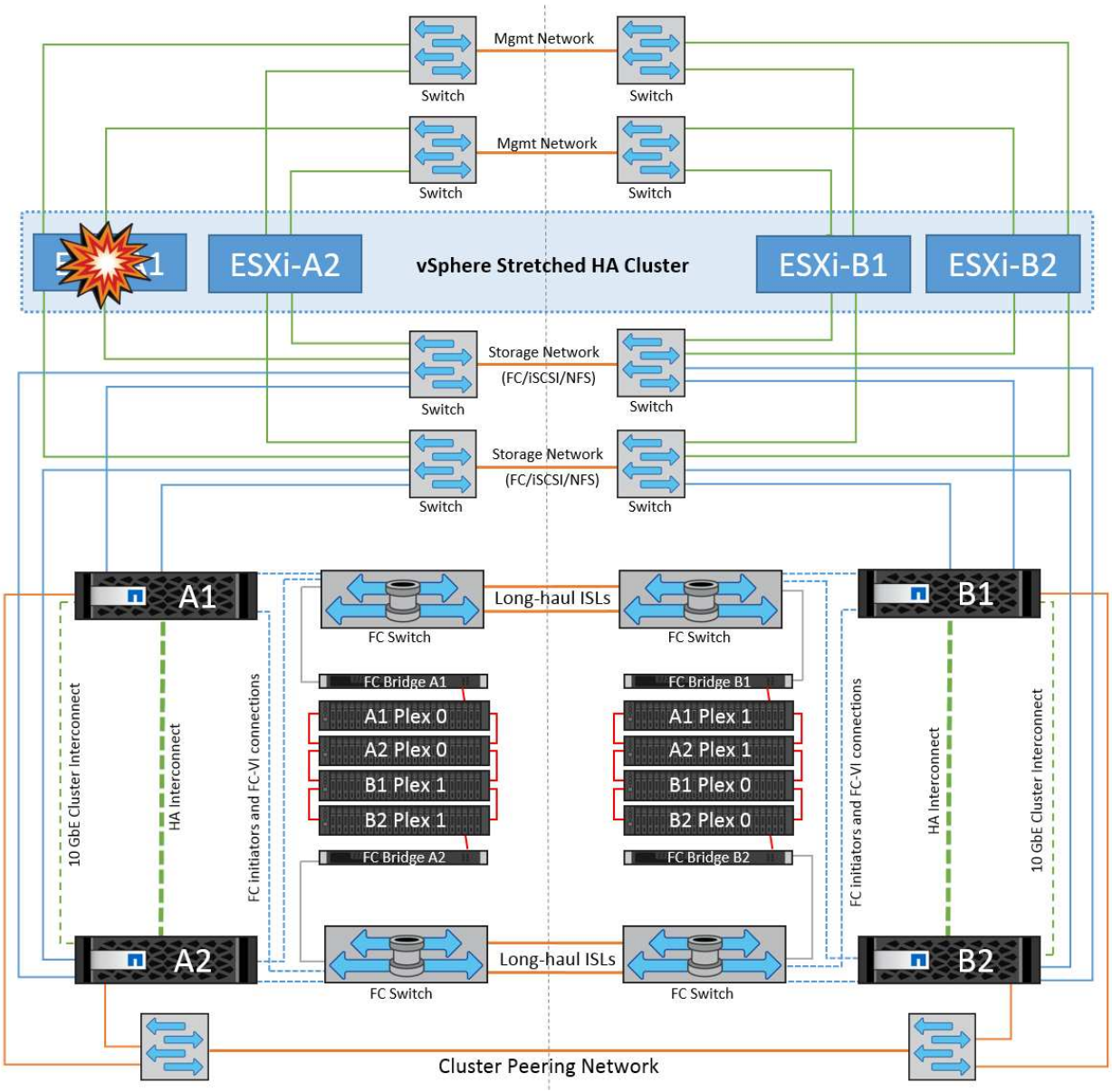
Also, for iSCSI LUNs, multipathing must be configured by binding the vmkernel interfaces to the iSCSI network adapters. For more information, refer to the vSphere storage documentation.

Best Practice

In environments in which Fibre Channel LUNs are used, NetApp recommends having at least two HBAs, which guarantees resiliency at the HBA/port level. NetApp also recommends single initiator to single target zoning as the best practice to configure zoning.

Virtual Storage Console (VSC) should be used to set multipathing policies because it sets policies for all new and existing NetApp storage devices.

Single ESXi Host Failure



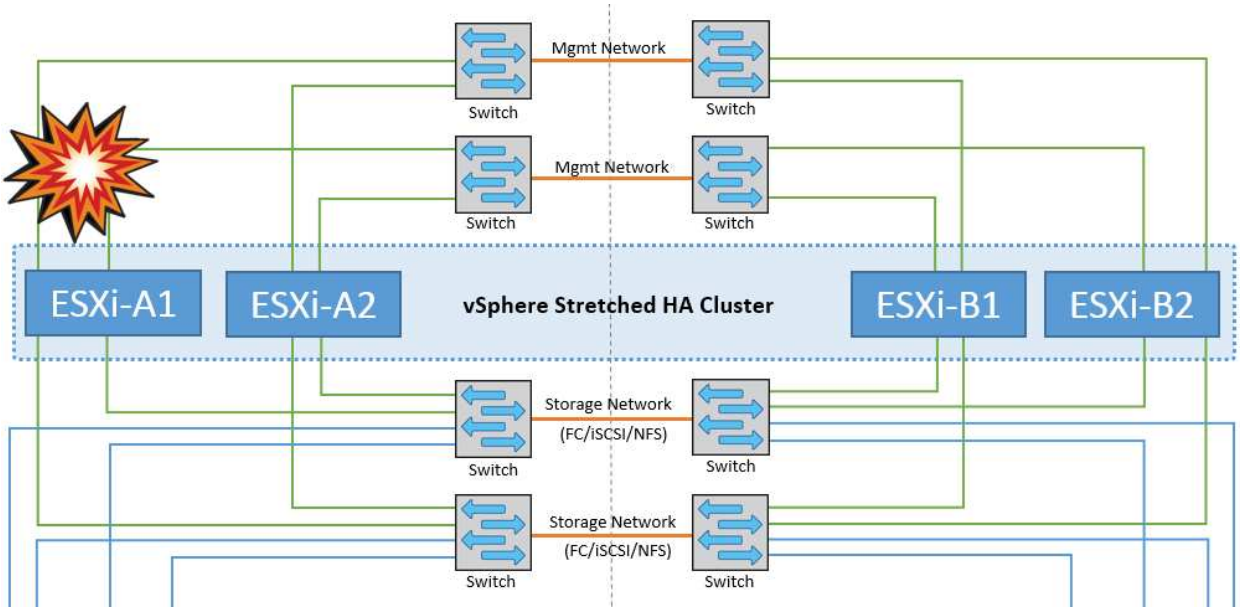
In this scenario, if there is an ESXi host failure, the master node in the VMware HA cluster detects the host failure since it no longer receives network heartbeats. To determine whether the host is really down or only a network partition, the master node monitors the datastore heartbeats and, if they are absent, it performs a final check by pinging the management IP addresses of the failed host. If all these checks are negative, then the master node declares this host a failed host and all the virtual machines that were running on this failed host are rebooted on the surviving host in the cluster.

If DRS VM and host affinity rules have been configured (VMs in VM group sitea_vms should run hosts in host group sitea_hosts), then the HA master first checks for available resources at site A. If there are no available hosts at site A, the master attempts to restart the VMs on hosts at site B.

It is possible that the virtual machines will be started on the ESXi hosts at the other site if there is a resource constraint in the local site. However, the defined DRS VM and host affinity rules will correct if any rules are violated by migrating the virtual machines back to any surviving ESXi hosts in the local site. In cases in which DRS is set to manual, NetApp recommends invoking DRS and applying the recommendations to correct the virtual machine placement.

There is no change in the MetroCluster behavior in this scenario and all the datastores continue to be intact from their respective sites.

ESXi Host Isolation

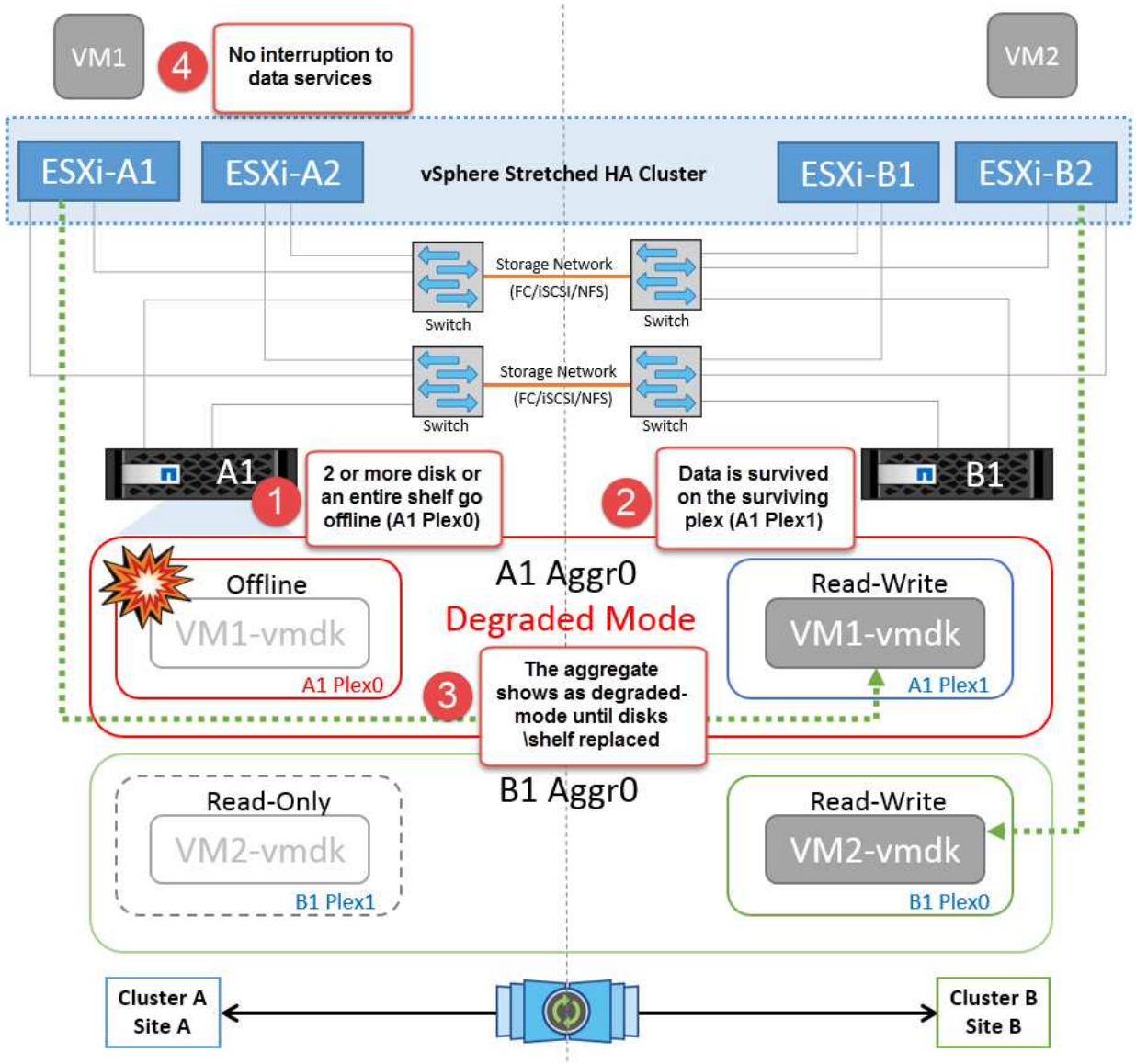


In this scenario, if the management network of the ESXi host is down, the master node in the HA cluster will not receive any heartbeats, and thus this host becomes isolated in the network. To determine whether it has failed or is only isolated, the master node starts monitoring the datastore heartbeat. If it is present then the host is declared isolated by the master node. Depending on the isolation response configured, the host may choose to power off, shut down the virtual machines, or even leave the virtual machines powered on. The default interval for the isolation response is 30 seconds.

There is no change in the MetroCluster behavior in this scenario and all the datastores continue to be intact from their respective sites.

Disk Shelf Failure

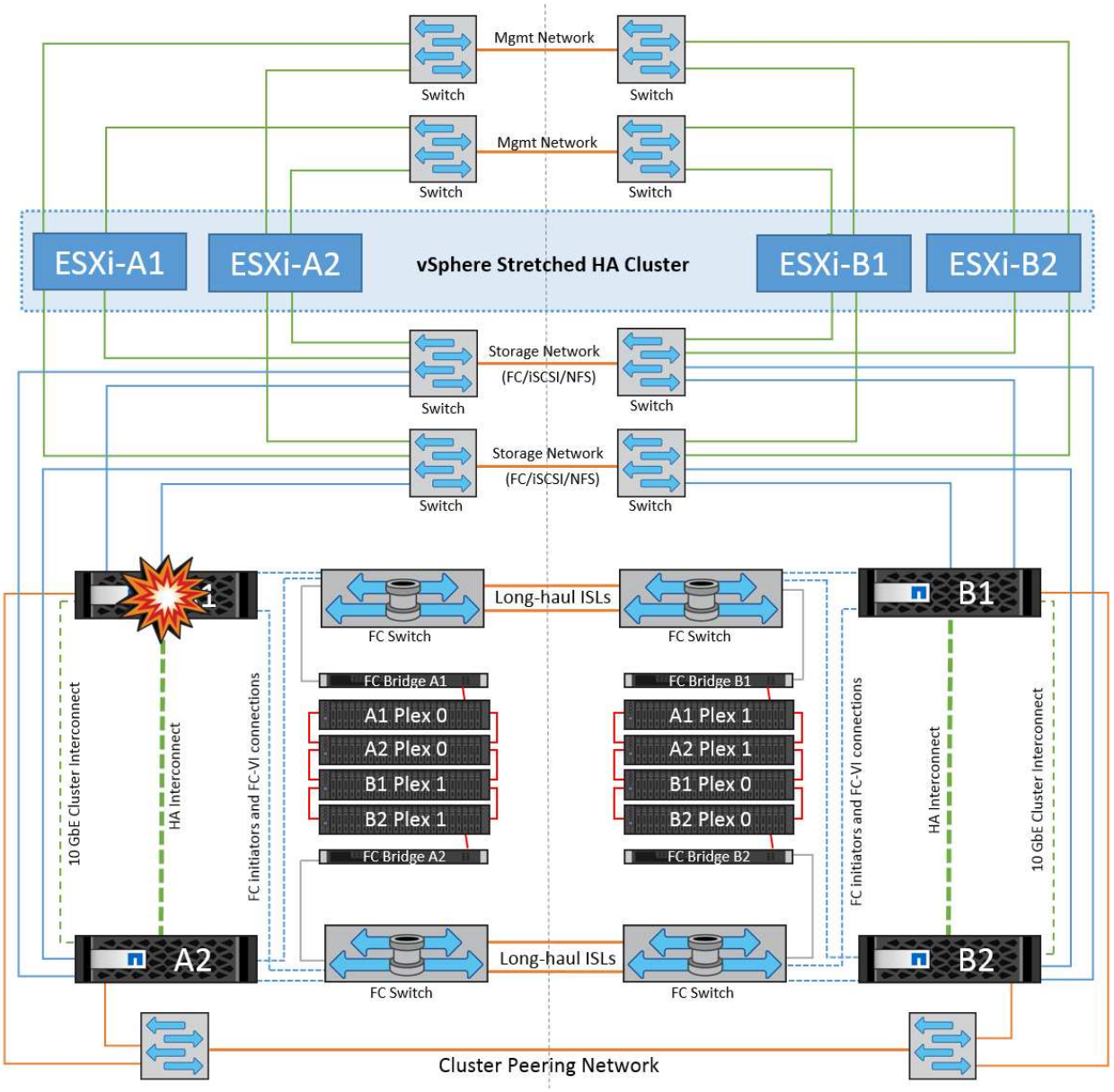
In this scenario, there is a failure of more than two disks or an entire shelf. Data is served from the surviving plex with no interruption to data services. The disk failure could affect either a local or remote plex. The aggregates will show as degraded mode because only one plex is active. Once the failed disks are replaced, the affected aggregates will automatically resync to rebuild the data. After resync, the aggregates will return automatically to normal mirrored mode. If more than two disks within a single RAID group have failed, then the plex has to be rebuilt from scratch.



Note: During this period, there is no impact on the virtual machine I/O operations, but there is degraded performance because the data is being accessed from the remote disk shelf through ISL links.

Single Storage Controller Failure

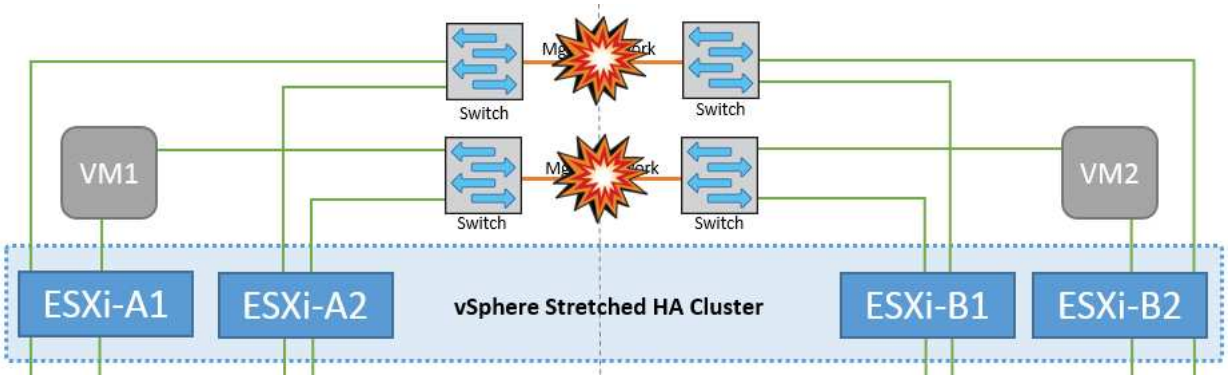
In this scenario, one of the two storage controllers fails at one site. Because there is an HA pair at each site, a failure of one node transparently and automatically triggers failover to the other node. For example, if node A1 fails, its storage and workloads are automatically transferred to node A2. Virtual machines will not be affected because all plexes remain available. The second site nodes (B1 and B2) are unaffected. In addition, vSphere HA will not take any action because the master node in the cluster will still be receiving the network heartbeats.



If the failover is part of a rolling disaster (node A1 fails over to A2), and there is a subsequent failure of A2, or the complete failure of site A, switchover following a disaster can occur at site B.

Interswitch Link Failures

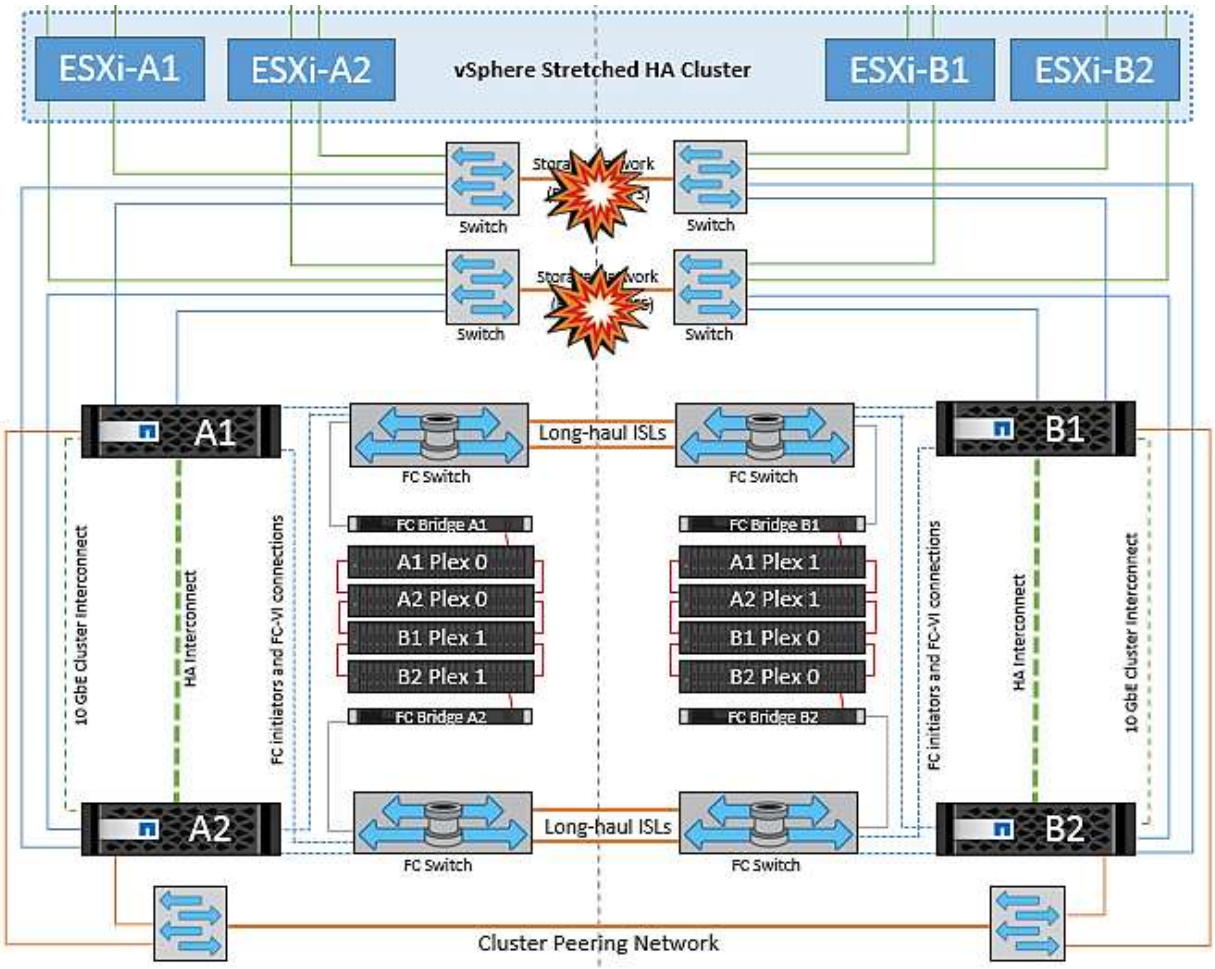
Interswitch Link Failure at Management Network



In this scenario, if the ISL links at the front-end host management network fail, the ESXi hosts at site A will not be able to communicate with ESXi hosts at site B. This will lead to a network partition because ESXi hosts at a particular site will be unable to send the network heartbeats to the master node in the HA cluster. As such, there will be two network segments because of partition and there will be a master node in each segment that will protect VMs from host failures within the particular site.

Note: During this period, the virtual machines remain running and there is no change in the MetroCluster behavior in this scenario. All the datastores continue to be intact from their respective sites.

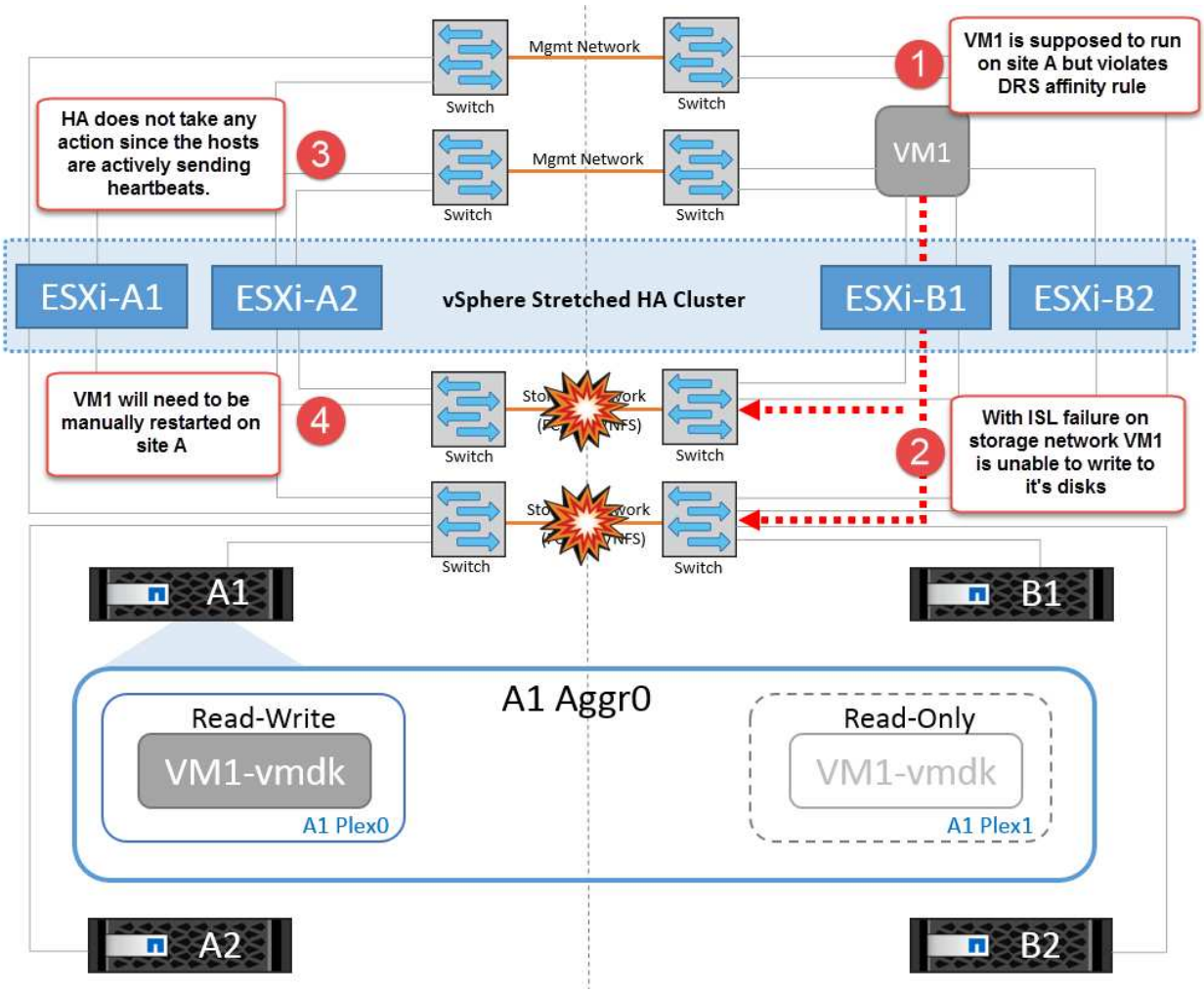
Interswitch Link Failure at Storage Network



In this scenario, if the ISL links at the backend storage network fail, the hosts at site A will lose access to the storage volumes or LUNs of cluster B at site B and vice versa. The VMware DRS rules are defined so that host-storage site affinity facilitates the virtual machines to run without impact within the site.

During this period, the virtual machines remain running in their respective sites and there is no change in the MetroCluster behavior in this scenario. All the datastores continue to be intact from their respective sites.

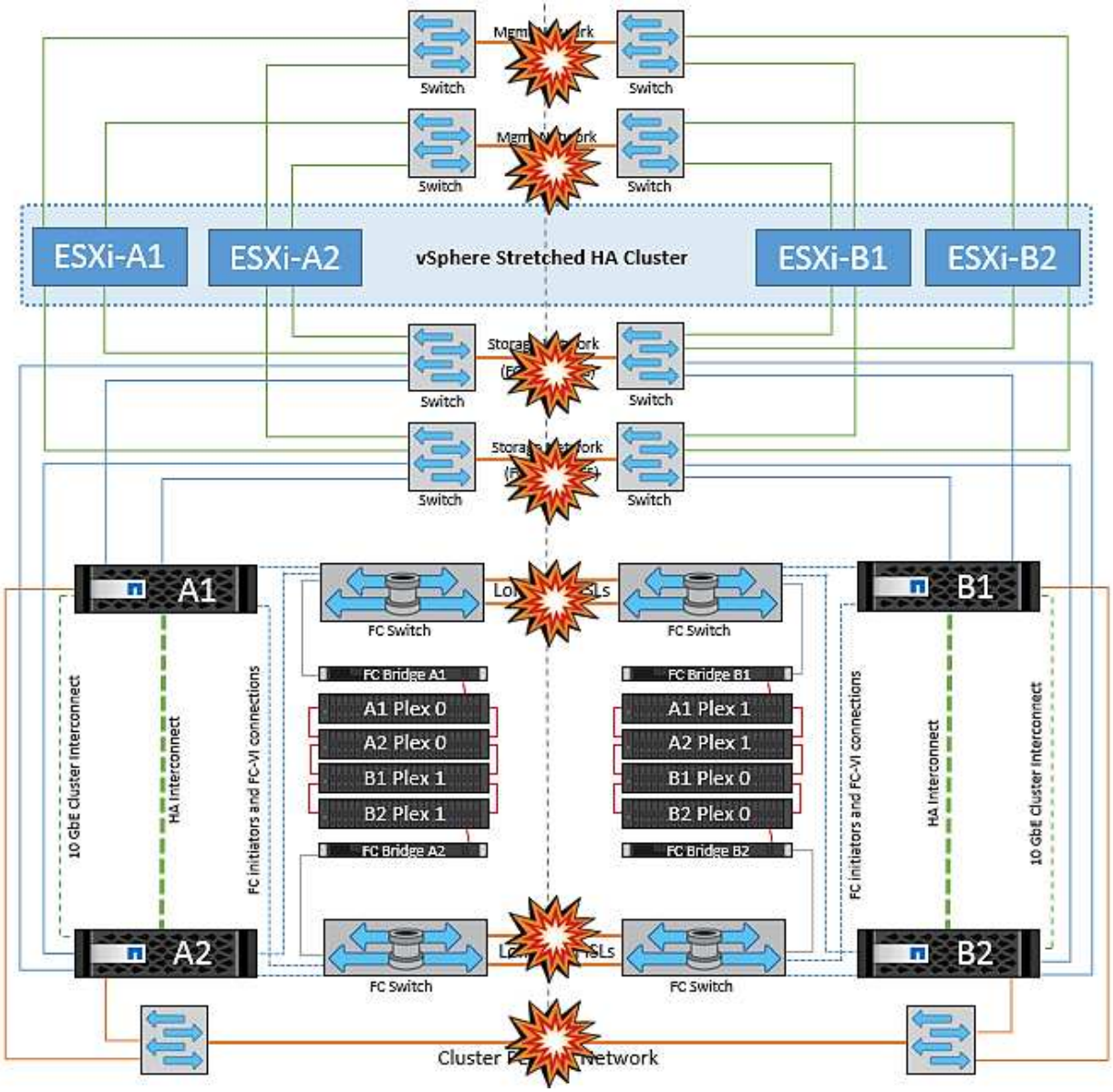
If for some reason the affinity rule was violated (for example, VM1, which was supposed to run from site A where its disks reside on local cluster A nodes, is running on a host at site B), the virtual machine's disk will be remotely accessed via ISL links. Because of ISL link failure, VM1 running at site B would not be able to write to its disks because the paths to the storage volume are down and that particular virtual machine is down. In these situations, VMware HA does not take any action since the hosts are actively sending heartbeats. Those virtual machines need to be manually powered off and powered on in their respective sites. The following figure illustrates a VM violating a DRS affinity rule.



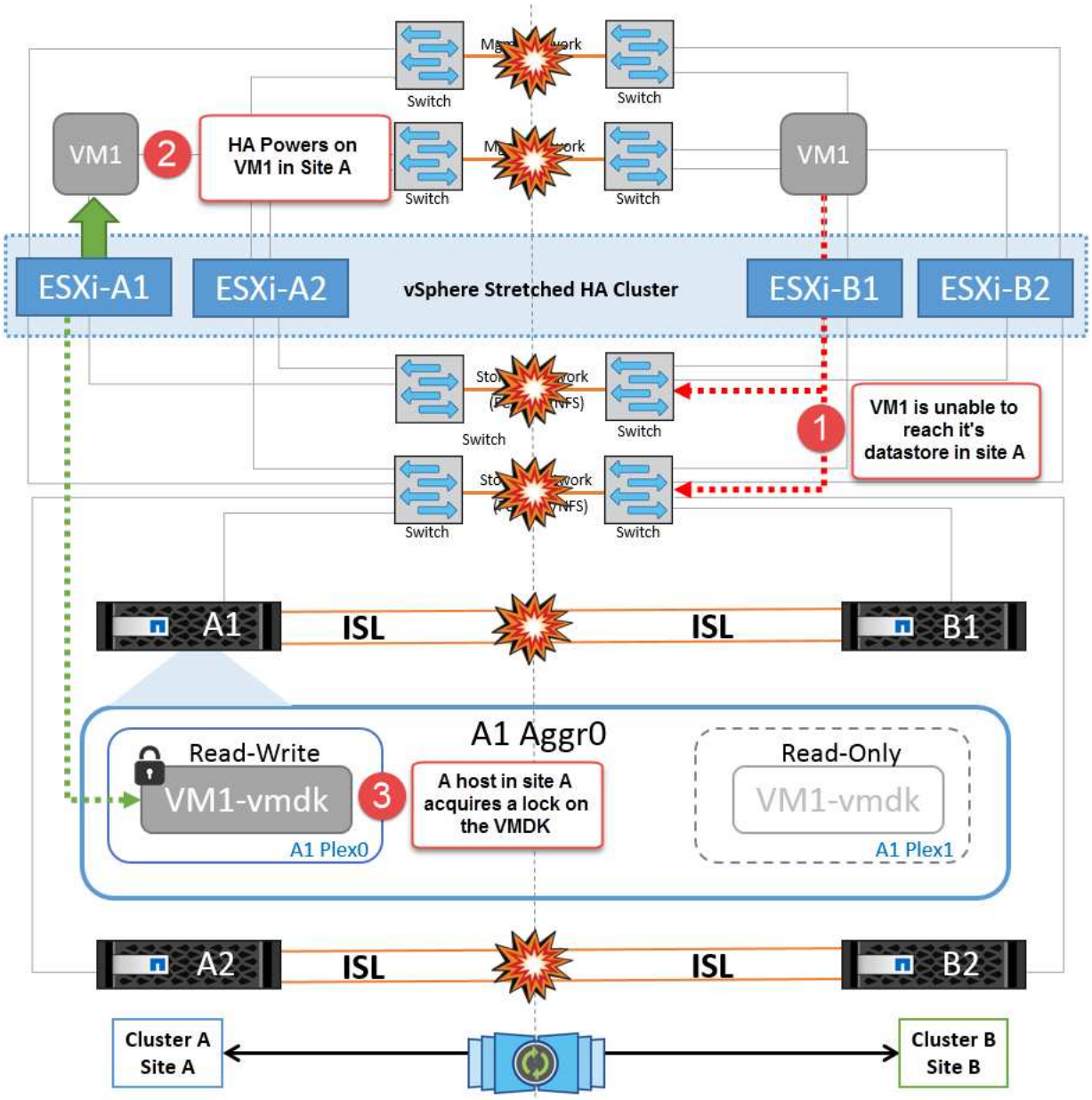
All Interswitch Failure or Complete Data Center Partition

In this scenario, all the ISL links between the sites are down and both the sites are isolated from each other. As discussed in earlier scenarios, such as ISL failure at the management network and at the storage network, the virtual machines are not affected in complete ISL failure.

After ESXi hosts are partitioned between sites, the vSphere HA agent will check for datastore heartbeats and, in each site, the local ESXi hosts will be able to update the datastore heartbeats to their respective read-write volume/LUN. Hosts in site A will assume that the other ESXi hosts at site B have failed because there are no network/datastore heartbeats. vSphere HA at site A will try to restart the virtual machines of site B, which will eventually fail because the datastores of site B will not be accessible due to storage ISL failure. A similar situation is repeated in site B.



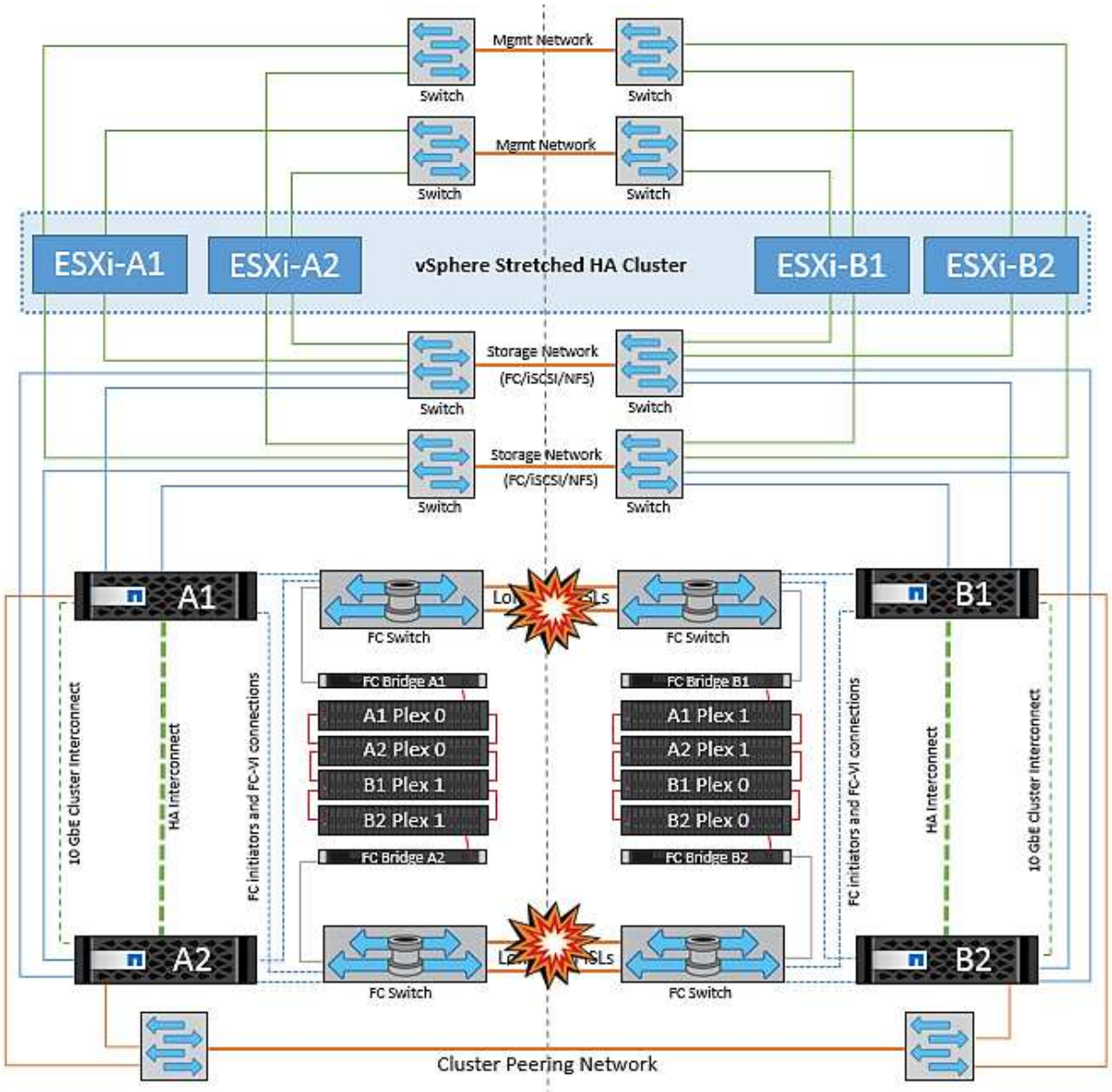
NetApp recommends determining if any virtual machine has violated the DRS rules. Any virtual machines running from a remote site will be down since they will not be able to access the datastore, and vSphere HA will restart that virtual machine on the local site. After the ISL links are back online, the virtual machine that was running in the remote site will be killed, since there cannot be two instances of virtual machines running with the same MAC addresses.



Interswitch Link Failure on Both Fabrics in NetApp MetroCluster

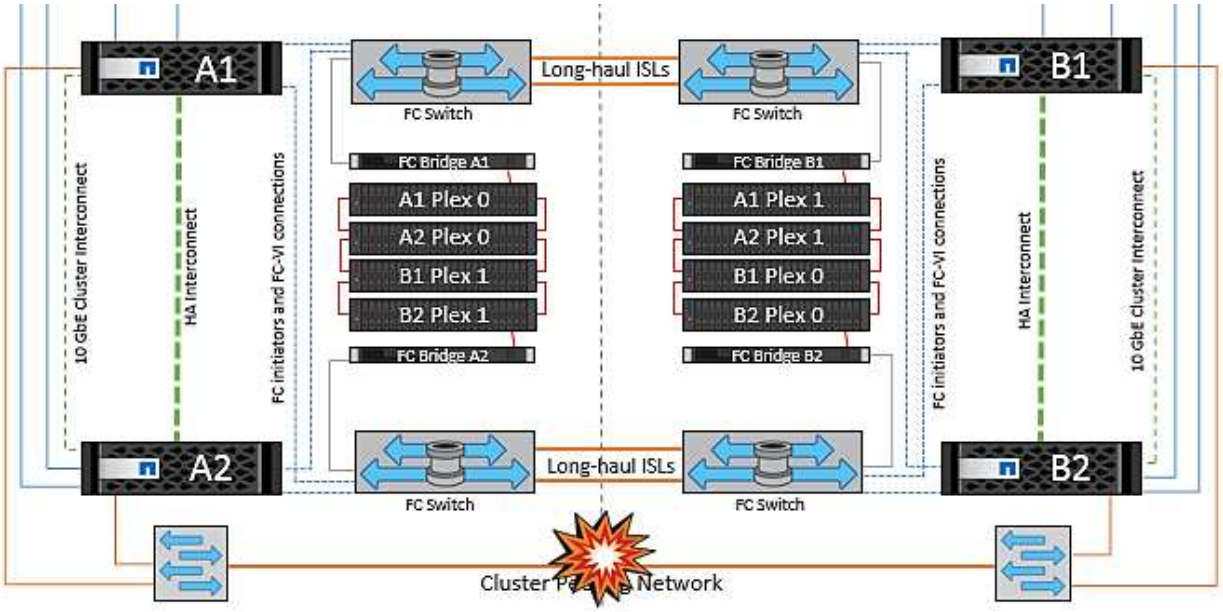
In a scenario of one or more ISLs failing, traffic continues through the remaining links. If all ISLs on both fabrics fail, such that there is no link between the sites for storage and NVRAM replication, each controller will continue to serve its local data. On restoration of a minimum of one ISL, resynchronization of all the plexes will happen automatically.

Any writes occurring after all ISLs are down will not be mirrored to the other site. A switchover on disaster, while the configuration is in this state, would therefore incur loss of the data that had not been synchronized. In this case, manual intervention is required for recovery after the switchover. If it is likely that no ISLs will be available for an extended period, an administrator can choose to shut down all data services to avoid the risk of data loss if a switchover on disaster is necessary. Performing this action should be weighed against the likelihood of a disaster requiring switchover before at least one ISL becomes available. Alternatively, if ISLs are failing in a cascading scenario, an administrator could trigger a planned switchover to one of the sites before all the links have failed.



Peered Cluster Link Failure

In a peered cluster link failure scenario, because the fabric ISLs are still active, data services (reads and writes) continue at both sites to both plexes. Any cluster configuration changes (for example, adding a new SVM, provisioning a volume or LUN in an existing SVM) cannot be propagated to the other site. These are kept in the local CRS metadata volumes and automatically propagated to the other cluster upon restoration of the peered cluster link. If a forced switchover is necessary before the peered cluster link can be restored, outstanding cluster configuration changes will be replayed automatically from the remote replicated copy of the metadata volumes at the surviving site as part of the switchover process.



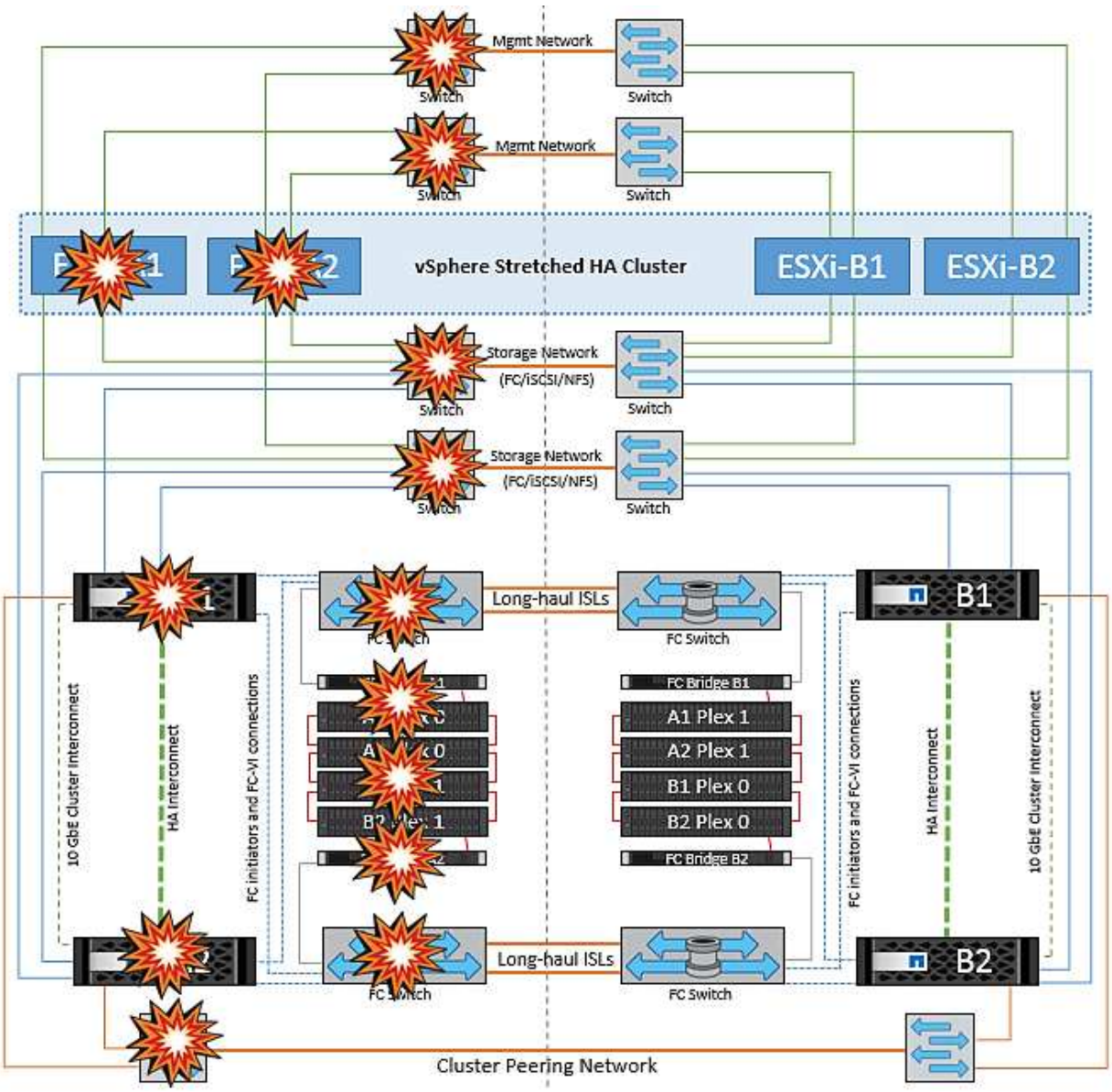
Complete Site Failure

In a complete site A failure scenario, the ESXi hosts at site B will not get the network heartbeat from the ESXi hosts at site A because they are down. The HA master at site B will verify that the datastore heartbeats are not present, declare the hosts at site A to be failed, and try to restart the site A virtual machines in site B. During this period, the storage administrator performs a switchover to resume services of the failed nodes on the surviving site which will restore all the storage services of site A at site B. After the site A volumes or LUNs are available at site B, the HA master agent will attempt to restart the site A virtual machines in site B.

If the vSphere HA master agent's attempt to restart a VM (which involves registering it and powering it on) fails, the restart is retried after a delay. The delay between restarts can be configured to up to a maximum of 30 minutes. vSphere HA attempts these restarts for a maximum number of attempts (six attempts by default).

Note: The HA master does not begin the restart attempts until the placement manager finds suitable storage, so in the case of a complete site failure, that would be after the switchover has been performed.

If site A has been switched over, a subsequent failure of one of the surviving site B nodes can be seamlessly handled by failover to the surviving node. In this case, the work of four nodes is now being performed by only one node. Recovery in this case would consist of performing a giveback to the local node. Then, when site A is restored, a switchback operation is performed to restore steady state operation of the configuration.


Mgmt Network
Switch
Mgmt Network
Switch
vSphere Stretched HA Cluster
ESXi-B1
ESXi-B2
Storage Network
(FC/iSCSI/NFS)
Switch
Storage Network
(FC/iSCSI/NFS)
Switch
Long-haul ISLs
FC Switch
B1
FC Bridge B1
A1 Plex 1
A2 Plex 1
B1 Plex 0
B2 Plex 0
FC Bridge B2
Long-haul ISLs
FC Switch
B2
10 GbE Cluster Interconnect
HA Interconnect
FC initiators and FC-VI connections
FC initiators and FC-VI connections
HA Interconnect
10 GbE Cluster Interconnect
Cluster Peering Network

Copyright information

Copyright © 2024 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.