Technical Report

MetroCluster for Clustered Data ONTAP 8.3.2

Roy Scaife, NetApp

Version 1.2.1
March 2016 | TR-4375

Abstract

This document provides technical information about NetApp® MetroCluster™ software in the NetApp Data ONTAP® 8.3.2 operating system.

Data Classification

Public

Version History

Version	Date	Document Version History
Version 1.2.1	March 2016	Roy Scaife: Revised distance for Stretch MetroCluster for ATTO 7500N between clusters.
Version 1.2	February 2016	Roy Scaife: Clustered Data ONTAP 8.3.2 updates for FCIP configurations and new features.
Version 1.1	September 2015	Roy Scaife: Clustered Data ONTAP 8.3.1 updates for two-node configurations and minor corrections.
Version 1.0	April 2015	Charlotte Brooks: Initial version.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction to NetApp MetroCluster

Starting with NetApp clustered Data ONTAP 8.3, NetApp MetroCluster provides continuous data availability across geographically separated data centers for mission-critical applications. MetroCluster high-availability and disaster recovery software runs on the clustered Data ONTAP storage operating system. Fabric and stretch MetroCluster are actively used by thousands of enterprises worldwide for high availability, zero data loss, and nondisruptive operations both within and beyond the data center.

This technical report focuses on MetroCluster for the clustered Data ONTAP 8.3.x operating system. Fabric MetroCluster and stretch MetroCluster in Data ONTAP operating in 7-Mode continue to be supported through Data ONTAP 8.2.x. Unless otherwise explicitly stated, the term “MetroCluster” in this paper refers to MetroCluster in clustered Data ONTAP 8.3 and later. For more information about MetroCluster in 7-Mode, see TR-3548: MetroCluster Version 8.2.1 Best Practices for Implementation.

It is assumed that the reader is familiar with the clustered Data ONTAP architecture and its capabilities. TR-3982: NetApp Clustered Data ONTAP 8.3.x and 8.2.x – An Introduction presents clustered Data ONTAP concepts.

1.1 Features

In today’s enterprises, the IT department must meet increasing service-level demands while maintaining cost and operational efficiency. As data volumes explode and applications consolidate and move to shared virtual infrastructures, the need for continuous availability for both mission-critical and other business applications dramatically increases. With data and application consolidation, the storage infrastructure itself becomes a critical asset. For some enterprises, perhaps no single application warrants the “mission-critical” designation. However, for all enterprises, the loss of the storage infrastructure as a whole for even a short period has substantial adverse effects on the company’s revenue and reputation.

MetroCluster maintains the availability of the storage infrastructure and provides the following key benefits:

- Transparent recovery from failures:
  - The clustered Data ONTAP storage operating system provides nondisruptive operations within the data center. It withstands component, node, and network failures and allows planned hardware and software upgrades.
  - MetroCluster extends business continuity and continuous availability beyond the data center to a second data center. MetroCluster configurations provide automatic takeover (for local high availability) and manual switchover (from one data center to the other).
- Combines array-based clustering with synchronous mirroring to deliver zero data loss:
  - It provides a recovery point objective (RPO), which is the maximum amount of acceptable data loss, of zero.
  - It provides a recovery time objective (RTO) of 120 seconds or less for planned switchover and switchback. RTO is defined as the maximum acceptable time period that is required to make the storage and associated data available in a correct operational state after a switchover to the other data center.
- Reduced administrative overhead:
  - After the initial setup, subsequent changes on one cluster are automatically replicated to the second cluster.
  - Ongoing management and administration are almost identical to a clustered Data ONTAP environment by using NetApp OnCommand® System Manager, OnCommand Unified Manager, and OnCommand Workflow Automation.
  - Zero (or minimal) changes are required to applications, hosts, and clients. MetroCluster is designed to be transparent and agnostic to any front-end application environment. Connection

paths are identical before and after switchover, so most applications, hosts, and clients (NFS and SAN) do not need to reconnect or rediscover their storage but instead automatically resume.

Note: SMB applications, including SMB 3 with continuous availability shares, need to reconnect after a switchover or a switchback. This need is a limitation of the SMB protocol.

- Complements the full power of clustered Data ONTAP:
  - MetroCluster provides multiprotocol support for a wide range of SAN and NAS client and host protocols.
  - Operations are nondisruptive for technology refresh, capacity, and performance management.
  - Quality of Service (QoS) can be implemented to restrict the performance of less critical workloads.
  - Data deduplication and compression work in both SAN and NAS environments.
  - Data management and replication are integrated with enterprise applications.
- Lower cost:
  - MetroCluster lowers your acquisition costs and the cost of ownership because of its easy-to-manage architecture. MetroCluster capability is integrated directly into clustered Data ONTAP and requires no additional license to be purchased or installed.
- Simplified disaster recovery:
  - During a total site loss event, services can be transitioned to the disaster recovery site with a single command within minutes. No complex failover scripts or procedures are required.

1.2 New Features in MetroCluster 8.3.1

Starting in MetroCluster 8.3.1, NetApp introduced support of the two-node architecture for clustered Data ONTAP. Enhancements include:

- Two-node fabric MetroCluster
- Two-node stretch MetroCluster
- NetApp All Flash FAS (AFF) without requiring policy-variance request (PVR) approval
- Reduced data aggregate requirements
- Improved utilization for the NetApp FAS8020 platform

1.3 New Features in MetroCluster 8.3.2

Starting in MetroCluster 8.3.2, the software delivers new features to enhance the two-node and four-node architecture for the clustered Data ONTAP operating system. Enhancements include:

- Increased Fibre channel (FC) ISL distance between the MetroCluster clusters
  - Up to 185 miles (300km) using the Brocade 6505 or 6510 switch
  - 185 miles (300km) is not supported using the Cisco FC switches
- Fibre Channel over IP (FCIP) Inter-Switch Link (ISL) connectivity between the MetroCluster clusters:
  - Uses the Cisco MDS 9250i FCIP fabric switch
  - Supports up to 120 miles (200 kilometers) between the MetroCluster clusters
- Starting in Data ONTAP 8.3.2 GA, the ATTO 7500N FibreBridge is supported for all MetroCluster fabric configurations.
  - Supports dual-power supplies
  - Provides increased performance (compared to the ATTO 6500N)
- New functionality for shared FC-VI and FC connectivity on the same initiator card (X1132A-EN-R6-C):
  - NetApp FAS8020 and AFF8020 only

1.4 Architecture and Supported Configurations

NetApp MetroCluster is designed for organizations that require continuous protection of their storage infrastructure and mission-critical business applications. By synchronously replicating data between geographically separated clusters, MetroCluster provides a zero-touch continuous availability solution that guards against faults inside and outside the array.

Standard MetroCluster Configuration

MetroCluster configurations protect data by using two distinct clusters that are separated by a distance of up to 185 miles (300km). Each cluster synchronously mirrors the data and configuration information of the other. Effectively, all storage virtual machines (SVMs) and their associated configurations are replicated. Independent clusters provide isolation and resilience to logical errors.

If a disaster occurs at one site, an administrator can perform a switchover, which activates the mirrored SVMs and resumes serving the mirrored data from the surviving site. In clustered Data ONTAP 8.3.x, the MetroCluster four-node configuration consists of a two-node high-availability (HA) pair at each site. This configuration allows the majority of planned and unplanned events to be handled by a simple failover and giveback within the local cluster. Full switchover to the other site is required only in the event of a disaster or for testing purposes. Switchover and the corresponding switchback operations transfer the entire clustered workload between the sites.

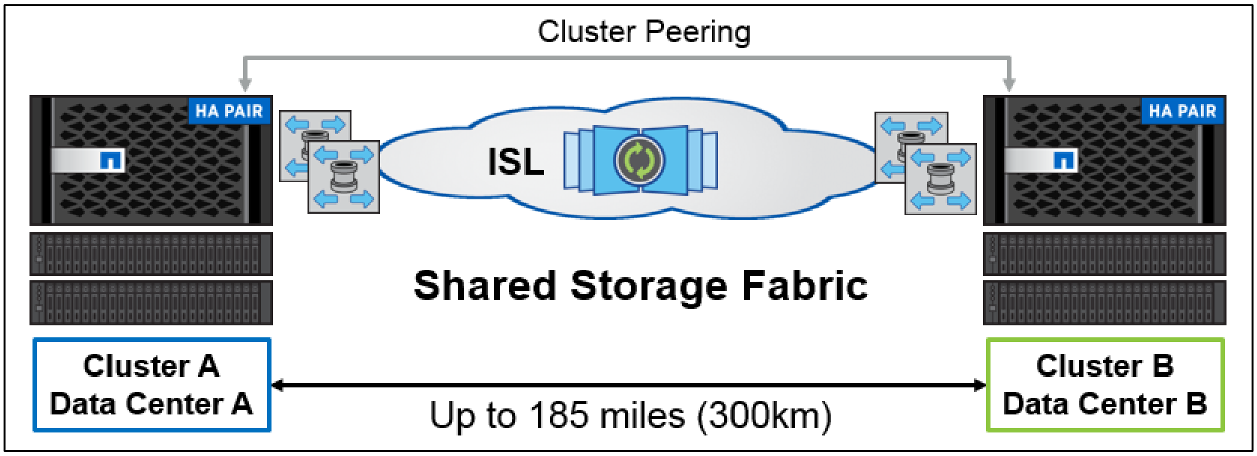
The MetroCluster two-node configuration has a one-node cluster at each site. Planned and unplanned events are handled by using switchover and switchback operations. Switchover and the corresponding switchback operations transfer the entire clustered workload between the sites.

Figure 1 shows the basic MetroCluster four-node configuration. The two data centers, A and B, are separated by a distance of up to 185 miles (300km) with ISLs running over dedicated FC links. The cluster at each site consists of two nodes in an HA pair. We use the following configuration and naming in this report:

- Cluster A: nodes A1 and A2
- Cluster B: nodes B1 and B2

The two clusters and sites are connected by two separate networks that provide the replication transport. The cluster peering network is an IP network that is used to replicate cluster configuration information between the sites. The shared storage fabric is an FC connection and is used for storage and NVRAM synchronous replication between the two clusters. All storage is visible to all controllers through the hared storage fabric.

Figure 1) Basic MetroCluster four-node configuration.

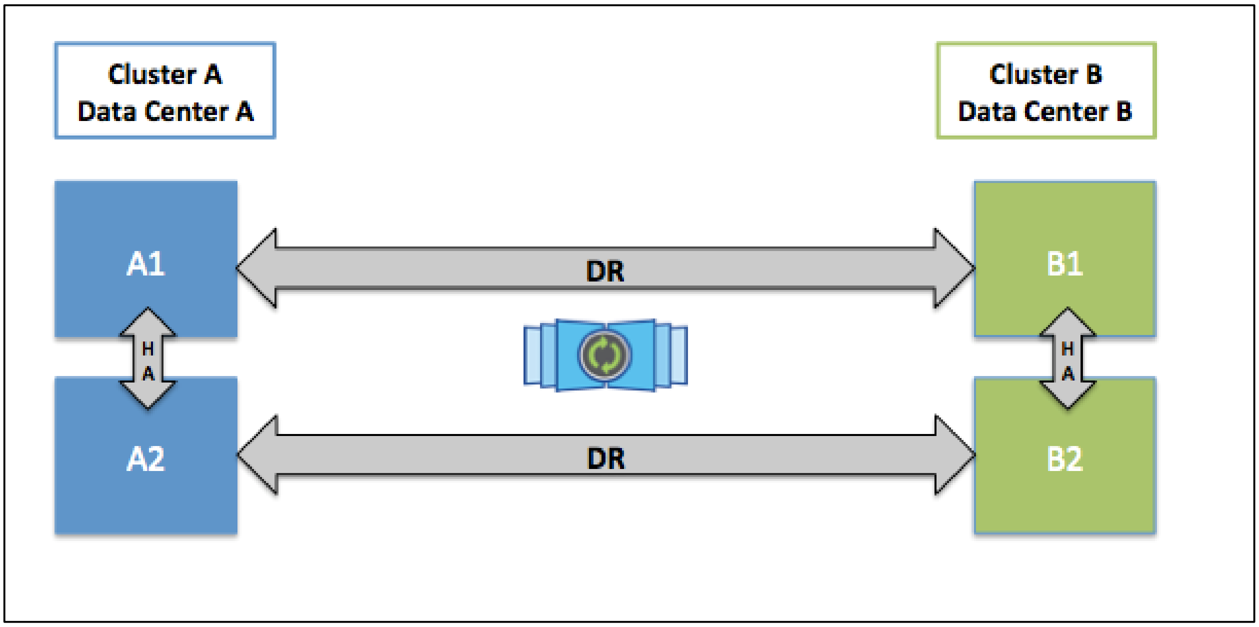


HA and DR Relationships

The four-node architecture provides both local HA failover and remote disaster recovery (DR) switchover. Each node has an *HA partner* in the same local cluster and a *DR partner* in the remote cluster, as shown in Figure 2. A1 and A2 are HA partners, as are B1 and B2. Node A1 and B1 are DR partners, as are A2 and B2. NVRAM is replicated to both the HA and the DR partner, as explained further in section 1.55. The DR partner for a node is automatically selected when MetroCluster is configured, and the partner is chosen according to system ID (NVRAM ID) order. See section 2.2 for more information about the selection of the DR partners.

In a two-node architecture, both HA failover and remote DR are accomplished by using MetroCluster switchover and switchback functionality. Each node is an independent cluster that functions as both the HA partner and a DR partner in the remote cluster. NVRAM is replicated to the remote partner, similar to a four-node configuration.

Figure 2) HA and DR partners.



In an HA failover, one of the nodes in the HA pair temporarily takes over the storage and services of it's HA partner. For example, node A2 takes over the resources of node A1. Takeover is enabled by mirrored NVRAM and multipathed storage between the two nodes. Failover can be planned, for example, to perform a nondisruptive clustered Data ONTAP upgrade, or it can be unplanned, such as a panic or hardware failure. Giveback is the reverse process: The failed node resumes its resources from the node that took over. Giveback is always a planned operation. Failover is always to the local HA partner, and either node can fail over to the other.

In a switchover, one cluster assumes the storage and services of the other cluster, while continuing to perform its own workloads. For example, if site A switches over to site B, cluster B nodes take temporary control of the storage and services owned by cluster A. After switchover, the SVMs from cluster A are brought online and continue running on cluster B. Switchover can be negotiated (planned), for example, to perform testing or site maintenance, or it can be forced (unplanned) in the event of a disaster that destroys one of the sites. Switchback is the process in which the surviving cluster sends the switched-over resources back to their original location to restore the steady operational state. Switchback is coordinated between the two clusters and is always a planned operation. Either site can switch over to the other.

It is also possible for a subsequent failure to occur while the site is in switchover. For example, after switchover to cluster B, suppose that node B1 then fails. B2 automatically takes over and services all workloads.

Active-Active or Active-Passive Configuration

MetroCluster is automatically enabled for symmetrical switchover and switchback; that is, either site can switch over to the other in the event of a disaster at either site. Therefore, an active-active configuration, in which both sites actively serve independent workloads, is intrinsic to the product.

An alternative configuration is active-standby or active-passive, in which only one cluster (say, cluster A) hosts application workloads in a steady state, and only one-way switchover, from site A to site B, is required. The nodes in cluster B still require their own mirrored root aggregates and metadata volumes, as described in the section “Configuration Replication Service” later in this document. If requirements later change and workloads are provisioned on cluster B, this change from active-passive to active-active does not require any change to the MetroCluster configuration. Any workloads (SVMs) that are defined at either site are automatically replicated and protected at the other site.

Another supported option is active-passive within the HA pair, so that only one of the two nodes hosts workloads. This option creates a very small configuration in which only a single data aggregate per cluster is required.

MetroCluster preserves the identity of the storage access paths on switchover. Logical interface (LIF) addresses are maintained after switchover, and NFS exports and SMB shares are accessed by using the same IP address. Also, LUNs have the same LUN ID, worldwide port name (WWPN), or IP address and target portal group tag. Because of this preserved identity, the front-end network must span both sites so that front-end clients and hosts can recognize the paths and connections. A layer 2 spanned Ethernet network and a single SAN fabric across the two sites are required.

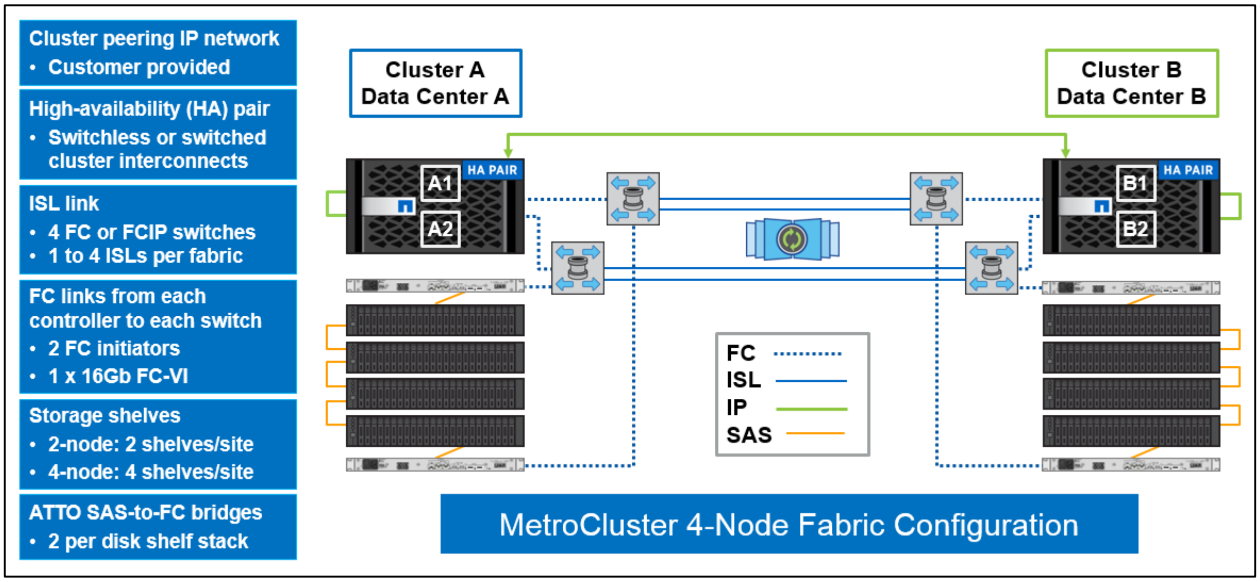
Hardware Configuration

MetroCluster is a fully redundant configuration, with identical hardware required at each site. Figure 3 shows the core components and connections for the four-node configuration. For details about currently supported hardware components, consult the Interoperability Matrix Tool.

In a four-node configuration, each cluster includes the standard clustered Data ONTAP cluster interconnects; typically it is a switchless, back-to-back connection between the two nodes. Four FC or FCIP switches, two at each site, connect to the nodes through both FC initiators and FC-VI connections and connect also to the storage through SAS-to-FC bridges. With this connectivity in place, all nodes in both clusters have visibility to all the storage. When using Brocade 6505 or 6510 FC switches, this configuration has a cluster-to-cluster range of 185 miles (300km).

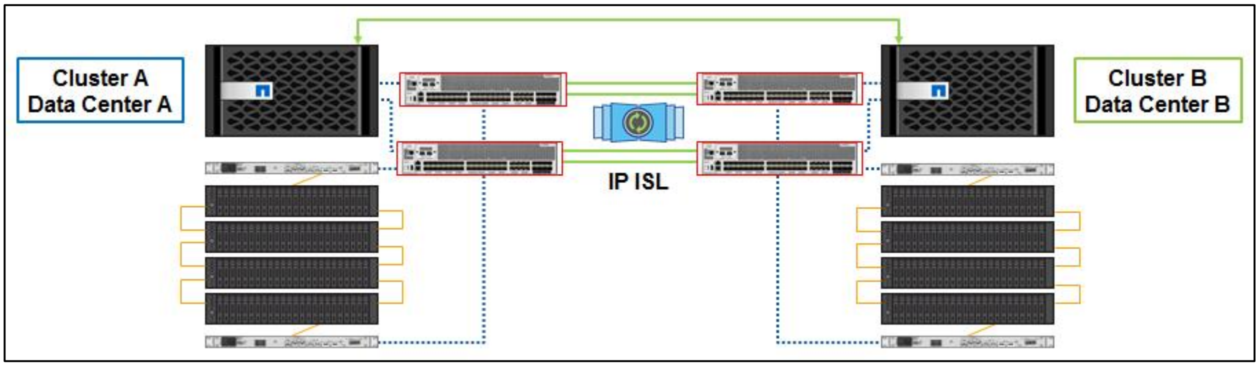
MetroCluster supports both switchless and cluster interconnect switches in clustered Data ONTAP. When using cluster interconnect switches, the HA pairs can use any cluster interconnect switch supported in clustered Data ONTAP. The cluster interconnect switch must be the same at both sites. For the list of supported cluster interconnect switches, see the Hardware Universe.

Figure 3) Cabling diagram for MetroCluster four-node configuration.



In an FCIP configuration, MetroCluster uses an IP ISL to connect to the remote MetroCluster cluster. This configuration uses four Cisco MDS 9250i FCIP switches, two at each site, to connect to the nodes through both FC initiators and FC-VI connections and to connect to the storage through SAS-to-FC bridges. With this connectivity in place, all nodes in both clusters have visibility to all the storage. FCIP configurations have a cluster-to-cluster range of 125 miles (200km). Figure 4 shows the FCIP configuration.

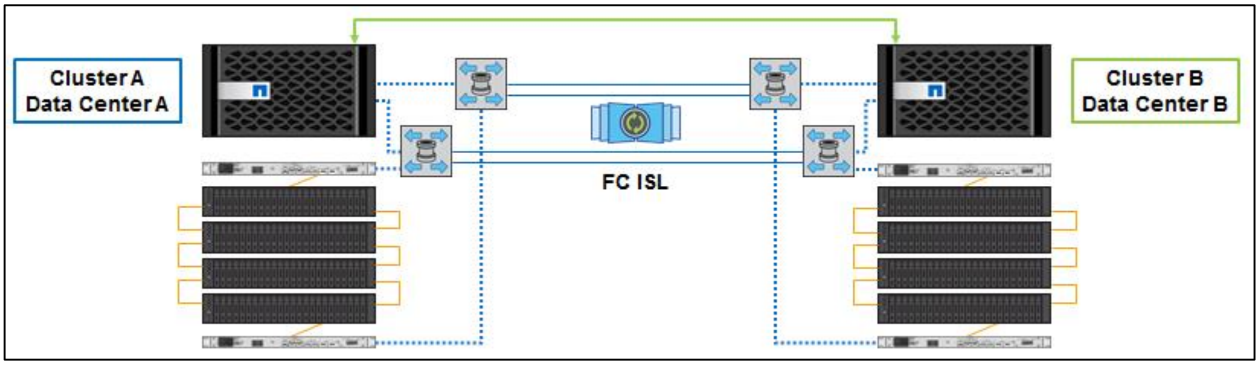
Figure 4) MetroCluster FCIP configuration in clustered Data ONTAP 8.3.2.



In a two-node configuration, for each (single-node) cluster, the standard cluster interconnects are not required. The hardware requirements for a two-node configuration are different for stretch and fabric implementations.

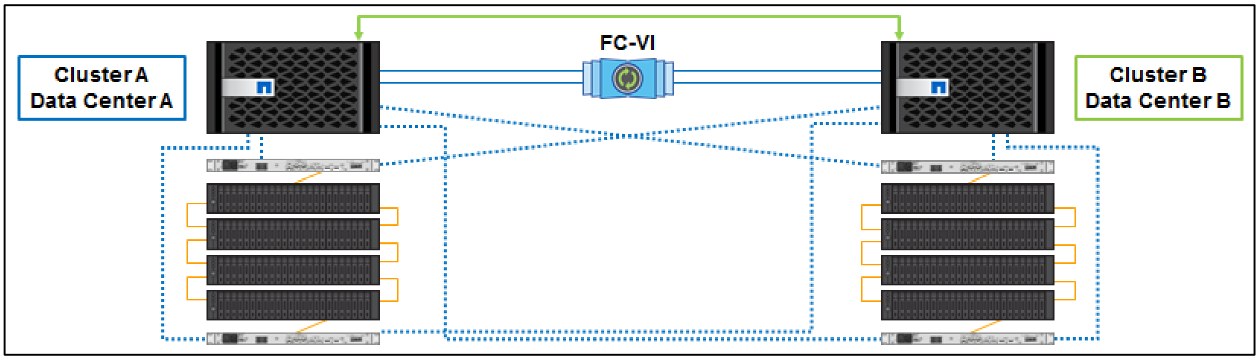
For a two-node fabric-attached configuration, four FC or FCIP switches, two at each site, connect to the nodes through both FC initiators and FC-VI connections and connect to the storage through SAS-to-FC bridges. With this connectivity in place, all nodes in both clusters have visibility to all the storage. When using Brocade 6505 or 6510 FC switches, this configuration has a cluster-to-cluster range of 185 miles (300km). Figure 5 shows the two-node fabric-attached configuration.

Figure 5) Cabling diagram for a MetroCluster two-node fabric-attached configuration.



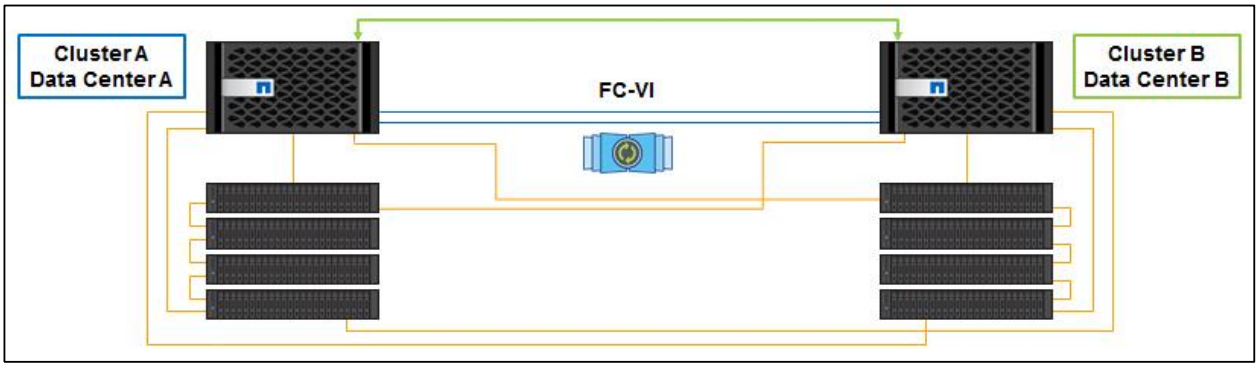
For a two-node stretch bridge-attached configuration, SAS-to-FC FibreBridges provide connectivity to the nodes. No FC or FCIP switches are required. All connectivity to the storage uses FC cables. All nodes in both clusters have visibility to all the storage. This configuration has a cluster-to-cluster range of 1,640 feet (500m) when using the ATTO 6500N FibreBridge and a cluster-to-cluster range of 885 feet (270m) with the ATTO 7500N FibreBridge. Figure 6 shows the two-node bridge-attached configuration.

Figure 6) MetroCluster two-node bridge-attached configuration.



For a two-node stretch direct-attached configuration, no FC or FCIP switches or SAS-to-FC bridges are required. All connectivity to the storage uses extended optical SAS or optical patch panel cables. All nodes in both clusters have visibility to all the storage. This configuration has a cluster-to-cluster range of 1,640 feet (500m). Figure 7 shows the two-node direct-attached configuration.

Figure 7) Cabling diagram for a MetroCluster two-node direct-attached configuration.



The maximum distance (up to 1,640 feet [500m]) for stretch MetroCluster configurations depends on the SFP speed and type. Table 1 shows the maximum distances for stretch MetroCluster configurations.

Table 1) Stretch MetroCluster maximum distance for 16Gbps SFPs.

Speed (Gbps)	Maximum distance (m)			
	16Gbps SW SFP			**16Gbps LW SFP**
	OM2	**OM3**	**OM3+/OM4**	**Single-Mode (SM) Fiber**
2	N/A	N/A	N/A	N/A
4	150	270	270	500
8	50	150	170	500
16	35	100	125	500

For updates to the maximum supported distances for stretch MetroCluster configurations, see the Interoperability Matrix Tool.

Table 2 describes the individual components in more detail.

Table 2) Required hardware components.

Component	Description
Two clustered Data ONTAP clusters: • Four-node: four controllers • Two-node: two controllers	One cluster is installed at each MetroCluster site. All controllers in both clusters must be the same FAS model, both within the HA pair (four-node) and across both sites. Each controller requires a 16GbFC-VI card (two ports, one connection to each local switch) and four FC initiators (8Gb or 16Gb: two connections to each local switch). FAS, AFF and V-Series controllers are supported.
Four FC switches (supported Brocade or Cisco models): • Not required for two-node direct-attached or bridge-attached configurations	The four switches are configured as two independent fabrics with dedicated ISLs between the sites for redundancy. A minimum of one ISL per fabric is required, and up to four ISLs per fabric are supported to provide greater throughput and resiliency. When more than one ISL fabric is configured, trunking is used. All switches must be purchased from and supported by NetApp.
Two SAS-to-FC bridges (ATTO 6500N or 7500N FibreBridges) per storage stack, except if storage arrays (array LUNs) are used: • Not required for two-node direct-attached configurations	The bridges connect the SAS shelves to the local FC or FCIP switches and bridge the protocol from the SAS to FC because only SAS shelves are supported. The FibreBridge is used only to attach NetApp disk shelves; storage arrays connect directly to the switch.

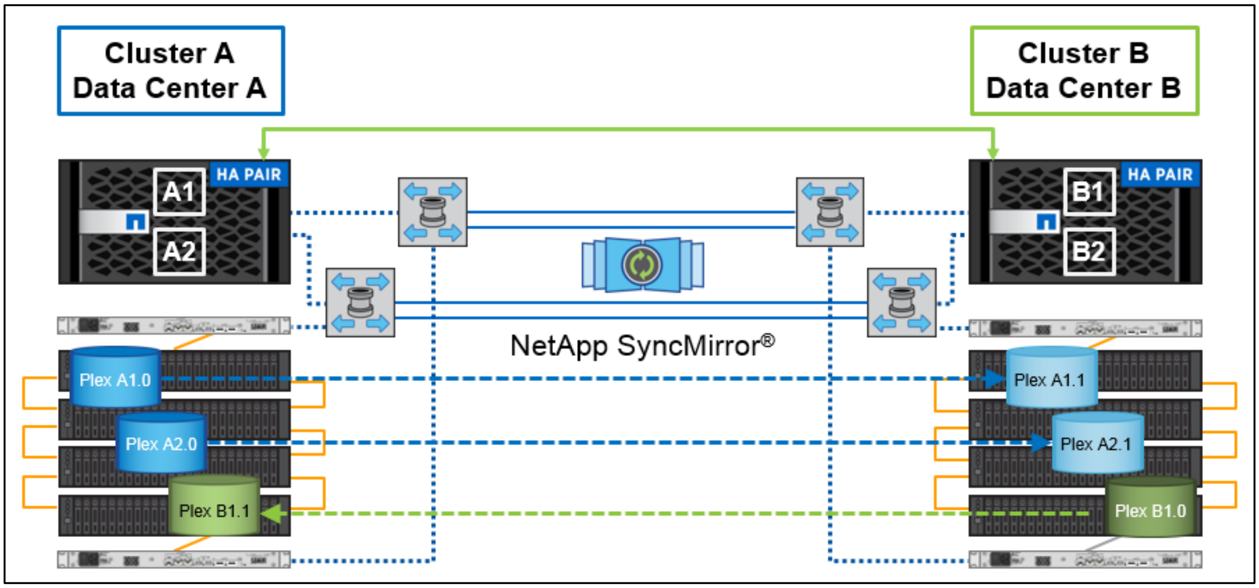
Component	Description
Recommended minimum SAS disk shelves per site (or equivalent storage array disks [array LUNs]) • Four-node: four disk shelves • Two-node: two disk shelves	The storage configuration must be identical at each site. In a four-node configuration, NetApp strongly recommends a minimum of four shelves at each site for performance and capacity and to allow disk ownership on a per-shelf basis. In a two-node configuration, NetApp recommends a minimum of two shelves per site. A minimum of two shelves (four-node configuration) or one shelf (two-node configuration) at each site is supported, but NetApp does not recommend it. See the Interoperability Matrix Tool for supported storage, number of shelves supported in a stack, and storage type mixing rules. All storage in the MetroCluster system must be visible to all nodes. All aggregates, including the root aggregates, must be created on the shared storage.

Disk Assignment

Before MetroCluster is installed, disks must be assigned to the appropriate pool. Each node has both a local pool (at the same site as the node) and a remote pool (at the other site). These pools are used to assign disks to the aggregate's mirrored plexes. See section 2.1 for more information about how aggregates are assigned to pools and across the shelves.

In the four-node MetroCluster configuration, there are a total of eight pools: a local (`Pool0`) and a remote (`Pool1`) pool for each of the four nodes, as shown in Table 3. Cluster A local pools and cluster B remote pools are located at site A. Cluster B local pools and cluster A remote pools are located at site B. Disk ownership is assigned so that node A1 owns all the disks in both its pools, and so on for the other nodes. This configuration is shown in Figure 8: Disks owned by cluster A are shown in blue, and disks owned by cluster B are shown in green.

Figure 8) MetroCluster four-node configuration local and remote pool layout.



In the recommended minimum configuration of four shelves at each site, each shelf contains disks from only one pool. This configuration allows per-shelf disk ownership assignment on original setup and automatic ownership of any failed disk replacements. If shelves are not dedicated to pools, manual disk ownership assignment is required during initial installation and for any subsequent failed disk replacements. NetApp recommends that each shelf in the entire MetroCluster configuration (across both sites) have a unique shelf ID. Table 3 shows the shelf assignments that NetApp recommends.

Table 3) Recommended shelf numbering schema.

Shelf ID Site A	Usage	Shelf ID Site B	Usage
Shelves 10 to 19	`A1:Pool0`	Shelves 20 to 29	`A1:Pool1`
Shelves 30 to 39	`A2:Pool0`	Shelves 40 to 49	`A2:Pool1`
Shelves 60 to 69	`B1:Pool1`	Shelves 50 to 59	`B1:Pool0`
Shelves 80 to 89	`B2:Pool1`	Shelves 70 to 79	`B2:Pool0`

To display the disks and pool assignments, use the following command. Storage stack 1 is at site A, and storage stack 2 is at site B.

```
tme-mcc-A: > disk show -fields home, pool
disk    home       pool
------- ---------- -----
1.10.0   tme-mcc-A1 Pool0
1.10.1   tme-mcc-A1 Pool0
1.10.2   tme-mcc-A1 Pool0
1.10.3   tme-mcc-A1 Pool0
…. <disks omitted>
1.30.0   tme-mcc-A2 Pool0
1.30.1   tme-mcc-A2 Pool0
1.30.2   tme-mcc-A2 Pool0
1.30.3   tme-mcc-A2 Pool0
…. <disks omitted>
1.60.0   tme-mcc-B1 Pool1
1.60.1   tme-mcc-B1 Pool1
1.60.2   tme-mcc-B1 Pool1
1.60.3   tme-mcc-B1 Pool1
…. <disks omitted>
1.80.0   tme-mcc-B2 Pool1
1.80.1   tme-mcc-B2 Pool1
1.80.2   tme-mcc-B2 Pool1
1.80.3   tme-mcc-B2 Pool1
…. <disks omitted>
2.20.0  tme-mcc-A1 Pool1
2.20.1  tme-mcc-A1 Pool1
2.20.2  tme-mcc-A1 Pool1
2.20.3  tme-mcc-A1 Pool1
…. <disks omitted>
2.40.0  tme-mcc-A2 Pool1
2.40.1  tme-mcc-A2 Pool1
2.40.2  tme-mcc-A2 Pool1
2.40.3  tme-mcc-A2 Pool1
…. <disks omitted>
2.50.0  tme-mcc-B1 Pool0
2.50.1  tme-mcc-B1 Pool0
2.50.2  tme-mcc-B1 Pool0
2.50.3  tme-mcc-B1 Pool0
…. <disks omitted>
2.70.0  tme-mcc-B2 Pool0
2.70.1  tme-mcc-B2 Pool0
2.70.2  tme-mcc-B2 Pool0
2.70.3  tme-mcc-B2 Pool0
…. <disks omitted>
```

Disk Ownership

Controllers are shipped from manufacturing with a default disk ownership assignment. Verify this assignment and adjust it for the desired node-to-disk layout in maintenance mode before the clusters are created so that the correct DR partner is chosen for each node. For more information, see section 2.2.

Disk ownership is updated temporarily during an HA failover or DR switchover. Clustered Data ONTAP needs to track which controller owns a particular disk and save its original owner so that ownership can be restored correctly after the corresponding giveback or switchback. To enable this tracking, MetroCluster introduces a new field, `dr-home`, for each disk, in addition to the `owner` and `home` fields. The `dr-home` field is set only after switchover and identifies a disk in an aggregate that has been switched over from the partner cluster. Table 4 shows how the fields change during the different events.

Table 4) Disk ownership changes.

	Value During			
Field	Normal Operation (All Nodes Up)	Local HA Failover (4-Node Configuration)	MetroCluster Switchover	HA Failover After Switchover
`owner`	Name of the node that has access to the disk	Name of the HA partner that temporarily has access to the disk	Name of the DR partner that temporarily has access to the disk	Name of the DR partner's HA partner that temporarily has access to the disk while in switchover
`home`	Name of the original owner of the disk within the cluster	Name of the original owner of the disk within the HA pair	Name of the DR partner	Name of the DR partner
`dr-home`	Unassigned	Unassigned	Name of the original owning node	Name of the original owning node

Table 5 shows the ownership changes of a disk in the A1 remote pool, `Pool1`. The disk is physically located on site B, but is owned in normal operation by node A1.

Table 5) Example of disk ownership changes.

MetroCluster State	Value of Ownership Fields			Notes
	`owner`	`home`	`dr-home`	
Normal operation: all nodes up	A1	A1	Unassigned	N/A
Local HA failover: A1 ➔ A2	A2	A1	N/A	**Note:** A2 takes over A1, so it has temporary ownership of A1's disks. After giveback to A1, disk ownership returns to normal operation.

MetroCluster State	Value of Ownership Fields			Notes
	owner	home	dr-home	
Site switchover: A1/A2 ➔ B1/B2	B1	B1	A1	**Note:** The original owning node name is saved in `dr-home`. B1 now owns A1's resources. After switchback to site A, disk ownership returns to normal operation.
Site switchover followed by HA takeover: A1/A2 ➔ B1/B2 B1 ➔ B2	B2	B1	A1	**Notes:** As the sole surviving node, B2 now owns A1's resources. Recovery from this scenario is a two-step process: 1. B2 gives back to B1 and restores `owner` to B1 (same state as the site switchover line). 2. Site B switches back to site A. Ownership returns to the normal operation state with `dr-home` once again unassigned.

1.5 MetroCluster Replication

Synchronous replication is the mechanism for providing a zero-RPO solution of the data between the two sites. With synchronous replication, changes made at one site are automatically propagated to the other site. The MetroCluster configuration provides replication at three levels: storage replication with NetApp SyncMirror® software, NVRAM replication, and cluster configuration replication.

SyncMirror Storage Replication

A clustered Data ONTAP system stores data in NetApp FlexVol® volumes that are provisioned from aggregates. Each aggregate contains a NetApp WAFL® (Write Anywhere File Layout) file system. In a configuration without MetroCluster, the disks in each aggregate consist of a single group, or *plex*. The plex resides in local storage attached to the controller.

In a MetroCluster configuration, each aggregate consists of two plexes that are physically separated: a local plex and a remote plex. All storage is shared and is visible to all the controllers in the MetroCluster configuration. The local plex must contain only disks from the local pool (`Pool0`), and the remote plex must contain only disks from the remote pool. The local plex is always `plex0`. Each remote plex has a number other than `0` to indicate that it is remote (for example, `plex1` or `plex2`).

Only mirrored aggregates can be created with MetroCluster; unmirrored aggregates are not supported. The `-mirror true` flag must therefore be used when creating aggregates after MetroCluster has been configured; if it is not specified, the `create` command fails. The number of disks that are specified by the `-diskcount` parameter is automatically halved. For example, to create an aggregate with 6 usable disks, 12 must be specified as the disk count. That way, the local plex is allocated 6 disks from the local pool and the remote plex is allocated 6 disks from the remote pool. The same process applies when adding disks to an aggregate: twice the number of disks must be specified as are required for capacity.

The following example shows how the disks have been assigned to the aggregates. Each node has a root aggregate and one data aggregate. Each root aggregate contains 8 drives, of which 4 are on the local plex and 4 are on the remote plex. Therefore, the available capacity of the aggregate is 4 drives. Similarly, each of the data aggregates contains 10 drives: 5 local and 5 remote.

```
tme-mcc-A::> storage aggr show -fields diskcount, plexes
aggregate            diskcount plexes
-------------------- --------- -------------------------------------------------------
aggr0_tme_A1         8         /aggr0_tme_mcc_A1/plex0,/aggr0_tme_mcc_A1/plex1
aggr0_tme_A2         8         /aggr0_tme_mcc_A2/plex0,/aggr0_tme_mcc_A2/plex1
aggr1_tme_A1         10        /aggr1_tme_mcc_A1/plex0,/aggr1_tme_mcc_A1/plex1
aggr1_tme_A2         10        /aggr1_tme_mcc_A2/plex0,/aggr1_tme_mcc_A2/plex1

tme-mcc-B::> storage aggr show -fields diskcount, plexes
aggregate            diskcount plexes
-------------------- --------- -------------------------------------------------------
aggr0_tme_B1         8         /aggr0_tme_mcc_B1/plex0,/aggr0_tme_mcc_B1/plex1
aggr0_tme_B2         8         /aggr0_tme_mcc_B2/plex0,/aggr0_tme_mcc_B2/plex1
aggr1_tme_B1         10        /aggr1_tme_mcc_B1/plex0,/aggr1_tme_mcc_B1/plex1
aggr1_tme_B2         10        /aggr1_tme_mcc_B2/plex0,/aggr1_tme_mcc_B2/plex1
```

In normal MetroCluster operation, both plexes are updated simultaneously at the RAID level. All writes, whether from client and host I/O or cluster metadata, generate two physical write operations: one to the local plex and one to the remote plex, using the ISL connection between the two clusters. Reads by default are fulfilled from the local plex.

MetroCluster requires a minimum number of root and data disks in an aggregate for SyncMirror storage replication. The required number of root and data disks depends on the RAID configuration and on the disk drive capacity. As for any Data ONTAP configuration, having more drives in an aggregate is likely to give better performance, and aggregates with few drives do not perform optimally.

For RAID 4 data aggregates, the minimum number of disks is 3 (2 data, 1 parity). For RAID DP, the minimum number of disks is 5 (3 data, 2 parity).

Plex Read Behavior

By default, all reads are from the local plex. A RAID option can be set so that read operations alternate between the local and remote plexes. Some workloads might experience a performance increase when reading from both plexes, particularly if the sites are only a short distance apart. The RAID option can be changed nondisruptively, without affecting application I/O. To set the option to read from alternate plexes, use the following command on each node in the HA pair:

```
storage raid-options modify -node <node-name> -name raid.mirror_read_plex_pref -value alternate
```

To set the option back to the default value, specify `–value local` on the same command.

NetApp strongly recommends that the option has the same setting on both nodes of an HA pair. There might be particular workload environments in which different read behavior on each node gives optimal performance. If the setting is different between the nodes, the value does not propagate to the other node in the event of an HA failover.

Aggregate Snapshot Copies

Automatic aggregate NetApp Snapshot® copies are taken, and, by default, 5% of the aggregate capacity is reserved for these Snapshot copies. These Snapshot copies are used as the baseline for resyncing the aggregates when necessary.

If one plex becomes unavailable (for example, because of a shelf or storage array failure), the unaffected plex continues to serve data until the failed plex is restored. The plexes are automatically resynchronized when the failing plex is repaired so that both plexes are consistent. The type of resync is automatically determined and performed: If both plexes share a common aggregate Snapshot copy, this Snapshot copy is used as the basis for a partial resync. If there is no common Snapshot copy shared between the plexes, a full resync is performed.

NVRAM Cache Mirroring

In a clustered Data ONTAP HA pair, each node mirrors its NVRAM to the other node by using the HA interconnect. NVRAM is split into two segments, one for each node's NVRAM. MetroCluster provides additional mirroring; each node has a DR partner node on the other site, and the NVRAM is mirrored to the DR partner over the ISL connection. Therefore, in a four-node configuration, each node's NVRAM is mirrored twice—to the HA partner and the DR partner—and each node's NVRAM is split into four segments.

Writes are staged to NVRAM before being committed to disk. A write operation is acknowledged as complete to the issuing host or application after all NVRAM segments have been updated. In a four-node configuration, this update includes the local NVRAM, the HA partner's NVRAM, and the DR partner's NVRAM. Updates to the DR partner's NVRAM are transmitted over the FC-VI connection through the ISL. FC-VI traffic is prioritized over storage replication by using switch QoS.

If the ISL latency increases, write performance can be affected because it takes longer to acknowledge the write to the DR partner's NVRAM. If all ISLs are down, or if the remote node does not respond after a certain time, the write is acknowledged as complete anyway. In that way, continued local operation is possible in the event of temporary site isolation. The remote NVRAM mirror resynchronizes automatically when at least one ISL becomes available. For more information about the scenario in which all the ISLs have failed, see section 3.1.

NVRAM transactions are committed to disk through a consistency point at least once every 10 seconds. When a controller boots, WAFL always uses the most recent consistency point on disk. This approach eliminates the need for lengthy file system checks after a power loss or system failure. The storage system uses battery-backed-up NVRAM to avoid losing any data I/O requests that might have occurred after the most recent consistency point. If a takeover or a switchover occurs, uncommitted transactions are replayed from the mirrored NVRAM, preventing data loss.

The `metrocluster interconnect show` command displays each node's NVRAM mirroring partners. The following output is for cluster B in a four-node configuration:

- Node B1: HA partner is B2, DR partner is A1.
- Node B2: HA partner is B1, DR partner is A2.

```
tme-mcc-B::> metrocluster interconnect show
                          Mirror   Mirror
                  Partner Admin    Oper
Node Partner Name Type    Status   Status  Adapter        Type   Status
---- ------------ ------- -------- ------- -----------    ------ ------
tme-mcc-B1
     tme-mcc-B2
                  HA      enabled  online
                                           cxgb3_0        iWARP  Up
                                           cxgb3_0        iWARP  Up
     tme-mcc-A1
                  DR      enabled  online
                                           fcvi_device_0  FC-VI  Up
                                           fcvi_device_1  FC-VI  Up
tme-mcc-B2
     tme-mcc-B1
                  HA      enabled  online
                                           cxgb3_0        iWARP  Up
                                           cxgb3_0        iWARP  Up
     tme-mcc-A2
                  DR      enabled  online
                                           fcvi_device_0  FC-VI  Up
                                           fcvi_device_1  FC-VI  Up
```

When issued from cluster A, the command output shows:

- Node A1: HA partner is A2, DR partner is B1.
- Node A2: HA partner is A1, DR partner is B2.

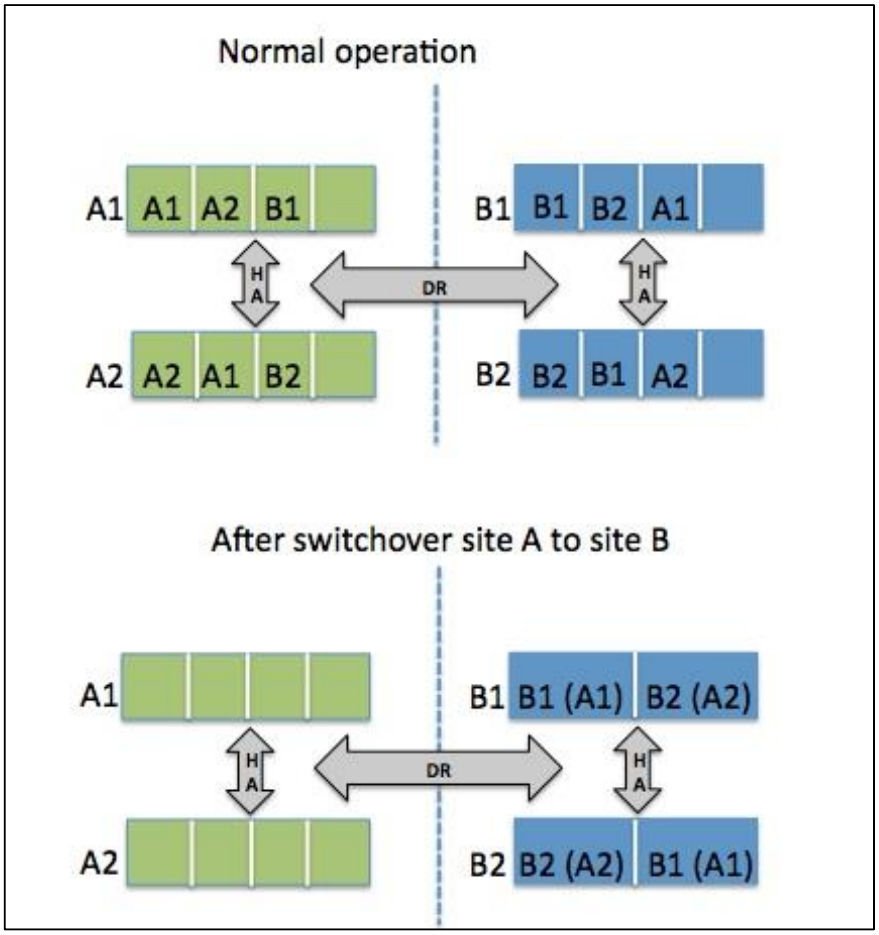
The NVRAM is split into the required four segments when the HA state is set to `mcc` as part of the MetroCluster installation. The `ha-config modify controller` and `ha-config modify chassis` commands are used as described in the MetroCluster Installation and Configuration Guide. In a four-node configuration, this variable must be set to `mcc`. In the two-node configuration, the variable must be set to `mcc-2n`.

In normal operation, NVRAM is allocated as follows:

- Segment 1: node's own NVRAM
- Segment 2: node's HA partner NVRAM
- Segment 3: node's DR partner NVRAM
- Segment 4: unused

The fourth segment is used in the event of switchover operations. For example, Figure 9 shows the layout of NVRAM on the four nodes before and after a switchover. After the switchover, the NVRAM uncommitted transactions from A1 and A2 are replayed in cluster B. The local NVRAM sections are extended to include the sections that were replayed, with the result that site B's nodes log the NVRAM transactions for all the aggregates.

Figure 9) NVRAM segment assignment before and after switchover.



Configuration Replication Service

A MetroCluster configuration consists of two clustered Data ONTAP clusters. Each cluster maintains its own metadata or configuration information, in an internal, highly resilient data structure known as the *replicated database* (RDB). Because each cluster has its own RDB, there is isolation and protection against propagation of errors from one cluster to the other.

When switchover occurs, the stopped cluster's metadata objects (SVMs, including their associated protocol information, volumes, LUNs, export policies, and so on) are activated on the surviving cluster so that the storage services continue to be available. This process means that these objects must have been previously transferred from the owning cluster to the other cluster, ready for activation when needed. A mechanism is required to transfer new and changed configuration objects from one cluster's RDB to the other RDB and to keep this information synchronized. The mechanism used for this transfer has three components:

- **Cluster peering.** This component uses the same peering method and intercluster LIFs as are used with clustered Data ONTAP SnapMirror® and SnapVault® software with intercluster LIFs. The connection between the two clusters is a customer-supplied TCP/IP connection, known as the *configuration replication network*. For reliability, NetApp highly recommends redundant IP connections with two intercluster LIFs per node.
- **Configuration replication service (CRS).** This service runs on each cluster. CRS replicates required metadata objects from the owning cluster and stores them in the other cluster's RDB.
- **Metadata volumes (MDVs).** MDVs are staging volumes for cluster metadata information. Two volumes, each 10GB in size, are created on each cluster when MetroCluster is configured. Each volume must be created on a separate **nonroot** aggregate; therefore, at least two data aggregates must exist on each cluster before you configure MetroCluster. For resiliency, NetApp highly recommends that each data aggregate be on a separate node.

It is a core feature of MetroCluster that changes to the configuration of one cluster **automatically** propagate to the other cluster so that switchover is achieved with zero data or configuration loss. (Example configuration changes include creating a new SVM or a new LIF, or provisioning a volume or LUN in an existing SVM.)

Because the update is automatic, almost no ongoing administration is required that is specific to a MetroCluster configuration. No administrator action is required as workloads are added to continue automatic synchronous protection, and it is not possible to forget to protect a newly added or changed storage resource. Whenever an object is created or updated, the information relating to that transaction is logged in an MDV on the cluster that is performing the update. Changes are not committed to the local RDB until the logging is complete. Updates are propagated near synchronously to the other cluster's RDB over the configuration replication network.

If changes cannot be propagated because of temporary errors in the configuration replication network, the changes are automatically sent to the other cluster after connectivity is restored. Changes are sent by replaying the logged transactions from the MDV. This catch-up in the configuration is automatic. If a forced switchover is necessary when the network is down and there are unpropagated RDB updates, the updates are processed from the mirrored copy of the MDV at the surviving site. To promote resiliency, NetApp recommends redundant networks for the cluster configuration network.

The MDVs are given system-assigned names and are visible on each cluster, as shown in the following example. Because the command was issued from cluster A, the first two volumes that are listed are the local MDVs, with the state of `online`. The second two MDVs belong to cluster B (note their hosting aggregate) and are offline unless a switchover is performed.

```
tme-mcc-A::> volume show -volume MDV*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
tme-mcc-A MDV_CRS_cd7628c7f1cc11e3840800a0985522b8_A
                       aggr1_tme_A1 online     RW         10GB     9.50GB    5%
tme-mcc-A MDV_CRS_cd7628c7f1cc11e3840800a0985522b8_B
                       aggr1_tme_A2 online     RW         10GB     9.50GB    5%
tme-mcc-A MDV_CRS_e8fef00df27311e387ad00a0985466e6_A
                       aggr1_tme_B1 -          RW            -          -     -
tme-mcc-A MDV_CRS_e8fef00df27311e387ad00a0985466e6_B
```

```
                        aggr1_tme_B2 -           RW              -          -     -
```

Created objects are automatically propagated to the other cluster over the cluster peering network by using CRS. Note that job schedules are not propagated, for example, to associate with a Snapshot policy. The following example shows the message received when creating or modifying a job schedule in a MetroCluster configuration. The same job schedule command **must** be run on the other cluster. If you associate a schedule that does not exist on both clusters with a Snapshot policy and then apply that policy to a volume, the replication is out of sync.

```
tme-mcc-A::> job schedule cron create -name Monday -hour 8,10 -minute 00 -dayofweek Monday

Warning: Because this is a MetroCluster configuration, an additional step is required. To
complete applying the changes for schedule "Monday", execute the same command on the remote
cluster.
```

2 Initial MetroCluster Setup

A NetApp MetroCluster solution consists of a combination of hardware and software. Specific hardware is required to create the shared storage fabric and intersite links. For supported hardware, consult the NetApp Interoperability Matrix Tool. On the software side, MetroCluster is completely integrated into the NetApp clustered Data ONTAP operating system. No separate tools or interfaces are required.

After the MetroCluster relationships have been established, data and configuration are automatically continuously replicated between the sites, so manual effort is not required to establish replication of newly provisioned storage. This capability not only simplifies the administrative effort required, but it also eliminates the possibility of forgetting to replicate storage for critical workloads.

Setup instructions for MetroCluster configurations are available in two product publications:

- The MetroCluster Installation and Configuration Guide provides worksheets and detailed instructions to configure your MetroCluster components. Follow the procedures closely and, as a best practice, use the same naming conventions and port assignments as those contained in this document.
- Alternatively, if the hardware is shipped as configured from the factory with Brocade FC switches and NetApp FAS shelves only (no array LUNs), you can use the MetroCluster Installation Express Guide. The Express Guide contains minimal background information and few options. If you're in doubt, use the MetroCluster Installation and Configuration Guide.

2.1 Hardware and Software Requirements

Requirements for Clustered Data ONTAP

The following information applies to four-node MetroCluster configurations:

- Four nodes are required, two nodes at each site. The four nodes are known collectively as a *DR group*. All four nodes must be the same FAS or V-Series model (for example, four FAS8020 systems or four FAS8060 systems). FAS and V-Series controllers cannot coexist in the same MetroCluster DR group.
- No additional or specific license is required. MetroCluster functionality, including NetApp SyncMirror, is included in the basic clustered Data ONTAP license. Protocols and other features such as NetApp SnapMirror require licenses if they are used in the cluster. Licenses must be symmetrical across both sites. For example, SMB, if used, must be licensed in both clusters. Switchover does not work unless both sites have the same licenses.
- All nodes should be licensed for the same node-locked features.
- Infinite volumes are not supported in a MetroCluster configuration.
- Advanced disk partitioning (for either the root aggregate or NetApp Flash Pool™ aggregates) is not supported.

- NetApp Storage Encryption (NSE) drives are not supported in a MetroCluster configuration. Clustered Data ONTAP systems can attach to NetApp E-Series storage arrays with full disk encryption (FDE) drives. The root aggregate can reside on encrypted drives. For the list of systems that support this configuration, see the NetApp Interoperability Matrix Tool.

The following information applies to two-node MetroCluster configurations:

- Two nodes are required, one node at each site. The two nodes are known collectively as both the *HA pair* and the *DR group*. Both nodes must be the same FAS or V-Series model (for example, two FAS8020 systems or two FAS8060 systems). FAS and V-Series controllers cannot coexist in the same MetroCluster DR group.
- No additional or specific license is required. MetroCluster functionality, including SyncMirror, is included in the basic clustered Data ONTAP license. Protocols and other features such as SnapMirror require licenses if they are used in the cluster. Licenses must be symmetrical across both sites. For example, SMB, if used, must be licensed in both clusters. Switchover does not work unless both sites have the same licenses.
- All nodes should be licensed for the same node-locked features.
- Infinite volumes are not supported in a MetroCluster configuration.
- Advanced disk partitioning (for either the root aggregate or Flash Pool aggregates) is not supported.
- NSE drives are not supported in a MetroCluster configuration. Clustered Data ONTAP systems can attach to NetApp E-Series storage arrays with FDE drives. The root aggregate can reside on encrypted drives. For the list of systems that support this configuration, see the NetApp Interoperability Matrix Tool.

Requirements for MetroCluster Configurations with Array LUNs

The following are requirements for setting up a MetroCluster configuration with array LUNs:

- The platforms and storage array must be listed in the Interoperability Matrix Tool as supported for MetroCluster configurations.

  Note: The Interoperability Matrix Tool contains details of MetroCluster configurations that use array LUNs. It includes information about supported storage arrays, switches, NetApp controllers, and clustered Data ONTAP versions that are supported for use with array LUNs. The Interoperability Matrix Tool is the authority for requirements and restrictions for MetroCluster configurations that use array LUNs. In the tool, select V-Series and FlexArray Virtualization for Fabric MetroCluster as the storage solution.

- All the clustered Data ONTAP systems in a MetroCluster configuration must be of the same model.
- Sharing of multiple FC initiator ports with a single target port is not supported in a MetroCluster configuration. Similarly, sharing of multiple target ports with a single FC initiator port is also not supported.
- Additional ports are required when mixing FAS and array LUN disk shelves.
- The FlexArray Virtualization Implementation Guide for Third-Party Storage and the FlexArray Virtualization Implementation Guide for E-Series Storage contain additional details about the supported storage array families. The storage arrays in the MetroCluster configuration must be symmetric, which means the following:
  - The two storage arrays must be from the same vendor family and must have the same firmware version installed.
  - You must have two sets of array LUNs: one set for the aggregate on the local storage array and another set of LUNs at the remote storage array for the mirror of the aggregate. The array LUNs must be of the same size for mirroring the aggregate.
  - The disk types (for example, SATA, SSD, or SAS) that are used for mirrored storage must be the same on both storage arrays.

  - The parameters for configuring storage arrays, such as RAID type and tiering, must be the same across both sites.
  - Effectively, MetroCluster configurations with array LUNs should be completely symmetrical across sites.

Requirements for FC Switches

The following are requirements for setting up a MetroCluster configuration with FC switches:

- The switches and switch firmware must be identified in the Interoperability Matrix Tool as being supported for MetroCluster configurations.
- In the two-node and four-node fabric configuration, each fabric must have two switches, making four switches in total. Two-node bridge- and direct-attached configurations do not require FC switches.
- All switches in the configuration must be the same model and from the same vendor, and they must be licensed for the same number of ports.
- All switches must be ordered and purchased from NetApp and must be dedicated to the MetroCluster configuration.
- Host and application traffic cannot share the same switches as those used for MetroCluster.
- To provide redundancy in case of device and path failures, each clustered Data ONTAP system must be connected to storage by using redundant components.
- Clustered Data ONTAP supports from one to four ISLs per fabric. Trunking is used if there are multiple ISLs per fabric. Verify that the xWDM vendor supports trunking on the FC switch. For a single ISL connection, trunking is not required. NetApp recommends a minimum of two ISLs.
- In-order delivery is the default and the setting that NetApp recommends, regardless of the number of ISLs per fabric. Out-of-order delivery is not supported.

  Note: For more information about basic switch configuration, ISL settings, and FC-VI configurations, see the MetroCluster Installation and Configuration Guide.

Requirements for FCIP Switches

Although FC is the clear choice for a mission-critical, high-performance, low-latency, highly reliable SAN fabric, many modern data centers have invested in IP technologies. FC over IP (FCIP) transparently interconnects FC over IP networks and is an important technology for linking FC SANs.

The FCIP MetroCluster configuration is the same as the FC MetroCluster configuration for all two-node and four-node architectures. The only difference is that the FC switches are replaced with the FCIP switches and an IP ISL. For FCIP MetroCluster deployments, NetApp supports the Cisco MDS 9250i. The Cisco MDS 9250i offers up to forty 16Gbps FC ports, two 1/10Gb Ethernet IP storage service ports, and eight 10Gb FC over Ethernet (FCoE) ports.

The cabling configuration remains the same. Only the Cisco 9250i is supported for IP ISL connectivity with the Data ONTAP 8.3.2 release. The FCIP configuration requires a dedicated IP network for the IP ISL and can be directly attached or can use intermediate switches.

The following are requirements for setting up a MetroCluster configuration with FCIP switches:

- The switch firmware must be identified in the Interoperability Matrix Tool as being supported for MetroCluster configurations.
- In the two-node and four-node fabric configuration, each fabric must have two switches, making four switches in total. Two-node bridge- and direct-attached configurations do not require FCIP switches.
- All switches in the configuration must be Cisco MDS 9250i, and they must be licensed for the same number of ports.
- All switches must be ordered and purchased from NetApp and must be dedicated to the MetroCluster configuration.

- Host and application traffic cannot share the same switches as those used for MetroCluster.
- To provide redundancy in case of device and path failures, each clustered Data ONTAP system must be connected to storage by using redundant components.
- To verify ISL fabric maximums and trunking requirements, see the Interoperability Matrix Tool.

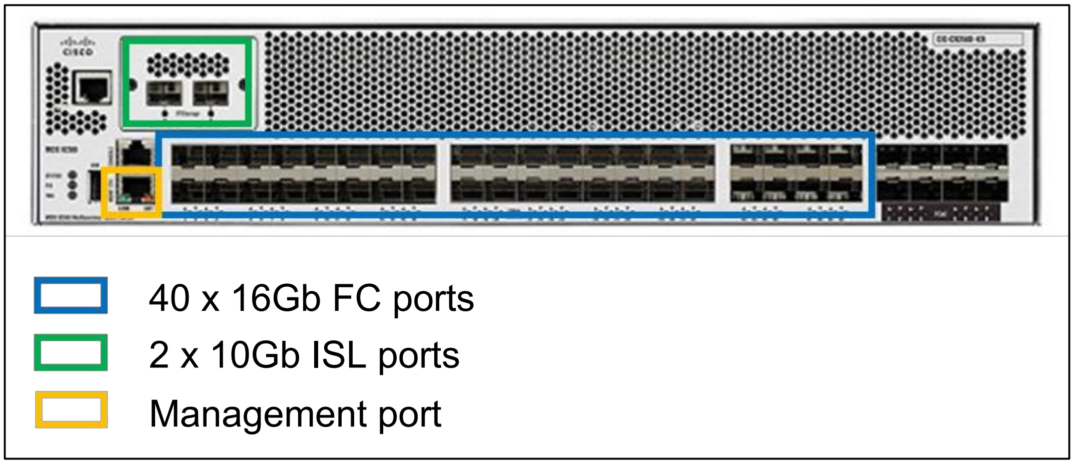
  Note: For more information about basic switch configuration, ISL settings, and FC-VI configurations, see the MetroCluster Installation and Configuration Guide.

Although the FC and FCIP switches perform a similar function, there are a few differences:

- The Cisco MDS 9250i cannot be connected to more than one MetroCluster cluster.
- Only 10Gbps FCIP ISL connectivity is supported.
- Only one FCIP port is supported.
- Link-write acceleration is not supported on the Cisco MDS 9250i.
- The FCIP ISLs between the MetroCluster clusters cannot be shared.

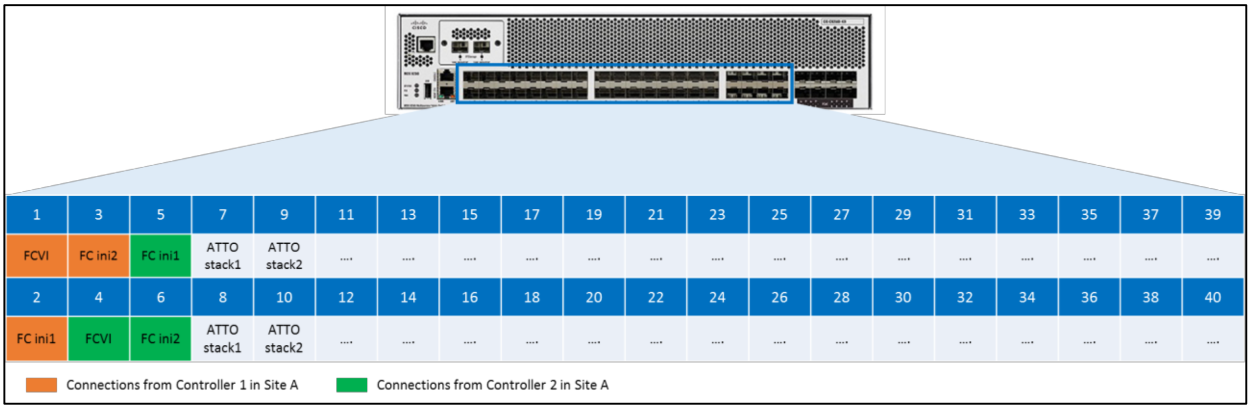
Figure 10 shows the switchports that are configured by factory default on the Cisco MDS 9250i. For more information, see the Cisco MDS 9250i datasheet.

Figure 10) Cisco MDS 9250i FCIP switchports.



The FC ports are assigned in a manner that is similar to the existing fabric MetroCluster configuration. Figure 11 shows the recommended way to configure the FC ports.

Figure 11) Cisco MDS 9250i FCIP switchport assignment.



1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	37	39
FCVI	FC ini2	FC ini1	ATTO stack1	ATTO stack2	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....
2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40
FC ini1	FCVI	FC ini2	ATTO stack1	ATTO stack2	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....

🟧 Connections from Controller 1 in Site A 🟩 Connections from Controller 2 in Site A

Requirements for FibreBridges

All four-node configurations with NetApp FAS shelves require FibreBridges. The two-node fabric and two-node stretch configurations with FibreBridge also require them. Only the two-node configuration with SAS optical cables does not use FibreBridges, because the storage connects directly to SAS ports on the two controllers.

Starting in Data ONTAP 8.3.2 GA, the ATTO 7500N FibreBridge is supported and recommended for attaching NetApp FAS disk shelves. The ATTO 7500N provides improved performance and lower latency than the ATTO 6500N FibreBridge does, and it is ideal for All Flash FAS MetroCluster deployments. In addition, the ATTO 7500N has dual power supplies for redundancy. A FibreBridge is not used for a MetroCluster configuration that uses solely array LUNs. Before Data ONTAP 8.3.2 GA, NetApp recommended the ATTO 6500N FibreBridge for attaching NetApp FAS shelves.

The key advantages of the ATTO 7500N include:

- Dual power supply
- Increased throughput performance
- Two 16Gb FC ports, both can be enabled
- Four 12Gb SAS ports, all four can be enabled
- Support for four SSD disk shelves with 96 SSDs that are connected to a pair of ATTO 7500Ns
- Support for four disk shelf stacks that are connected to a pair of ATTO 7500Ns

The ATTO 6500N and the ATTO 7500N both require two FibreBridges per disk stack. At least one stack is required at each site; therefore, a minimum of four FibreBridges is required. Each FibreBridge connects through one FC port to one switch and through one SAS port to the SAS disk stack. Up to 10 shelves per stack are supported when only HDDs are used. Mixing disk types in the stack is supported as described the storage subsystem requirements.

Table 6) Maximum supported SSD and HDD stack depths for the ATTO 7500N and 6500N.

		Total shelves per FibreBridge pair	
Disk Configuration	Disk Type	ATTO 6500N	ATTO 7500N
All SSDs	SSD	2	4
All HDDs	HDD	10	10

Mixed SSD + HDDs	SSD	1	2	1	2	3	4
	HDD	6	2	9	8	7	6

For more information, check the Hardware Universe and the Clustered Data ONTAP Maximum Configuration Guides.

If SSDs are in the FibreBridge stack, the supported maximum stack depths are shown in Table 6. The total number of shelves includes SSD-only shelves, mixed SSD-HDD shelves, and HDD-only shelves. Up to 96 SSDs can be in any single disk stack, and the SSDs can be distributed across any of the shelves. However, NetApp recommends that you not mix SSDs and HDDs in the same shelf, and that you instead configure SSD-only shelves.

Furthermore, for optimal performance, NetApp recommends that you place SSD-only shelves either in their own dedicated stack (no HDD shelves) or at the top or the bottom of a mixed SSD and HDD stack. In that way, the SSD shelves connect directly to the FibreBridge. In this instance, the SSD shelves likely contain pools from more than one node, so disk ownership must be manually assigned. For all-SSD aggregates, stacks should contain SSDs only, with no more than two SSD shelves in the stack. Optimal performance is achieved with one SSD shelf per stack.

Requirements for Zoning

The following are requirements for setting up zoning for the FC and FCIP switches:

- Single-initiator to single-target zoning must be followed for MetroCluster configurations. Single-initiator to single-target zoning limits each zone to a single FC initiator port.
- FC-VI ports must be zoned end to end across the fabric by using the virtual WWPN. The A ports of the FC-VI cards must be in one zone and the B ports must be in a separate zone.
- Sharing of multiple initiator ports with a single target port is not supported. Similarly, sharing of multiple target ports with a single initiator port is also not supported.

Requirements for SyncMirror and Storage

The following are requirements for the SyncMirror storage:

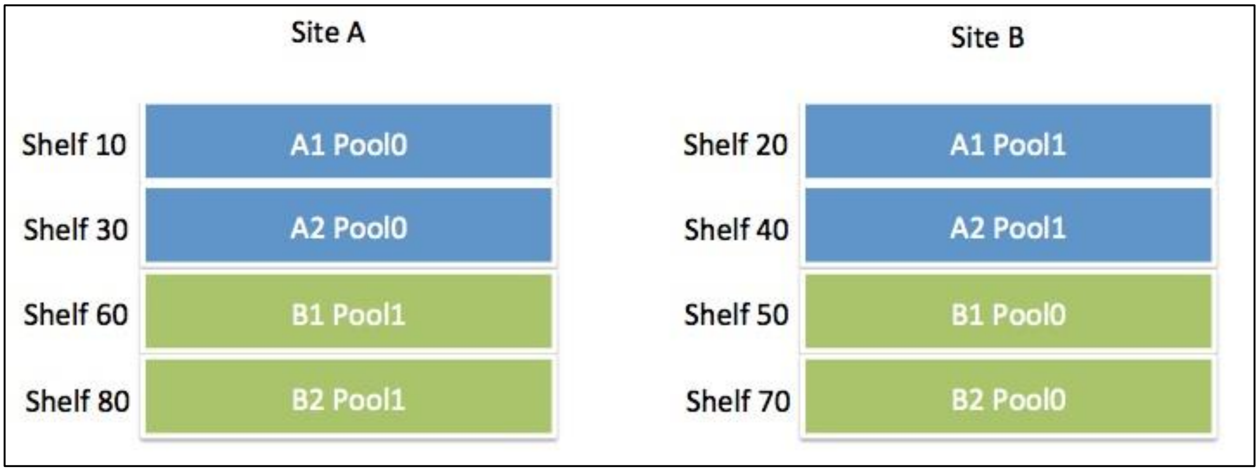
- The disk configuration must be identical between the two sites. This requirement includes NetApp Flash Pool configurations. If Flash Pool is required for a node’s workload, the same capacity of Flash Pool intelligent caching must exist in the mirrored plex for each Flash Pool aggregate.
- All aggregates in a MetroCluster configuration must be mirrored, including the root aggregates on all four controllers.
- RAID 4 and NetApp RAID DP® technology are both supported for the root and data aggregates.

The best practice is a minimum of four shelves per site. With four FAS shelves per site, each pool (local and remote for each node) has its own shelf, and the disks can be assigned on a per-shelf basis to each node. Future disk expansions should be planned so that the pool-shelf isolation is preserved. Although fewer than four shelves per site are supported, it makes disk assignment more complex and also significantly limits the amount of usable storage that is available.

Figure 12 shows a sample layout of pool-to-shelf assignment with shelf IDs assigned as recommended in Table 3) Recommended shelf numbering schema.

Separating pools on a per-shelf basis is not enforced; software disk ownership allows disks in any shelf to be assigned to any node. When multiple nodes own disks on the same shelf, per-shelf disk auto assignment is disabled, and the scope of impact of a shelf failure is also increased.

Figure 12) MetroCluster pool-to-shelf assignment.



If SSDs are included in FibreBridge-attached stacks, see the Interoperability Matrix Tool for the supported maximum stack depths. Up to 48 SSDs can be present in any single disk stack, and the SSDs can be distributed across any of the shelves. However, NetApp recommends that you not mix SSDs and HDDs in the same shelf, and that you instead configure SSD-only shelves.

Further, for optimal performance, NetApp recommends that you place SSD shelves either in their own dedicated stack (no HDD shelves) or at the top or the bottom of a mixed SSD and HDD stack. In that way, the SSD shelves connect directly to the FibreBridge. In this instance, the SSD shelves likely contain pools from more than one node, so disk ownership must be manually assigned. For all-SSD aggregates, stacks should contain SSDs only, with no more than two SSD shelves in the stack. Optimal performance is achieved with one SSD shelf per stack.

After the initial installation, storage can be added one shelf at a time to each cluster; there is no requirement to add multiple shelves. However, the same storage must be added at both sites to preserve mirroring symmetry.

Cabling and Switch Configuration Best Practices

NetApp highly recommends that you use the switch configuration files that are available on the NetApp Support site. To use the configuration files, see the section “Configuring the FC Switches by Running a Configuration File” in the MetroCluster Installation and Configuration Guide in the MetroCluster documentation. The configuration files set up the ports as shown in the table “Reviewing the FC Switch Port Assignments” in the MetroCluster Installation and Configuration Guide.

The recommended assignments have 4 ports reserved for ISLs. Four ports are assigned for storage stacks for all the supported switches, except for the Brocade 6505. On the 6505, 2 ports for storage stacks are assigned, because by default only 12 ports are enabled on the switch, with 1 port on demand (PoD). If you do not use the recommended port assignments, or if you are installing more storage shelves than are included in the standard configuration, then you need to manually configure the switches as per the installation guide.

All FC-VI A ports must be cabled to one fabric, and all FC-VI B ports must be cabled to the other fabric. Make sure that the cabling is done this way, even if you have deviated from the recommended port assignments.

Attach the FC initiators in each controller such that the two attachments to each fabric on a controller use separate ASICs; for example:

`0a/0c`: fabric 1, port 1 and port 2

`0b/0d`: fabric 2, port 1 and port 2

Note: The distributed reference configuration files for the Brocade 6505 switch assign ports 8, 9, 10, and 11 for ISLs and assign 2 ports for storage stacks. If you need more than 2 stacks at a site, and use only 1 or 2 ISLs per fabric, you can use ports 10 and 11 for the ISLs, leaving ports 8 and 9 free for the additional 2 storage stacks. The first 2 storage stacks are already connected on ports 6 and 7. Making this connection requires manual switch configuration, because the supplied configuration files reserve ports 8 to 11 for ISLs. By using ports 8 and 9 for storage stacks, you avoid the need for an additional PoD (12-port) license to support the added storage. If you need all the ISL ports and more than 2 stacks, then an additional PoD license is required. For more information about the options for the 6505, see the pertinent NetApp Knowledge Base article.

2.2 Summary of Installation and Setup Procedure

The MetroCluster Installation and Configuration Guide provides worksheets and instructions to configure your MetroCluster components.

For reference, following is a summary of the key setup and configuration steps for the MetroCluster four-node architecture. The two-node configuration uses a similar setup. The precise tasks that are required depend on whether the system was factory configured before shipment and whether existing equipment is being redeployed (for example, in a transition scenario). Consult the documentation for detailed procedures:

1. Rack the hardware, then install and connect interdevice and intersite cabling (ISLs and cluster peering connections). This step includes the clustered Data ONTAP nodes (FAS or V-Series controllers, cluster management switch, or cluster interconnect switch if used), SAS disk shelves, FC or FCIP switches, and FibreBridges. FibreBridges are not used with array LUNs.
2. Configure the FC or FCIP switches. These switches might have been configured at the factory. If not, NetApp recommends that you use the configuration files from the NetApp Support site. Configuration files are available for the switches that are supported in MetroCluster. If manual configuration is required, follow the steps in the MetroCluster Installation and Configuration Guide.
3. Do not connect the ISL links until directed to do so in the product documentation. If the links are already up, local HA cluster setup does not work.
4. In maintenance mode:
   a. Verify disk ownership. Equipment received from manufacturing should have preassigned disk ownership equally across the nodes. However, if the assignment does not match your requirement, manually reassign the disks before creating the clusters. See the “MetroCluster Installation and Configuration Guide” for a planning worksheet for disk assignment. Each node requires a local pool (`Pool0`, located at the same site as the node) and a remote pool (`Pool1`, located at the other site), with disks owned by that node.

   Note: Each node’s DR partner is automatically selected by the `metrocluster configure` command, based on its system ID (NVRAM ID). The lower-ordered node in one HA pair is DR-partnered with the corresponding lower-ordered node in the other cluster, and so on. The DR partner assignment cannot be changed after the initial configuration. Particularly if the nodes in an HA pair are configured with different storage, check that the corresponding nodes with a lower system ID in each cluster have matching disk configurations. Perform the same step for the two nodes with higher system IDs. Adjust the disk ownership as needed before you continue so that the DR partners are assigned correctly. The assignment of DR partners also affects the ports that LIFs use on switchover. In normal circumstances, LIFs switch over to their DR partner node. For more information, see the section “Networking and LIF Creation Guidelines for MetroCluster Configurations” in the MetroCluster Installation and Configuration Guide.

   b. Verify that the controller and chassis components are set to `mcc-2n` or `mcc`. This step enables NVRAM to be correctly partitioned for replication.
5. In normal mode:
   a. Set up the cluster on each site by using system setup or the CLI `cluster setup` wizard.

b. Create intercluster LIFs and peer the two clusters. Intercluster LIFs can be on dedicated ports or on shared data ports.

c. Mirror the root aggregates.

d. Create a mirrored data aggregate on each node. A minimum of two data aggregates is required in each cluster that hosts the MDVs. This aggregate, created before initial setup, can be used for data (volumes and LUNs); MDVs do not need a dedicated aggregate.

e. Install any additional clustered Data ONTAP feature licenses that are required. Licenses must be symmetrical on both clusters to achieve correct client and host access after switchover. Switchover is vetoed if licenses are not symmetrical.

f. Enable ISLs and zoning.

g. Initialize MetroCluster from one of the nodes by using the command `metrocluster configure –node-name <node-initiating-the-command>`.

h. Add the switches and FibreBridges as monitored devices in the health monitor (`storage switch add` and `storage bridge add` commands). Health monitoring provides information for monitoring and alerting to NetApp OnCommand Unified Manager (see section 4.2 for more details).

i. Verify the MetroCluster configuration with the `metrocluster check run` command. Follow the additional verification steps in the MetroCluster Installation and Configuration Guide.

6. Install OnCommand Unified Manager if it is not already available in the environment. Add the MetroCluster clusters to the managed storage systems.
7. Run Config Advisor against the configuration and check and correct any errors found. Your NetApp or partner representative can provide Config Advisor.
8. NetApp recommends that you install OnCommand Performance Manager to monitor MetroCluster performance.
9. The configuration is now ready for testing. To check HA takeover and giveback and site switchover, healing, and giveback, follow the steps in the MetroCluster Installation and Configuration Guide.

2.3 Postsetup Configuration and Administration

After the MetroCluster configuration is complete, ongoing administration is almost identical to the administration for a clustered Data ONTAP environment without MetroCluster. Configure SVMs with the required protocols and create the LIFs, volumes, and LUNs that are required on the cluster that runs these services in normal operation. You can configure SVMs by using the CLI, OnCommand System Manager, or OnCommand Workflow Automation, as you prefer. These objects are automatically replicated to the other cluster over the cluster peering network.

Aggregate resync NetApp Snapshot copies. Aggregate Snapshot copies are taken at regular intervals for SyncMirror operation. The default interval is 60 minutes. NetApp recommends reducing this interval to 15 minutes. Also, if Flash Pool is used, this interval should be reduced to 5 minutes. See section 4.10 for the command to modify the aggregate Snapshot interval.

The following sequence of commands represents the state of the two clusters after additional aggregates and SVMs have been created on each cluster and have been configured for protocol access. For brevity, CLI output is used, but the same output is visible by using OnCommand System Manager.

Viewing Aggregates

In normal operation, each cluster sees only its own aggregates, as shown in the output from each cluster.

```
tme-mcc-A::> aggr show


Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
```

```
aggr0_tme_A1
            1.38TB   741.9GB   48% online       1 tme-mcc-A1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr0_tme_A2
            1.38TB   741.9GB   48% online       1 tme-mcc-A2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_A1
            2.07TB    2.06TB    1% online       4 tme-mcc-A1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_A2
            2.07TB    2.06TB    0% online       1 tme-mcc-A2       raid_dp,
                                                                   mirrored,
                                                                   normal
```

```
tme-mcc-B::> aggr show


Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
aggr0_tme_B1
            1.38TB   741.9GB   48% online       1 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr0_tme_B2
            1.38TB   741.9GB   48% online       1 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_B1
            2.07TB    2.06TB    1% online       2 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_B2
            2.07TB    2.06TB    1% online       3 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
```

Viewing SVMs

Data SVMs in clustered Data ONTAP 8.3.x are differentiated by the property subtype. Without MetroCluster, the subtype is set to a default value. In MetroCluster, the subtype is either `sync-source` or `sync-destination`. Any data SVM is of type `sync-source` on its owning cluster. The equivalent SVM object that is replicated to the other cluster is of type `sync-destination`, with the suffix `–mc` added to its name. In normal operation, `sync-source` SVMs have the operational state of `running`, and `sync-destination` SVMs have the operational state of `stopped`.

Consider the example output from the `vserver show` command that was run on each of the clusters. The administrator had previously created an SVM on cluster A, `svm1_mccA`, and an SVM on cluster B, `svm1_mccB`.

The cluster A output shows `svm1_mccA`, with type `sync-source`. The entry SVM `svm1_mccB-mc` represents the replicated SVM from cluster B. Therefore, its subtype is `sync-destination`, and it is in a stopped state.

```
tme-mcc-A::> vserver show –type data
                               Admin      Operational Root
Vserver     Type    Subtype    State      State       Volume          Aggregate
----------- ------- ---------- ---------- ----------- ----------      ----------
svm1_mccA   data    sync-source           running     svm1_mccA_root  aggr1_tme_A1
                               running
svm1_mccB-mc
            data    sync-destination      stopped     svm1_mccB_root  aggr1_tme_B1
```

```
                               running
```

The following cluster B output shows the reverse. The running SVM is `svm1_mccB` (`sync-source`). The SVM replicated from cluster A is `svm1_mccA-mc` (`sync-destination`), and it is in a stopped state.

```
tme-mcc-B::> vserver show -type data
                               Admin      Operational Root
Vserver     Type    Subtype    State      State       Volume          Aggregate
----------- ------- ---------- ---------- ----------- ----------      ----------
svm1_mccA-mc
            data    sync-destination      stopped     svm1_mccA_root aggr1_tme_A1
                               running
svm1_mccB   data    sync-source           running     svm1_mccB_root  aggr1_tme_B1
                               running
```

Upon switchover, the surviving cluster starts all `sync-destination` SVMs.

Viewing Volumes

In normal operation, all volumes in data SVMs from both clusters are visible, along with all the MDVs from both clusters, and the root volumes from the local cluster only. Note that on cluster A, only SVM `svm1_mccA`'s volumes are in an online state; the volumes in cluster B's SVM are present but are not accessible. They are brought online only after a switchover. For clarity, MDVs and root volumes are omitted from this output.

```
tme-mcc-A::> vol show
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
svm1_mccA svm1_mccA_lun1_vol
                       aggr1_tme_A1 online     RW       1.24GB    248.2MB   80%
svm1_mccA svm1_mccA_root
                       aggr1_tme_A1 online     RW          1GB    972.5MB    5%
svm1_mccA vol1         aggr1_tme_A1 online     RW          2GB     1.85GB    7%
svm1_mccB-mc
          svm1_mccB_root
                       aggr1_tme_B1 -          RW            -          -     -
svm1_mccB-mc
          vol1         aggr1_tme_B2 -          RW            -          -     -
svm1_mccB-mc
          vol2         aggr1_tme_B2 -          RW            -          -     -
```

On cluster B, the reverse is true. Its SVM volumes are online, and volumes from cluster A's SVM are not.

```
tme-mcc-B::> vol show
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
svm1_mccA-mc
          svm1_mccA_lun1_vol
                       aggr1_tme_A1 -          RW            -          -     -
svm1_mccA-mc
          svm1_mccA_root
                       aggr1_tme_A1 -          RW            -          -     -
svm1_mccA-mc
          vol1         aggr1_tme_A1 -          RW            -          -     -
svm1_mccB svm1_mccB_root
                       aggr1_tme_B1 online     RW          1GB    972.5MB    5%
svm1_mccB vol1         aggr1_tme_B2 online     RW          2GB     1.85GB    7%
svm1_mccB vol2         aggr1_tme_B2 online     RW          1GB    972.5MB    5%
```

Viewing LUNs

LUNs are visible on the cluster only where they are active. In this configuration, only cluster A has an SVM with a LUN defined.

```
tme-mcc-A::> lun show
```

```
Vserver   Path                            State   Mapped   Type        Size
--------- ------------------------------- ------- -------- -------- --------
svm1_mccA /vol/svm1_mccA_lun1_vol/svm1_mccA_lun1
                                          online  mapped   windows_2008
                                                                      1.00GB
```

In normal operation, cluster B therefore does not display any LUNs.

```
tme-mcc-B::> lun show
This table is currently empty.
```

Viewing LIFs

Cluster LIFs, cluster management LIFs, and intercluster LIFs are visible only on the local cluster. Data LIFs are visible on both clusters within the scope of their SVM. The following output shows only these data LIFs. The LIF replicated from cluster B's SVM (`svm1_mccB_nfs_lif1`) shows an operational state of `down` because it is not currently usable on cluster A. Replicated LIFs are created by default on a node's DR partner.

Note the IP address and port assignment of each of the LIFs; these settings are preserved after switchover, as shown in the corresponding "Viewing LIFs" in section 3.2. For FC SAN configurations, each node must be logged in to the correct fabric in the front-end SAN. If this information is not correct, then LIFs cannot be created and assigned correctly on the partner cluster, and switchover is not possible.

```
tme-mcc-A::> network interface show -vserver svm*
            Logical    Status     Network            Current       Current Is
Vserver     Interface  Admin/Oper Address/Mask       Node          Port    Home
----------- ---------- ---------- ------------------ ------------- ------- ----
svm1_mccA
            svm1_mccA_iscsi_lifA1_1
                         up/up    10.228.22.68/24    tme-mcc-A1    e0a     true
            svm1_mccA_iscsi_lifA1_2
                         up/up    10.228.22.69/24    tme-mcc-A1    e0b     true
            svm1_mccA_iscsi_lifA2_1
                         up/up    10.228.22.97/24    tme-mcc-A2    e0a     true
            svm1_mccA_iscsi_lifA2_2
                         up/up    10.228.22.98/24    tme-mcc-A2    e0b     true
            svm1_mccA_nas_A1_1
                         up/up    10.228.22.62/24    tme-mcc-A1    e0a     true
svm1_mccB-mc
            svm1_mccB_nfs_lif1
                         up/down  10.228.22.74/24    tme-mcc-A1    e0a     true
```

Similarly, on cluster B, the reverse is true. LIFs replicated from cluster A are in an operational state of `down`.

```
tme-mcc-B::> network interface show -vserver svm*
            Logical    Status     Network            Current       Current Is
Vserver     Interface  Admin/Oper Address/Mask       Node          Port    Home
----------- ---------- ---------- ------------------ ------------- ------- ----
svm1_mccA-mc
            svm1_mccA_iscsi_lifA1_1
                         up/down  10.228.22.68/24    tme-mcc-B1    e0a     true
            svm1_mccA_iscsi_lifA1_2
                         up/down  10.228.22.69/24    tme-mcc-B1    e0b     true
            svm1_mccA_iscsi_lifA2_1
                         up/down  10.228.22.97/24    tme-mcc-B2    e0a     true
            svm1_mccA_iscsi_lifA2_2
                         up/down  10.228.22.98/24    tme-mcc-B2    e0b     true
            svm1_mccA_nas_A1_1
                         up/down  10.228.22.62/24    tme-mcc-B1    e0b     true
svm1_mccB
            svm1_mccB_nfs_lif1
                         up/up    10.228.22.74/24    tme-mcc-B1    e0b     true
```

3 Resiliency for Planned and Unplanned Events

NetApp MetroCluster enhances the high availability (HA) and nondisruptive operations of NetApp hardware and clustered Data ONTAP configurations, providing sitewide protection for the entire storage environment. Whether the application environment is composed of standalone servers, HA server clusters, or virtualized servers, MetroCluster seamlessly maintains storage availability in the event of failure at one site. Storage is available whether that failure is caused by a loss of power, cooling, or network connectivity; destruction of hardware; or operational error.

A MetroCluster configuration provides three basic methods for continued data availability in response to planned or unplanned events:

- Redundant components for protection against single component failure
- Local HA takeover for events that affect a single controller
- Complete site switchover: rapid resumption of service by moving storage and client access from the failed cluster to the surviving cluster

As seen earlier, key components of the MetroCluster infrastructure are redundant: two FibreBridges per stack, two switch fabrics, two FC-VI connections per node, two FC initiators per node per fabric, and multiple ISL links per fabric. This setup means that operations continue seamlessly in the event of single component failure, and the systems return to redundant operation when the failed component is repaired or replaced.

HA takeover and giveback functionality is inherent in all NetApp clustered Data ONTAP clusters, apart from single-node clusters in a two-node configuration that uses switchover and switchback for redundant operations. In a four-node configuration, controllers are organized into HA pairs, in which each of the two nodes is locally attached to the storage.

Takeover is the process in which one node automatically takes over the other’s storage so that its data services are preserved. Giveback is the reverse process to resume normal operation. Takeover can be planned (for example, when performing hardware maintenance or a clustered Data ONTAP upgrade on a node) or unplanned, such as for a node hardware or software failure. During a takeover, NAS LIFs are also automatically failed over. SAN LIFs do not fail over; hosts automatically use the direct path to the LUNs. Because HA takeover and giveback functionality is not specific to MetroCluster, for more information, see the Data ONTAP High Availability Guide.

Site switchover occurs when one cluster is offline. The remaining site assumes ownership of the offline cluster’s storage resources (disks and aggregates). The offline cluster’s SVMs are brought online and are restarted on the surviving site, preserving their full identity for client and host access.

3.1 MetroCluster with Unplanned and Planned Operations

Table 7 lists potential unplanned events and the behavior of MetroCluster configurations in these scenarios.

Table 7) Unplanned operations and MetroCluster response and recovery methods.

Unplanned Operation	Recovery Method
One or two disks fail	Automatic RAID recovery. No failover or switchover; both plexes remain available in all aggregates. Rebuilt disks from spares are automatically assimilated into the aggregate. NetApp RAID DP aggregates can survive two disk failures. RAID 4 aggregates are also supported but survive only a single disk failure. Both RAID DP and RAID 4 aggregates are supported with MetroCluster.

Unplanned Operation	Recovery Method
More than two disks fail, including shelf failure	Data is served from the surviving plex; there is no interruption to data services. The disk failure could affect either a local or a remote plex. The aggregate is placed in degraded mode because only one plex is active. If the failure is due to power loss on the shelf, when power is restored, the affected aggregate automatically resyncs itself to catch up any changes. If disks need to be replaced, the administrator deletes the failed plex (`storage aggregate plex delete` command) and then remirrors the affected aggregate (`storage aggregate mirror` command). This action begins the automatic resync process. After resync, the aggregate returns automatically to normal mirrored mode.
Switch fails	Data continues to be served on the surviving path; all plexes remain available. Because there are two fabrics, one fabric could completely fail and operation would be preserved.
FibreBridge fails	All traffic continues to the affected stack by the surviving bridge; all plexes remain available.
Single node fails	**Four-node configuration:** 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. All plexes remain available. The second site nodes (B1 and B2) are unaffected. If the failover is part of a rolling disaster, forced switchover can be performed to site B. An example is if node A1 fails over to A2 and there is a subsequent failure of A2 or of the complete site A. **Two-node configuration:** In a two-node configuration, there is no local HA pair, so a failure of a node requires a switchover to the remote MetroCluster partner node. Switchover is automatic if the mailbox disks are accessible.

Unplanned Operation	Recovery Method
ISL loss between sites	If one or more ISLs fail, I/O 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 continues to operate and serve its local data. The likelihood of all ISLs failing should be extremely low, given that two independent fabric connections are required (which can use independent network providers), and up to four separate ISLs per fabric can be configured. After a minimum of one ISL is restored, the plexes resynchronize automatically. Any writes that occur while all ISLs are down are not mirrored between the sites. A forced switchover while in this state could mean loss of the data that was not synchronized on the surviving site. In this case, manual intervention is required for recovery after the switchover; see section 3.5 for more details. If it is likely that no ISLs will be available for an extended period, an administrator can shut down all data services to avoid the risk of data loss if a forced switchover is necessary. Performing this action should be weighed against the likelihood that a disaster requiring switchover will occur before at least one ISL is available. Alternatively, if ISLs are failing in a cascading scenario, an administrator can trigger a planned switchover to one of the sites before all the links fail.
Peered cluster link (CRS) fails	Because the 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 or provisioning a volume or LUN in an existing SVM) cannot be propagated to the other site. These changes are kept in the local CRS metadata volumes and are automatically propagated to the other cluster upon restoration of the peered cluster link. Sometimes a forced switchover might be necessary before the peered cluster link can be restored. In that case, outstanding cluster configuration changes are replayed automatically from the remote replicated copy of the metadata volumes at the surviving site as part of the switchover process.
All nodes at a site fail, or the complete site is destroyed	The administrator performs a forced switchover to resume services of the failed nodes on the surviving site. Forced switchover is a manual operation; see section 4.4 for more information. After the failed nodes or sites are restored, a switchback operation is performed to restore steady-state operation of the configuration. In a four-node configuration, if the configuration was switched over, then a subsequent failure of one of the surviving nodes can be seamlessly handled by failover to the surviving node. The work of four nodes is then performed by only one node. Recovery in this case consists of performing a giveback to the local node, then performing a site switchback.

Table 8 lists standard maintenance (planned) events and how they are performed in a MetroCluster configuration.

Table 8) Planned operations with MetroCluster.

Planned Operation	Nondisruptive Process
Upgrading clustered Data ONTAP	For four-node configurations, this operation is performed by using nondisruptive upgrade (NDU: failover and giveback within the HA pair) as for any clustered Data ONTAP upgrade. For a minor upgrade (for example, 8.3 to 8.3.x), you can upgrade each of the two sites' clusters independently. You can use automated NDU for this upgrade. The upgrades of each cluster do not have to be synchronized, and operation and resilience continue even while the two clusters have slightly different clustered Data ONTAP versions. However, NetApp recommends that you finish the upgrade process as soon as possible for steady-state operation. A planned switchover should be delayed until all nodes are upgraded to the same clustered Data ONTAP version. Major updates (for example, 8.3 to 8.x) and two-node upgrades require orchestration; that is, both clusters should be upgraded together. Switchover and switchback operations are not possible when the clusters are running different clustered Data ONTAP versions.
Upgrading controller hardware by using aggregate relocation (ARL)	You can use ARL to nondisruptively upgrade controller models in place: for example, to upgrade FAS8040 to FAS8060. All nodes in the cluster must be upgraded to the same controller model. Storage replication is not affected while the controllers are being upgraded; however, NVRAM replication between the sites must be disabled during the entire process. The two clusters must be upgraded in sequence: 1. Disable NVRAM replication on all nodes. 2. Upgrade the nodes in the first cluster. 3. Upgrade the nodes in the second cluster. 4. Enable NVRAM replication on all nodes. The upgraded controller nodes have different serial numbers and system IDs. MetroCluster automatically retrieves the new system IDs and correctly identifies the new HA and DR partners. Disk mirroring continues during the ARL process, but NVRAM must be disabled. Therefore, the aggregates stay in sync; if a forced switchover is necessary during this time, at most the last 10 seconds of transactions might be missing. Planned switchover is not possible while a controller upgrade is in progress. The command to disable/enable NVRAM mirroring is: `metrocluster interconnect modify -node <nodename> -partner-type DR -mirror-status OFFLINE|ONLINE` The ARL process is lengthy and has many steps. For the detailed steps for this process, see the controller upgrade guide. Follow the steps closely.

3.2 Performing Planned (Negotiated) Switchover

There are two types of switchover: planned switchover and forced switchover after a disaster. A planned or negotiated switchover can be executed when the requirement to switch over is known in advance and all nodes are operational. The planned switchover gracefully transfers resource ownership, shutting down all services on the site that is switched over and then resuming them on the surviving site. The nodes being switched over are also shut down cleanly.

Before a planned switchover is executed, the MetroCluster configuration must be operating in a steady state. This requirement includes the following items:

- All nodes are operational with no failovers or givebacks in progress.
- A clustered Data ONTAP software update is not in progress, and all nodes are running the same release.
- The cluster peering network is up and operational.
- At least one ISL is up.
- There are no long-running tasks in progress that need to be restarted.
- CPU utilization is such that switchover can complete in a reasonable time frame.

Note: This list is not complete. Prechecks are automatically made for any other conditions that prevent a planned switchover. This approach is by design, because a planned switchover is a “clean” operation and relies on the system’s being in a consistent state.

If one of these conditions fails the prechecks and it is necessary nevertheless to perform a switchover, you can perform a forced switchover that is not subject to these requirements and vetoes. However, NetApp recommends that you wait until the vetoing condition is cleared and then proceed with the planned switchover.

A planned switchover is useful for testing purposes or, for example, if site or other planned maintenance is performed. To execute a planned switchover, issue the `metrocluster switchover` command on the cluster that assumes the resources. Following is a summary of the procedure, assuming that cluster A is switched over to cluster B. Cluster B is the surviving site:

1. Send a NetApp AutoSupport® message to alert NetApp Support that planned maintenance or testing is taking place, as advised in section 4.3.
2. Verify that the environment is ready for a switchover. Both clusters should be in a steady state before you issue a planned switchover. After you make any changes to either cluster configuration, wait at least several minutes before you issue the switchover command so that the changes can be replicated. Do not make any further configuration changes until the switchover is complete.
3. Execute `metrocluster check run` to verify that all components are OK. If any errors are reported, correct them before you continue.
4. On cluster B’s clustershell, execute `metrocluster switchover –simulate` (in advanced mode) to verify that the cluster can be switched over. This command runs all the prechecks for any conditions that preclude a planned switchover, but it does not take any of the actual switchover steps. You can perform prechecks for a planned switchover because a planned switchover is typically not as urgent as a forced switchover is. You should see the following message:

```
[Job 1234] Job succeeded: Switchover simulation is successful.
```

   Any other message indicates that the system is not in a steady state for a planned switchover. Correct the reported condition and try again.

5. Perform the switchover on cluster B: `metrocluster switchover`. The entire process can take several minutes to complete; however, actual client or host pauses in I/O should take less than two minutes. Use the `metrocluster operation show` command to monitor the progress. It is helpful to monitor the consoles of cluster A during this process to confirm the node shutdown. Switchover requires only one administrator command, and the following tasks are automatically performed, with no further intervention necessary:
   a. It checks the configuration for conditions that can be vetoed by using the same rules as in the `metrocluster switchover –simulate` command.
   b. It flushes cluster A’s NVRAM to disk for I/O consistency.

c. All of cluster A’s volumes and aggregates are taken offline. At this point, client and host I/O is paused. SVMs and LIFs remain on, meaning that paths to cluster A’s LUNs are kept available as long as possible to allow SAN hosts to respond to path inquiries.

d. Cluster B takes ownership of cluster A’s owned disks (both pools). The nonroot (data) aggregates are assimilated into NetApp WAFL, and their volumes come online.

e. Cluster A offlines its SVMs and LIFs.

f. SVMs from cluster A are brought online at cluster B. LIFs come online, and protocol services resume. Paused I/O from clients, hosts, and applications automatically resumes.

g. Cluster A’s nodes are shut down and wait at the `LOADER>` prompt. Storage remains available by default. If it is necessary to fence off the shutdown site entirely, the storage, bridge, and switches can be powered off, leaving only one plex available. NetApp recommends fencing the storage only if necessary: for example, if power is cut to the data center or ISLs are down. Leaving the storage up means higher resiliency because both plexes remain available. It also reduces the time to switch back because little or no resync is necessary.

The command output should look like the following. Use the command `metrocluster operation show` after job completion to confirm that the operation was successful.

```
tme-mcc-B::> metrocluster switchover

Warning: negotiated switchover is about to start. It will stop all the data Vservers on cluster
"tme-mcc-A" and automatically re-start them on cluster "tme-mcc-B". It will
         finally gracefully shutdown cluster "tme-mcc-A".
Do you want to continue? {y|n}: y
[Job 2839] Job succeeded: Switchover is successful.
```

6. When the message `Switchover is successful` is displayed, verify that site B is stable, as described in the sections “Confirming That the DR Partners Have Come Online” and “Reestablishing SnapMirror or SnapVault SVM Peering Relationships” in the MetroCluster Management and Disaster Recovery Guide. As noted in section 4.7, all SnapMirror or SnapVault relationships with a destination volume in the switched-over cluster (cluster A, in our example) must be manually re-created after any switchover or switchback operation. SnapMirror and SnapVault relationships with a source on a MetroCluster configuration continue without intervention.

```
tme-mcc-B::> metrocluster node show
DR                               Configuration  DR
Group Cluster Node               State          Mirroring Mode
----- ------- ------------------ -------------- --------- --------------------
1     tme-mcc-B
              tme-mcc-B1         configured     enabled   switchover completed
              tme-mcc-B2         configured     enabled   switchover completed
      tme-mcc-A
              tme-mcc-A1         unreachable    -         switched over
              tme-mcc-A2         unreachable    -         switched over
```

7. At this point, testing and verification or other planned maintenance can proceed, depending on the purpose of the planned switchover. When testing or other maintenance is complete, a switchback can be initiated, as described in section 3.3.

Following are the equivalent commands and checks from section 2.3, but executed after switchover. They show how the object view changes and verify that all the resources were switched over. Because cluster A is shut down, the commands are run only on the surviving cluster B.

Viewing Aggregates

After a planned switchover, all aggregates are visible on site B. Site A’s aggregates are displayed as switched-over aggregates. Only the data aggregates are online. Root aggregates are switched over, but are not brought online. All the online aggregates display as mirrored and with normal RAID status, because in this planned switchover, no storage was powered off. Both plexes are therefore available.

```
tme-mcc-B::> aggr show
```

```
tme-mcc-A Switched Over Aggregates:
Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
aggr0_tme_A1    0B        0B    0% offline      0 tme-mcc-B2       raid_dp,
                                                                   mirror
                                                                   degraded
aggr0_tme_A2    0B        0B    0% offline      0 tme-mcc-B1       raid_dp,
                                                                   mirror
                                                                   degraded
aggr1_tme_A1
            2.07TB    2.06TB    1% online       4 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_A2
            2.07TB    2.06TB    0% online       1 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal

tme-mcc-B Aggregates:
Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
aggr0_tme_B1
            1.38TB   741.9GB   48% online       1 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr0_tme_B2
            1.38TB   741.9GB   48% online       1 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_B1
            2.07TB    2.06TB    1% online       2 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_B2
            2.07TB    2.06TB    1% online       3 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
```

Viewing SVMs

Both SVMs now run on cluster B. The `sync-destination` SVM from cluster A, `svm1_mccA-mc`, has changed to a running state.

```
tme-mcc-B::> vserver show -type data
                               Admin      Operational Root
Vserver     Type    Subtype    State      State       Volume           Aggregate
----------- ------- ---------- ---------- ----------- --------------   -------------
svm1_mccA-mc
           data    sync-destination      running     svm1_mccA_root    aggr1_tme_A1
                               running
svm1_mccB   data    sync-source           running     svm1_mccB_root    aggr1_tme_B1
                               running
```

Viewing Volumes

Looking at a four-node configuration, we can see that all four MDVs are online (compared with the output shown in the section "Configuration Replication Service"). And because the data SVM from cluster A was brought online, the volumes from that SVM are also now online on cluster B. Note that, as in the previous output, the local root volumes are omitted. Each cluster sees only its own root volumes, regardless of the switchover state.

```
tme-mcc-B::> vol show
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
svm1_mccA-mc
```

```
          svm1_mccA_lun1_vol
                       aggr1_tme_A1 online     RW       1.24GB    248.2MB   80%
svm1_mccA-mc
          svm1_mccA_root
                       aggr1_tme_A1 online     RW          1GB    972.5MB    5%
svm1_mccA-mc
          vol1         aggr1_tme_A1 online     RW          2GB     1.85GB    7%
svm1_mccB svm1_mccB_root
                       aggr1_tme_B1 online     RW          1GB    972.5MB    5%
svm1_mccB vol1         aggr1_tme_B2 online     RW          2GB     1.85GB    7%
svm1_mccB vol2         aggr1_tme_B2 online     RW          1GB    972.5MB    5%
tme-mcc-B MDV_CRS_cd7628c7f1cc11e3840800a0985522b8_A
                       aggr1_tme_A1 online     RW         10GB     9.50GB    5%
tme-mcc-B MDV_CRS_cd7628c7f1cc11e3840800a0985522b8_B
                       aggr1_tme_A2 online     RW         10GB     9.50GB    5%
tme-mcc-B MDV_CRS_e8fef00df27311e387ad00a0985466e6_A
                       aggr1_tme_B1 online     RW         10GB     9.50GB    5%
tme-mcc-B MDV_CRS_e8fef00df27311e387ad00a0985466e6_B
                       aggr1_tme_B2 online     RW         10GB     9.50GB    5%
```

Viewing LUNs

Because cluster A had a LUN provisioned in its SVM, it is now visible and accessible on cluster B.

```
tme-mcc-B::> lun show
Vserver   Path                            State   Mapped   Type        Size
--------- ------------------------------- ------- -------- -------- --------
svm1_mccA-mc
          /vol/svm1_mccA_lun1_vol/svm1_mccA_lun1
                                          online  mapped   windows_2008
                                                                      1.00GB
```

Viewing LIFs

Finally, the switched-over LIFs from cluster A are available with the operational state of `up` on cluster B. All client and host access to these LIF addresses is through cluster B. The IP addresses and port assignments are the same, reproduced exactly on cluster B. After switchover, the MetroCluster system preserves the identity of the storage access resources, including LIF IP address, LUN target ID, SCSI reservations, WWPN, and WWN. Therefore, clients and hosts do not need to change any of their access details. As for port assignment, the best practice is to have identical network ports available on the equivalent DR partners. This approach allows port assignment to also be mapped identically after switchover.

```
tme-mcc-B::> network interface show -vserver svm*
            Logical    Status     Network            Current       Current Is
Vserver     Interface  Admin/Oper Address/Mask       Node          Port    Home
----------- ---------- ---------- ------------------ ------------- ------- ----
svm1_mccA-mc
            svm1_mccA_iscsi_lifA1_1
                         up/up    10.228.22.68/24    tme-mcc-B1    e0a     true
            svm1_mccA_iscsi_lifA1_2
                         up/up    10.228.22.69/24    tme-mcc-B1    e0b     true
            svm1_mccA_iscsi_lifA2_1
                         up/up    10.228.22.97/24    tme-mcc-B2    e0a     true
            svm1_mccA_iscsi_lifA2_2
                         up/up    10.228.22.98/24    tme-mcc-B2    e0b     true
            svm1_mccA_nas_A1_1
                         up/up    10.228.22.62/24    tme-mcc-B1    e0b     true
svm1_mccB
            svm1_mccB_nfs_lif1
                         up/up    10.228.22.74/24    tme-mcc-B1    e0b     true
```

Operations While in Switchover Mode

Changes can be made to existing SVMs while in switchover mode. For example, new volumes, LUNs, or LIFs can be created in the switched-over cluster's SVMs. New SVMs for the switched-over cluster cannot be created, however. For example, while cluster A has switched over to cluster B, the administrator cannot create a new SVM to be owned by cluster A, but the administrator can add a volume to one of cluster A's SVMs. Operations for cluster B's own resources are unaffected.

If SnapMirror or SnapVault relationships have been defined within the MetroCluster configuration, see section 4.7 for important considerations.

NetApp SnapManager® and SnapProtect® software (version 10 SP9 release) are also verified for MetroCluster operation; see their documentation for more information.

New aggregates that are owned by the switched-over cluster cannot be created. Therefore, in our example, a new aggregate for cluster A could not be created. New aggregates can be created for the surviving cluster; however, if the remote storage on the other cluster is not available, these aggregates are not mirrored. It is necessary to mirror these aggregates (`storage aggregate mirror` command) after they are healed and before performing the switchback.

It is possible to extend an aggregate from either cluster while in switchover mode. If only one plex is available (that is, the storage from the switched-over cluster has been fenced off), unmirror the aggregate (`storage aggregate plex delete` command). Then extend it by adding disks from the corresponding plex on the surviving cluster. After healing the aggregates and before switchback, use the `storage aggregate mirror` command to remirror the extended aggregates.

3.3 Performing Switchback

Switchback is always a planned operation; that is, it must be initiated by the administrator. A sequence of three commands is used to coordinate bringing up the shutdown site and handing over its resources. Presented here is a summary of the steps; follow the complete process as documented in the "Healing the Configuration" and "Performing a Switchback" sections of the MetroCluster Management and Disaster Recovery Guide:

1. If storage was powered off at site A as part of the switchover testing, power on the shelves now, including bridges and switches if necessary. **Do not** power on the nodes.
2. The cluster must be in a steady state. The surviving cluster must have both nodes available and must not be in takeover mode. The cluster peering network and at least one ISL must be up.
3. Heal the data aggregates with the following command:

```
tme-mcc-B::> metrocluster heal -phase aggregates
[Job 2853] Job succeeded: Heal Aggregates is successful.
```

4. If storage at the shutdown site was offline during the switchover, resync is automatically initiated to propagate any I/O that occurred during switchover. Wait for the resync to be complete; the next phase of switchback cannot execute until the data aggregates are resynced. To monitor the status, use the following command:

```
tme-mcc-B::> aggr show-resync-status
                                    Complete
Aggregate Resyncing Plex            Percentage
--------- ------------------------- ----------
aggr0_tme_A1
          plex0                              -
aggr0_tme_A1
          plex2                              -
aggr0_tme_A2
          plex0                              -
aggr0_tme_A2
          plex2                              -
aggr0_tme_B1
          plex0                              -
```

```
aggr0_tme_B1
          plex2                              -
aggr0_tme_B2
          plex0                              -
aggr0_tme_B2
          plex2                              -
aggr1_tme_A1
          plex0                             70%
aggr1_tme_A1
          plex1                              -
aggr1_tme_A2
          plex0                              -
aggr1_tme_A2
          plex1                              -
aggr1_tme_B1
          plex0                              -
aggr1_tme_B1
          plex1                              -
aggr1_tme_B2
          plex0                              -
aggr1_tme_B2
          plex1                              -
```

5. After resync is complete, verify the aggregate status once more. All aggregates should show RAID status as `mirrored, normal`.

```
tme-mcc-B::> storage aggregate show

tme-mcc-A Switched Over Aggregates:
Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
aggr0_tme_A1    0B        0B    0% offline      0 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr0_tme_A2    0B        0B    0% offline      0 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_A1
            2.07TB    2.06TB    1% online       4 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_A2
            2.07TB    2.06TB    0% online       1 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal

tme-mcc-B Aggregates:
Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
aggr0_tme_B1
            1.38TB   741.9GB   48% online       1 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr0_tme_B2
            1.38TB   741.9GB   48% online       1 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_B1
            2.07TB    2.06TB    1% online       2 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_B2
            2.07TB    2.06TB    1% online       3 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
```

6. Heal the root aggregates. This step gives back the root aggregates to the rejoining cluster.

```
tme-mcc-B::> metrocluster heal -phase root-aggregates
[Job 2854] Job succeeded: Heal Root Aggregates is successful.
```

7. Cluster A’s root aggregates are now no longer visible on the switched-over cluster, cluster B, because they have been returned to cluster A. Note that cluster A’s data aggregates are still visible on cluster A. At this point, all the SVMs and data services still run on cluster B.

```
tme-mcc-B::> aggr show
tme-mcc-A Switched Over Aggregates:
Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
aggr0_tme_A1     -         -     - unknown      - tme-mcc-B2       -
aggr0_tme_A2     -         -     - unknown      - tme-mcc-B1       -
aggr1_tme_A1
            2.07TB    2.06TB    1% online       4 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_A2
            2.07TB    2.06TB    0% online       1 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal

tme-mcc-B Aggregates:
Aggregate     Size Available Used% State   #Vols  Nodes            RAID Status
--------- -------- --------- ----- ------- ------ ---------------- ------------
aggr0_tme_B1
            1.38TB   741.9GB   48% online       1 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr0_tme_B2
            1.38TB   741.9GB   48% online       1 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_B1
            2.07TB    2.06TB    1% online       2 tme-mcc-B1       raid_dp,
                                                                   mirrored,
                                                                   normal
aggr1_tme_B2
            2.07TB    2.06TB    1% online       3 tme-mcc-B2       raid_dp,
                                                                   mirrored,
                                                                   normal
```

8. Verify that the configuration is ready for the next step.

```
tme-mcc-B::> metrocluster node show
DR                               Configuration  DR
Group Cluster Node               State          Mirroring Mode
----- ------- ------------------ -------------- --------- --------------------
1     tme-mcc-B
              tme-mcc-B1         configured     enabled   heal roots completed
              tme-mcc-B2         configured     enabled   heal roots completed
      tme-mcc-A
              tme-mcc-A1         unreachable    -         switched over
              tme-mcc-A2         unreachable    -         switched over
```

9. Check whether there are failed disks on the cluster A nodes, using the `disk show –broken` command. Remove any failed disks from the shelves before you continue.

10. Power on or boot the cluster A nodes and wait for the cluster and MetroCluster configuration to return to a stable state. The following output on cluster A shows that it has not completely synchronized after booting. Node A2 is not yet healthy, and the peering to cluster B has not been reestablished. An attempt to switch back fails at this point.

```
tme-mcc-A::> cluster show
Node                  Health  Eligibility
--------------------- ------- ------------
tme-mcc-A1            true    true
tme-mcc-A2            false   true
2 entries were displayed.

tme-mcc-A::> cluster peer show
Peer Cluster Name         Cluster Serial Number Availability   Authentication
------------------------- --------------------- -------------- --------------
```

```
tme-mcc-B                 1-80-000011           Unavailable    absent
```

11. Verify that the nodes are in the correct state.

```
tme-mcc-B::> metrocluster node show
DR                                Configuration  DR
Group Cluster Node                State          Mirroring Mode
----- ------- ------------------ -------------- --------- --------------------
1     tme-mcc-B
              tme-mcc-B1          configured     enabled   heal roots completed
              tme-mcc-B2          configured     enabled   heal roots completed
      tme-mcc-A
              tme-mcc-A1          configured     enabled   waiting for switchback recovery
              tme-mcc-A2          configured     enabled   waiting for switchback recovery
```

12. Wait for cluster A to be stable and for the cluster peering to be fully operational. This step might take several minutes. The Availability field for cluster peering should change from unavailable to pending to available:

```
tme-mcc-A::> cluster show
Node                  Health  Eligibility
--------------------- ------- ------------
tme-mcc-A1            true    true
tme-mcc-A2            true    true
tme-mcc-A::> cluster peer show
Peer Cluster Name         Cluster Serial Number Availability   Authentication
------------------------- --------------------- -------------- --------------
tme-mcc-B                 1-80-000011           Pending        absent

tme-mcc-A::> cluster peer show
Peer Cluster Name         Cluster Serial Number Availability   Authentication
------------------------- --------------------- -------------- --------------
tme-mcc-B                 1-80-000011           Available      absent
```

13. Verify that cluster peering is healthy from cluster B.

```
tme-mcc-B::*> cluster peer show
Peer Cluster Name         Cluster Serial Number Availability   Authentication
------------------------- --------------------- -------------- --------------
tme-mcc-A                 1-80-000011           Available      absent
```

14. Run `switchback –simulate` in advanced mode on cluster B as a final check. This command runs all the prechecks for any conditions that preclude a switchback, but it does not perform any of the actual switchback steps. If the command reports that the switchback is vetoed, clear this condition before proceeding. For example, if one node in cluster B was taken over, giveback must be performed before proceeding.

```
tme-mcc-B::*> metrocluster switchback -simulate
[Job 2855] Job succeeded: Switchback simulation is successful.

tme-mcc-B::*> metrocluster operation show
  Operation: switchback-simulate
      State: successful
 Start Time: 2/27/2015 16:31:51
   End Time: 2/27/2015 16:32:20
     Errors: -
```

15. Finally, perform the switchback. The switchback command automatically performs the following steps with no further intervention necessary:
    a. It checks the configuration for conditions that cause switchback to be vetoed by using the same rules as in the `metrocluster switchback –simulate` command.
    b. It sends a complete copy of all the required cluster configurations (reverse baseline) over the cluster peering network. This step enables any configuration change that occurred during switchover to be replicated back to the rejoining cluster.
    c. It flushes NVRAM to disk for I/O consistency.

   d. Cluster B takes all of cluster A’s volumes and aggregates offline. At this point, client and host I/O is paused. SVMs and LIFs remain on, meaning that paths to cluster A’s LUNs are kept available as long as possible to allow SAN hosts to respond to path inquiries.
   e. Disk ownership of the data aggregates is restored to its preswitched-over state. Cluster A assimilates its returned aggregates into WAFL and brings its volumes online.
   f. Cluster B offlines cluster A’s SVMs and LIFs.
   g. Cluster A brings its SVMs online, brings its LIFs online, and resumes protocol services. Paused I/O from clients, hosts, and applications automatically resumes on cluster A.

```
tme-mcc-B::> metrocluster switchback
[Job 2856] Job succeeded: Switchback is successful.

tme-mcc-B::> metrocluster operation show
  Operation: switchback
      State: successful
 Start Time: 2/27/2015 17:21:36
   End Time: 2/27/2015 17:22:38
     Errors: -

tme-mcc-B::> metrocluster node show
DR                               Configuration  DR
Group Cluster Node               State          Mirroring Mode
----- -------- ------------------ -------------- --------- --------------------
1     tme-mcc-B
              tme-mcc-B1         configured     enabled   normal
              tme-mcc-B2         configured     enabled   normal
      tme-mcc-A
              tme-mcc-A1         configured     enabled   normal
              tme-mcc-A2         configured     enabled   normal
```

16. Switchback is complete.

The MetroCluster Management and Disaster Recovery Guide provides detailed instructions to perform switchover for tests or maintenance.

3.4 Performing Forced Switchover

A forced switchover differs from a negotiated switchover in that the nodes and storage from the failed site either are not available or must first be completely fenced off (isolated). The surviving cluster performs all the steps, without access to any resources from site A. Here is a summary of the steps for an unplanned or forced switchover. Follow the detailed instructions in the MetroCluster Management and Disaster Recovery Guide:

1. Fence off the disaster-affected site (see the section “Fencing Off the Disaster Site” in the MetroCluster Management and Disaster Recovery Guide).
2. Unlike in a planned switchover, you must perform the forced switchover on the surviving site: `metrocluster switchover –force-on-disaster true`. The following steps are performed automatically to resume services on the surviving site:
   a. The surviving cluster nodes take ownership of the failed site’s `Pool1` disks.
   b. The surviving cluster assimilates the failed site’s aggregates and brings its volume online.
   c. Any uncommitted I/O is replayed.
   d. The surviving cluster brings the failed cluster’s SVMs and LIFs online and starts the protocol services, resuming storage services. Client, host, and application I/O from the failed site resumes.
3. Manually reestablish the SVM peering (if the destination of the relationship is in the MetroCluster configuration).

3.5 Protecting Volumes After Forced Switchover

In rare circumstances, volumes might be inconsistent after switchover. A rolling failure could cause this situation. For example, suppose that the two sites are first isolated from each other by failure of all the intersite links. This event is subsequently followed by a complete hardware failure at site A. In the intervening period between the failure of the links and the failure of the hardware, it is possible that some writes could occur at site A that are not propagated to site B. If switchover to site B is subsequently triggered, the data at site B is inconsistent with the data at site A.

A MetroCluster configuration detects this scenario by checking for NVRAM inconsistencies in the nodes. Any affected volumes can be fenced off so that application-level recovery can be performed as necessary. In normal circumstances in which no NVRAM inconsistency is detected, volumes are not fenced and are available automatically after switchover. In unfenced volumes, LUNs are brought online automatically, and file system IDs (FSIDs) for NFS do not change so that client mounts remain current. If volumes are fenced, then LUNs remain offline, and FSIDs change, returning stale file handles to NFS clients. This approach allows the administrator to decide when to make the data available to clients and applications after taking appropriate action to recover the lost data. SMB clients are not affected.

If data inconsistency is detected after switchover, any SVM root volumes and volumes that contain LUNs automatically set a flag known as *NVFAIL*. To enable NVFAIL to be set if required on new or existing volumes, specify the `–nvfail` flag on either the `volume create` or `volume modify` command, respectively. When the NVFAIL flag is set after switchover, the volumes are fenced and are not accessible to clients until they are manually unfenced. To unfence a volume, use the `-in-nvfailed-state` parameter of the `volume modify` command in advanced mode. If the root volume is in NVFAIL state, the root recovery procedure must be used.

NetApp recommends making sure that NVFAIL is set for any volumes that are part of a database application. For more information about this topic, see the MetroCluster Management and Disaster Recovery Guide.

3.6 Recovery from a Forced Switchover, Including Complete Site Disaster

The recovery process depends on the nature of the event that triggered the forced switchover. If it was a sudden, lengthy power outage at one site, then, after power is restored, the switchback can be performed because no hardware replacement is required. However, if hardware was destroyed (up to and including destruction of the data center itself), it must be replaced before you perform a switchback. In this instance, the site might be operating in switched-over mode for an extended period of time.

While the system is operating in switched-over mode, a surviving cluster HA pair could experience a local takeover event, leaving only one node operational. Clearly at this point the remaining node represents a single final point of failure. A further disaster affecting the surviving site correspondingly affects the storage availability. In that case, there is no disaster protection until the original disaster site is restored.

While the system is in switched-over mode, take extra care to monitor and recover from any subsequent failures as quickly as possible. Do not perform a major clustered Data ONTAP upgrade on the surviving cluster, because it adds significant operational complexity. You can perform a minor upgrade while the system is in switched-over mode. However, NetApp recommends that you seek guidance from Support beforehand to verify specific recommendations and to check for any considerations that are necessary when you perform the eventual switchback. For example, you must upgrade the down nodes at the disaster site to the same clustered Data ONTAP release before you perform the switchback.

Recovery steps after a forced switchover or disaster are documented in the MetroCluster Management and Disaster Recovery Guide. If hardware has to be replaced, see the section “Recovering from a Disaster When Both Controllers Failed.” If hardware replacement is not necessary, see the section “Recovering from a Site Failure When No Controllers Were Replaced.”

4 Interoperability

Several management tools are available to monitor the NetApp MetroCluster configuration, including NetApp OnCommand System Manager, OnCommand Unified Manager, MetroCluster Tiebreaker software, and Config Advisor.

To highlight their compatibility and interactions with MetroCluster configurations, a number of standard NetApp clustered Data ONTAP features, including QoS, SnapMirror and SnapVault, SnapLock, volume move, volume rehost, and Flash Pool, are also discussed here.

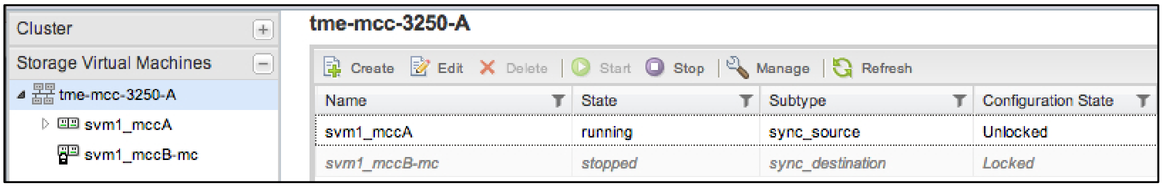
In general, after initial setup and testing, the MetroCluster configuration is managed as two independent clusters. You should create and configure SVMs (including the associated volumes, LIFs, LUNs, access policies, and so on) on either cluster as required. You can use the management interface that you prefer, for example, OnCommand System Manager, the CLI, OnCommand Workflow Automation (WFA), or another API-based tool.

4.1 OnCommand System Manager

OnCommand System Manager is a web-based on-box application in the clustered Data ONTAP 8.3.x operating system and is accessed by specifying the cluster management LIF as the URL. After you set up the MetroCluster configuration, you can use System Manager to create and configure the SVMs and their associated objects. Of course, the CLI, WFA, and other API-based methods are also available and supported.

All SVMs are visible on both clusters; however, when the clusters are in a steady state (both clusters are operational), only SVMs of the subtype state `sync_source` can be administered or updated on a cluster. In Figure 13, SVM `svm1_mccA` is the `sync_source` SVM on cluster A and shows the state as running and the configuration state as unlocked. SVM `svm1_mccB-mc` is the `sync_destination` copy of SVM `svm1_mccB`, and it shows the state as stopped and the configuration state as locked on cluster B.

Figure 13) Cluster A: The `sync_source` SVM is unlocked, and the `sync_destination` SVM is locked.

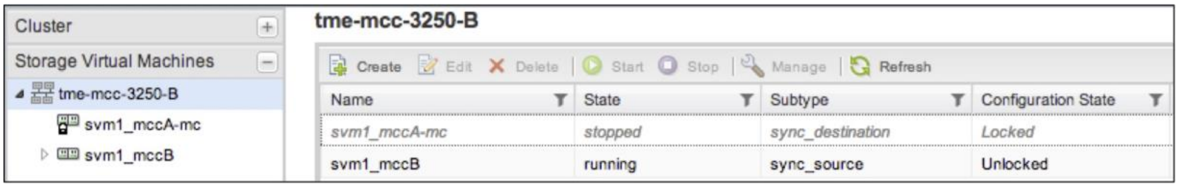


tme-mcc-3250-A

Name	State	Subtype	Configuration State
svm1_mccA	running	sync_source	Unlocked
svm1_mccB-mc	*stopped*	*sync_destination*	*Locked*

The equivalent SVM display from System Manager on cluster B shows SVM `svm1_mccA-mc` in the state of stopped and in the configuration state of locked (subtype `sync_destination`). SVM `svm1_mccB` is in the state of running and in the configuration state of unlocked (subtype `sync_source`).

Figure 14) Cluster B: The `sync_source` SVM is unlocked, and the `sync_destination` SVM is locked.

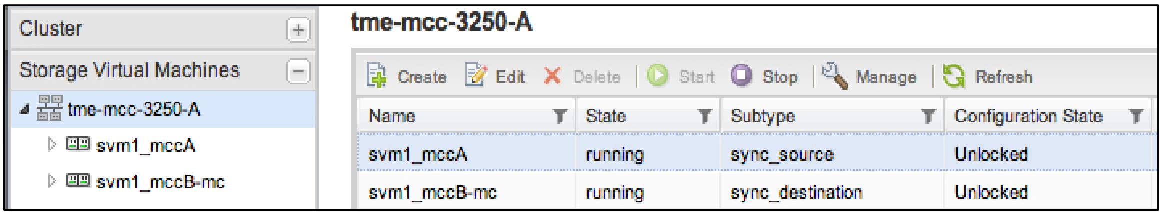


tme-mcc-3250-B

Name	State	Subtype	Configuration State
svm1_mccA-mc	*stopped*	*sync_destination*	*Locked*
svm1_mccB	running	sync_source	Unlocked

If a switchover is performed, `sync_destination` SVMs are unlocked and can be updated on the surviving cluster, for example, to provision new volumes and LUNs or to create or update export policies. Figure 15 shows the status after switching over cluster B to cluster A; the SVM `svm1_mccB-mc` is now running and unlocked on cluster A.

Figure 15) SVMs after switchover: All SVMs are unlocked.



Name	State	Subtype	Configuration State
svm1_mccA	running	sync_source	Unlocked
svm1_mccB-mc	running	sync_destination	Unlocked

4.2 OnCommand Unified Manager and Health Monitors

OnCommand Unified Manager (OCUM) 6.2 supports discovery, monitoring, and alerting of MetroCluster topology and configuration, including the nodes, links, bridges, switches, and storage. OCUM leverages the following capabilities, which are provided by the built-in system health monitors for MetroCluster:

- SNMP monitoring of the switches and FibreBridges
- MetroCluster and cluster topology details, including the status of NetApp SyncMirror storage replication and NVRAM replication
- Alerts if issues are encountered

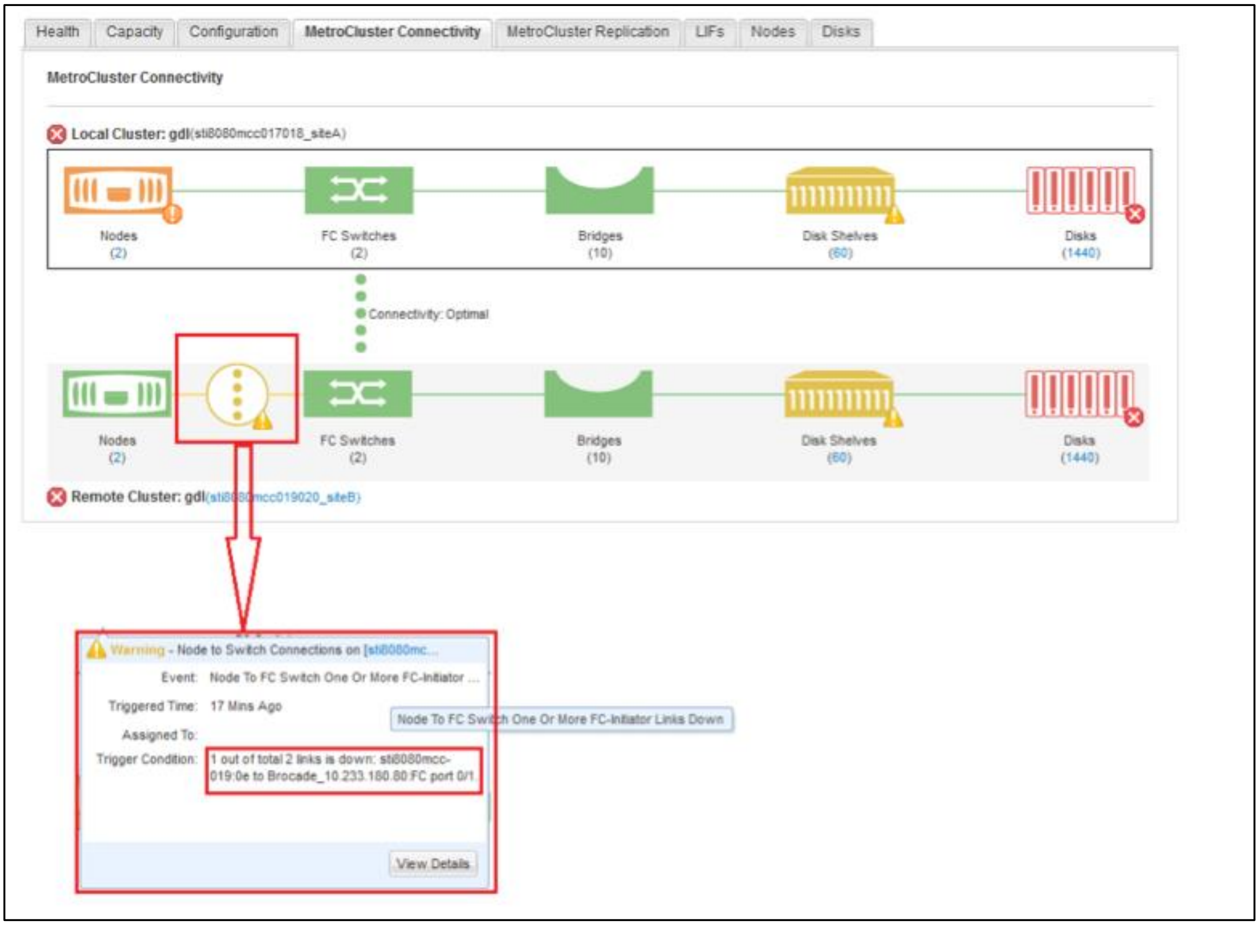
NetApp highly recommends OCUM for monitoring every MetroCluster installation. It can be used in conjunction with enterprise monitoring tools.

OnCommand Unified Manager uses the cluster topology information to build an end-to-end graphical representation of the MetroCluster configuration. This representation is built automatically when OCUM discovers the clusters that form a MetroCluster system. OCUM then periodically polls the health monitors for detected faults and translates them to OCUM alerts. An alert that is raised by OCUM includes information about the likely cause and suggested remediation actions. It can be assigned to system administrators for appropriate handling. In OCUM 6.2 with clustered Data ONTAP 8.3.x, the alerts are discovered by using a polling process. As a result, depending on timing, it can take several minutes for an alert to be displayed.

OCUM uses the information that is collected by the MetroCluster health monitors to gather information about the configuration and to collect events related to the components. The health monitors use SNMP to monitor the switches and bridges. OnCommand Unified Manager creates a logical, graphical representation of the components. It also monitors the devices end to end and monitors the state of the synchronous mirroring (SyncMirror for the aggregates and NVRAM mirroring). Issues with the devices or the links raise events that can be managed and assigned through the OCUM interface.

With connectivity monitoring, OCUM monitors and reviews the health of the hardware in the MetroCluster configuration and raises alerts for issues that are related to the physical connectivity between devices. Alerts show the likely cause and the impact of the issue and suggest remediation.

Figure 16) OCUM device and link monitoring.



Warning - Node to Switch Connections on [sti8080mc...

Event: Node To FC Switch One Or More FC-Initiator ...

Triggered Time: 17 Mins Ago

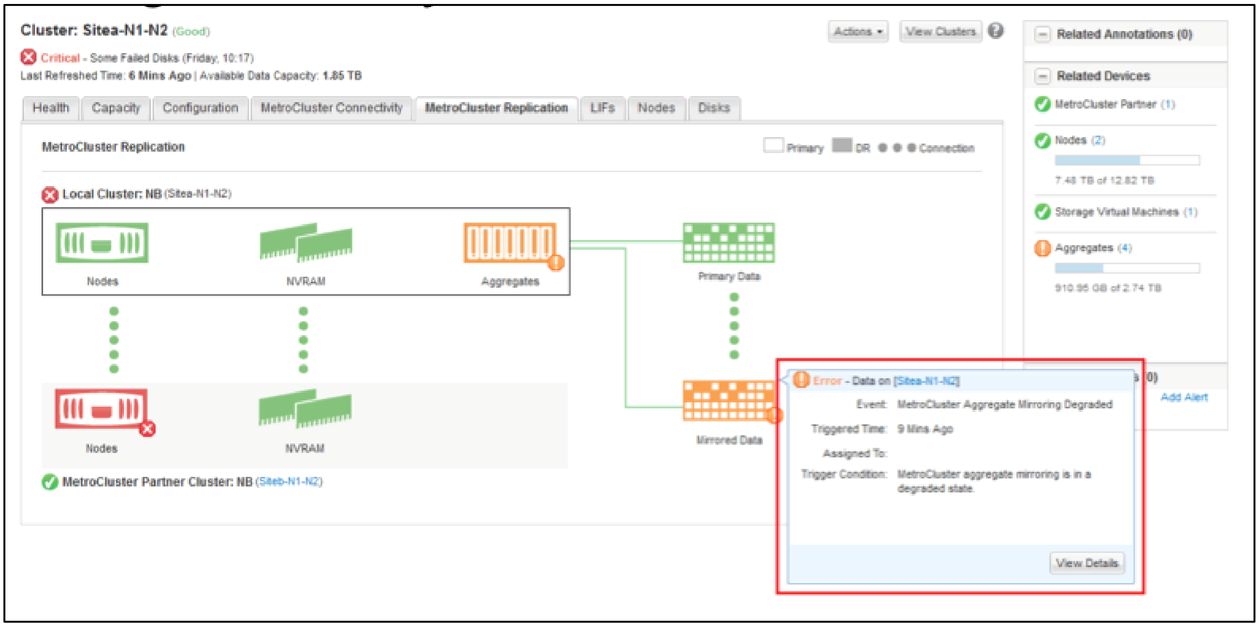
Assigned To:

Trigger Condition: 1 out of total 2 links is down: sti8080mcc-019:0e to Brocade_10.233.180.80:FC port 0/1

Node To FC Switch One Or More FC-Initiator Links Down

Replication monitoring displays the health of the synchronous relationships in MetroCluster: SyncMirror for aggregates and the NVRAM replication.

Figure 17) OCUM replication monitoring.



Cluster: Sitea-N1-N2 (Good)

Critical - Some Failed Disks (Friday, 10:17)

Last Refreshed Time: 6 Mins Ago | Available Data Capacity: 1.85 TB

Local Cluster: NB (Sitea-N1-N2)

MetroCluster Partner Cluster: NB (Siteb-N1-N2)

Error - Data on [Sitea-N1-N2]

Event: MetroCluster Aggregate Mirroring Degraded

Triggered Time: 9 Mins Ago

Assigned To:

Trigger Condition: MetroCluster aggregate mirroring is in a degraded state.

Related Annotations (0)

Related Devices

MetroCluster Partner (1)

Nodes (2)

7.48 TB of 12.82 TB

Storage Virtual Machines (1)

Aggregates (4)

910.95 GB of 2.74 TB

4.3 AutoSupport

AutoSupport messages that are specific to MetroCluster configurations are also automatically sent. NetApp strongly recommends the use of AutoSupport for all MetroCluster configurations. In a MetroCluster configuration, a case is automatically opened in response to certain events, including SyncMirror plex failure and switchover or switchback failures. This feature allows NetApp Support to respond quickly and proactively.

If you are performing MetroCluster operations solely for testing purposes or planned operations (such as verifying switchover and switchback capabilities), you should send user-triggered AutoSupport messages to alert NetApp Support that testing is taking place. This notification prevents an automatic case from being escalated and lets NetApp Support know that a real disaster event has not occurred. NetApp Knowledge Base article 1015155 (logging in to the NetApp Support site is required) has more information about using AutoSupport for this purpose.

My AutoSupport includes a dashboard and visualization of MetroCluster configurations, including system status, physical connectivity, storage usage, a health summary, and more.

4.4 MetroCluster Tiebreaker Software

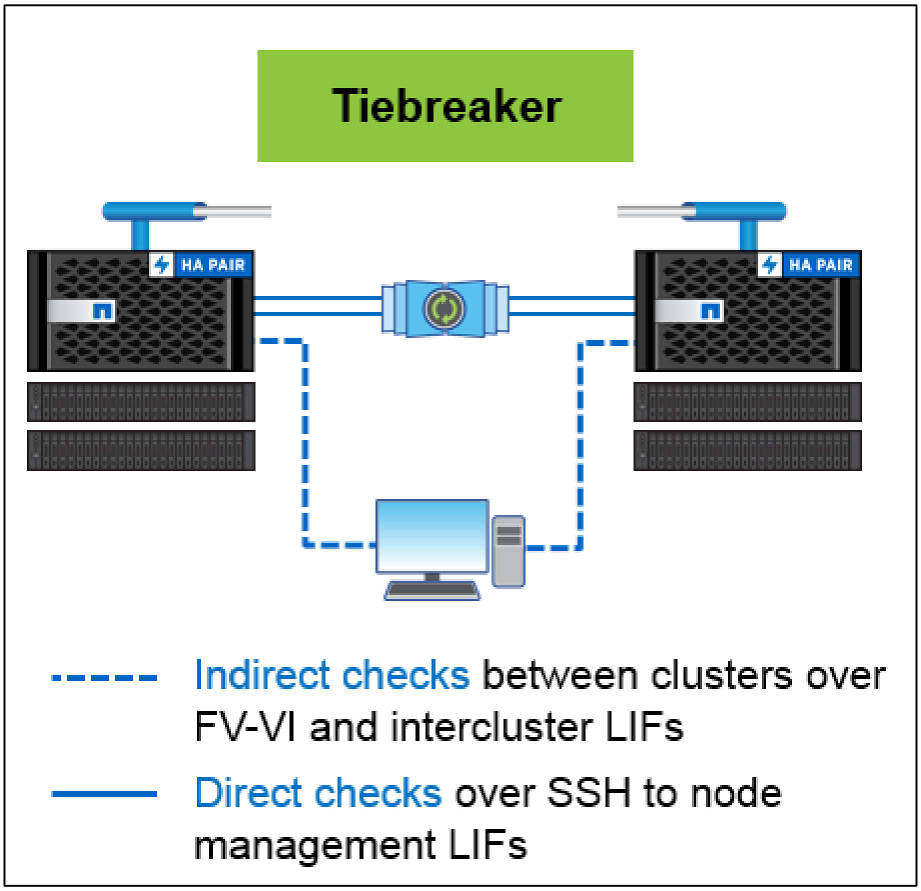
The MetroCluster configuration itself does not detect and initiate a switchover after site failure. You cannot rely on each site's monitoring the other for site failure. A lack of response from one cluster to the other could be caused by a genuine site failure or be caused by failure of all the intersite links. If all the links fail, the MetroCluster configuration continues to operate, providing local I/O but without any remote synchronization. After at least one intersite link is restored, replication automatically resumes and catches up with any changes that occurred in the interim. An automatic switchover is not desirable in this scenario because each cluster thinks that the other has failed, and both might try to perform the switchover, leading to the scenario known as "split brain."

The need for switchover, then, can be either human determined or application led. NetApp provides a fully supported capability, MetroCluster Tiebreaker software, which is installed at a third site with independent connections to each of the two clusters. The purpose of the Tiebreaker software is to monitor and detect both individual site failures and intersite link failure. MetroCluster Tiebreaker software can raise an SNMP alert in the event of a site disaster. It operates in observer mode and can detect and send an alert if a disaster requiring switchover occurs. The switchover then can be issued manually by the administrator. Optionally, through a policy-variance request, Tiebreaker software can be configured to automatically issue the command for switchover on disaster.

To create an aggregated, logical view of a site's availability, Tiebreaker software monitors relevant objects at the node, HA pair, and cluster level. It uses a variety of direct and indirect checks on the cluster hardware and links to update its state, as shown in Figure 18. The update indicates whether Tiebreaker detected an HA takeover event, a site failure, or failure of all intersite links.

A direct link check is through SSH to a node's management LIF. Failure of all direct links to a cluster indicates a site failure, which is characterized by the cluster's ceasing to serve any data (all SVMs are down). An indirect link check determines whether a cluster can reach its peer by any of the intersite (FC-VI) links or intercluster LIFs. If the indirect links between clusters fail and the direct links to the nodes succeed, this situation indicates that the intersite links are down and that the clusters are isolated from each other. MetroCluster continues operating in this scenario.

Figure 18) MetroCluster Tiebreaker software operation.

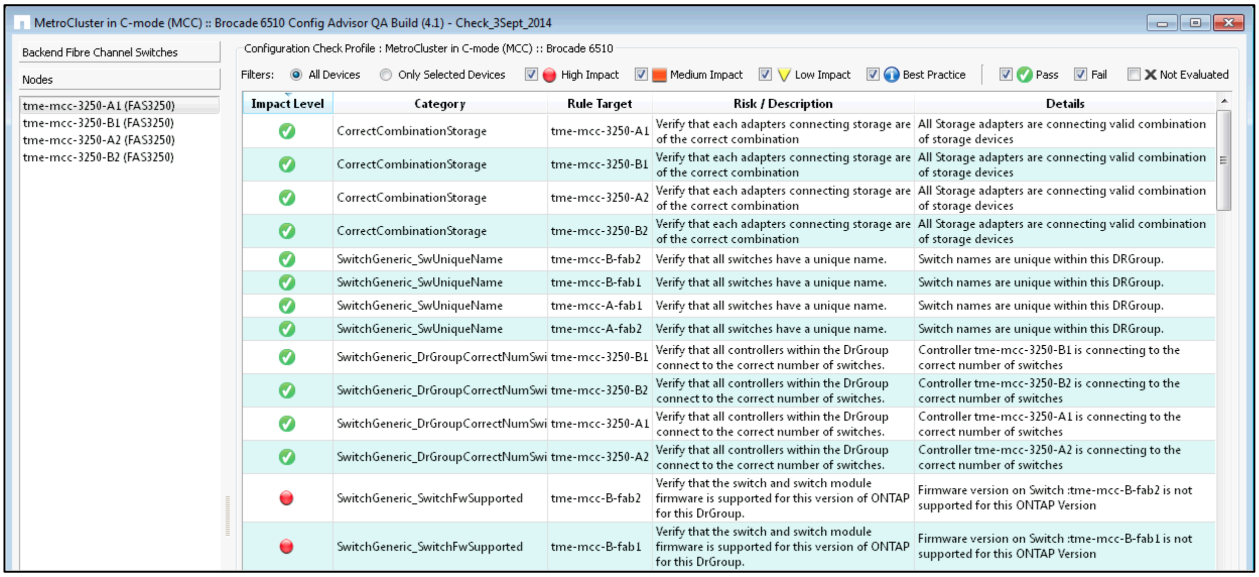

Tiebreaker
HA PAIR
HA PAIR
----- Indirect checks between clusters over FV-VI and intercluster LIFs
—— Direct checks over SSH to node management LIFs

MetroCluster Tiebreaker software is distributed on the NetApp Support site and is a standalone application that runs on a Linux host or VM. To download MetroCluster Tiebreaker software, go to the software downloads section of the NetApp Support site at http://mysupport.netapp.com/NOW/cgi-bin/software/ and select MetroCluster Tiebreaker. The "Installation and Configuration Guide" and "System Prerequisites" are also available at this link. No additional clustered Data ONTAP license is required for the Tiebreaker software.

4.5 Config Advisor

Config Advisor is a configuration validation and health check tool for NetApp systems. Config Advisor 4.1, and later, supports MetroCluster configurations, with over 40 rules that are specific to MetroCluster. To download Config Advisor, the MetroCluster plug-in, and documentation, go to the Config Advisor page of the NetApp Support site. An example of output when Config Advisor is run against a MetroCluster system is shown in Figure 19.

Figure 19) Config Advisor sample output.



MetroCluster in C-mode (MCC) :: Brocade 6510 Config Advisor QA Build (4.1) - Check_3Sept_2014

Backend Fibre Channel Switches
Nodes
tme-mcc-3250-A1 (FAS3250)
tme-mcc-3250-B1 (FAS3250)
tme-mcc-3250-A2 (FAS3250)
tme-mcc-3250-B2 (FAS3250)

Configuration Check Profile : MetroCluster in C-mode (MCC) :: Brocade 6510

Filters: ● All Devices ○ Only Selected Devices ☑ High Impact ☑ Medium Impact ☑ Low Impact ☑ Best Practice ☑ Pass ☑ Fail □ Not Evaluated

Impact Level	Category	Rule Target	Risk / Description	Details
✓	CorrectCombinationStorage	tme-mcc-3250-A1	Verify that each adapters connecting storage are of the correct combination	All Storage adapters are connecting valid combination of storage devices
✓	CorrectCombinationStorage	tme-mcc-3250-B1	Verify that each adapters connecting storage are of the correct combination	All Storage adapters are connecting valid combination of storage devices
✓	CorrectCombinationStorage	tme-mcc-3250-A2	Verify that each adapters connecting storage are of the correct combination	All Storage adapters are connecting valid combination of storage devices
✓	CorrectCombinationStorage	tme-mcc-3250-B2	Verify that each adapters connecting storage are of the correct combination	All Storage adapters are connecting valid combination of storage devices
✓	SwitchGeneric_SwUniqueName	tme-mcc-B-fab2	Verify that all switches have a unique name.	Switch names are unique within this DRGroup.
✓	SwitchGeneric_SwUniqueName	tme-mcc-B-fab1	Verify that all switches have a unique name.	Switch names are unique within this DRGroup.
✓	SwitchGeneric_SwUniqueName	tme-mcc-A-fab1	Verify that all switches have a unique name.	Switch names are unique within this DRGroup.
✓	SwitchGeneric_SwUniqueName	tme-mcc-A-fab2	Verify that all switches have a unique name.	Switch names are unique within this DRGroup.
✓	SwitchGeneric_DrGroupCorrectNumSwi	tme-mcc-3250-B1	Verify that all controllers within the DrGroup connect to the correct number of switches.	Controller tme-mcc-3250-B1 is connecting to the correct number of switches
✓	SwitchGeneric_DrGroupCorrectNumSwi	tme-mcc-3250-B2	Verify that all controllers within the DrGroup connect to the correct number of switches.	Controller tme-mcc-3250-B2 is connecting to the correct number of switches
✓	SwitchGeneric_DrGroupCorrectNumSwi	tme-mcc-3250-A1	Verify that all controllers within the DrGroup connect to the correct number of switches.	Controller tme-mcc-3250-A1 is connecting to the correct number of switches
✓	SwitchGeneric_DrGroupCorrectNumSwi	tme-mcc-3250-A2	Verify that all controllers within the DrGroup connect to the correct number of switches.	Controller tme-mcc-3250-A2 is connecting to the correct number of switches
●	SwitchGeneric_SwitchFwSupported	tme-mcc-B-fab2	Verify that the switch and switch module firmware is supported for this version of ONTAP for this DrGroup.	Firmware version on Switch :tme-mcc-B-fab2 is not supported for this ONTAP Version
●	SwitchGeneric_SwitchFwSupported	tme-mcc-B-fab1	Verify that the switch and switch module firmware is supported for this version of ONTAP for this DrGroup.	Firmware version on Switch :tme-mcc-B-fab1 is not supported for this ONTAP Version

4.6 Quality of Service (QoS)

QoS can be used in MetroCluster configurations to extend its typical use cases in a Data ONTAP cluster. QoS policies can be dynamically applied and modified as necessary. Some examples for using QoS in MetroCluster environments are:

- In normal operation when both clusters are active, QoS policies can be applied if periods of high traffic over the ISLs are observed. Limiting the application I/O necessarily lowers the ISL traffic for disk and NVRAM replication and prevents temporary overloading of the ISLs.
- When the configuration is running in switchover mode, fewer system resources are available because only half the nodes are active. Depending on the headroom applied to the system sizing, the reduction in available resources could affect client and application workloads. QoS policies can be configured to apply a ceiling (input/output operations per second [IOPS] or throughput) to noncritical workloads to provide more resource availability to critical workloads. The policies can be disabled after switchback when normal operation is resumed.

For more information about using QoS, see the white paper Clustered Data ONTAP Quality of Service Performance Isolation for Multi-Tenant Environments. See also the section “Managing Workload Performance by Using Storage QoS” in the “Data System Administration Guide for Cluster Administrators” in the clustered Data ONTAP product documentation.

4.7 SnapMirror and SnapVault

SnapMirror and SnapVault relationships can be created by using MetroCluster protected volumes as either the source or the destination. The data protection relationships can be within the MetroCluster environment (in the same or the other cluster) or to and from other clustered Data ONTAP clusters without MetroCluster. The following considerations apply when creating these relationships with MetroCluster protected volumes:

- When you use a cluster separate from the MetroCluster configuration as either the source or the destination of the relationship, you must peer both clusters to it. This step is necessary so that replication can continue after a switchover or a switchback. For example, if the MetroCluster environment consists of clusters A and B and a volume on cluster A is mirrored by using SnapMirror to cluster E, then both cluster A and cluster B must be peered with cluster E.
- You can run SnapMirror or SnapVault operations only from the cluster running the SVM that contains the volume. Because volumes are online only on one cluster at a time, you cannot use the copy of the volume that was mirrored with SyncMirror on the other cluster for any purpose. This limitation includes NetApp FlexClone®, SnapMirror, SnapVault, or any other operation that requires the volume to be brought online. The other cluster accesses the volumes only when a switchover is performed.

  In our preceding example, in steady state, the SnapMirror relationship is from the volume in cluster A to a volume in cluster E. Even though the volume is present offline in the equivalent `sync_destination` SVM on cluster B, that volume is not available for use.
- For MetroCluster volumes as a SnapMirror or SnapVault **source**, the peering relationships are automatically updated on switchover or switchback, and replication resumes automatically at the next scheduled time. Manually initiated replication operations must be explicitly restarted.
- For MetroCluster volumes as a SnapMirror or SnapVault **destination**, verifying and re-creating the peering relationships are necessary after switchover and switchback. Each replication relationship must be re-created after every switchover or switchback by using `snapmirror create`. A SnapMirror or SnapVault rebaseline is not required. A WFA workflow, “Re-create SnapMirror and SnapVault Protection After MetroCluster Switchover and Switchback,” is available to automate the re-creation of the relationships.

4.8 SnapLock

MetroCluster operating in clustered Data ONTAP does not support NetApp SnapLock® software.

4.9 Volume Move

Data mobility by using volume move (NetApp DataMotion™ for Volumes) is one of the core clustered Data ONTAP nondisruptive operations. Nondisruptively moving volumes helps to balance capacity and performance, as well as technology refresh. Volumes can be nondisruptively moved between any aggregates in a cluster, while remaining within the same SVM. Volume move cannot be used to transfer a volume to another SVM or to another cluster.

Volume move is initiated on the cluster that owns the volume. In a MetroCluster environment, the local and remote plexes in the source and destination aggregates are automatically synchronized by using SyncMirror. When the volume move completes, its new aggregate location is propagated to the other cluster through the cluster peering network.

If a volume move job is in progress when a switchover command is issued and the critical cutover phase of the job has not been reached, the job is automatically terminated. You must then manually delete the associated temporary (TMP) volume and restart the volume move job from the beginning. If, however, the commit phase has been reached, the job commit phase resumes after the aggregates are switched over. In this case, an EMS is logged, and the original source volume must be manually deleted.

It is not possible to do a switchback while volume move is in flight. The switchback command is vetoed until all volume move jobs are complete.

An MDV can be moved in advanced mode, but NetApp recommends that you do so only with the guidance of NetApp Support. The following warning message is issued. If the operation is confirmed, the volume move proceeds.

```
tme-mcc-A::*> vol move start -vserver tme-mcc-A -
volume  MDV_CRS_e8fef00df27311e387ad00a0985466e6_A -destination-aggregate aggr1_tme_mcc_B1
```

```
Warning: You are about to modify the system volume
"MDV_CRS_e8fef00df27311e387ad00a0985466e6_A".  This may cause severe performance or stability
problems.  Do not proceed unless directed to do so by support.  Do you want to proceed? {y|n}:
```

The main reason for a volume move of an MDV is that the aggregate containing it must be deleted, for example, when replacing a storage shelf. Because this action requires complete evacuation of the aggregate, the MDV should be moved to another aggregate that does not already contain an MDV. For resiliency reasons, the two MDVs in each cluster should reside on separate aggregates, preferably on separate nodes.

4.10 Volume Rehost

MetroCluster operating in clustered Data ONTAP does not support volume rehost.

4.11 Flash Pool

Flash Pool is supported in MetroCluster configurations. Each Flash Pool aggregate must contain an identical plex on each site. The maximum usable cache size in Flash Pool in a MetroCluster DR group is half the supported size of an equivalent model HA pair. For example, with a FAS8080 without MetroCluster, the maximum cache size is 144TiB. In MetroCluster, it is 72TiB.

Flash Pool operation is transparent to MetroCluster. The cache is kept in sync between the two plexes, just as for any other aggregate. If switchover is performed, the Flash Pool cache is in sync and is therefore warm.

The aggregate NetApp Snapshot copy resync time is the time interval between automatic aggregate Snapshot copies. It should be set to 5 minutes (from the default of 60 minutes) for Flash Pool aggregates that use SyncMirror or MetroCluster. This interval prevents data from being pinned longer than needed in flash storage. Use the following command:

```
storage aggregate modify -aggregate <aggrname> -resyncsnaptime 5
```

Advanced disk partitioning for SSDs is not supported with Flash Pool in MetroCluster configurations.

4.12 All Flash FAS (AFF)

NetApp All Flash FAS systems use 8000 series controller hardware and the clustered Data ONTAP 8.3.x operating system. They deliver submillisecond latency and up to 4 million IOPS throughput for business applications that demand exceptionally fast response times. The systems deliver high application data throughput at low latency while providing leading data management features.

All Flash FAS systems use high-performance SSDs – available in the capacity points 400GB, 800GB, 800GB NSE, and 1.6TB – which provide configuration flexibility to meet specific cost and density needs.

MetroCluster is fully compatible with All Flash FAS to deliver high performance and low latency. All Flash FAS is transparent to MetroCluster and is supported on all two-node and four-node MetroCluster configurations.

4.13 7-Mode Transition Tool (7MTT)

7MTT 2.0 or later supports transition from Data ONTAP operating in 7-Mode (including fabric MetroCluster and stretch MetroCluster) to MetroCluster 8.3.x. For 7-Mode MetroCluster migration to clustered Data ONTAP, you can use the 7MTT copy-based transition (CBT) method. With CBT, data is copied from the 7-Mode controllers to the MetroCluster 8.3.x destination cluster. Copying a volume into MetroCluster 8.3.x automatically syncs the data to the remote plex. The 7MTT prechecks advise that any volumes that are in unmirrored aggregates be transitioned to mirrored aggregates. The cutover requires a brief application outage. For more information, see the 7MTT Copy-Based Transition Guide.

Note: Copy-free transition in 7MTT is not supported for MetroCluster.

5 Transition to MetroCluster

For the purposes of this technical report, *transition* refers to the process of moving to a NetApp MetroCluster configuration from the following environments:

- 7-Mode configuration without MetroCluster
- MetroCluster 8.2.1 and earlier (7-Mode), either fabric or stretch MetroCluster
- Clustered Data ONTAP

This technical report does not consider transition from third-party systems. The general process is to set up the MetroCluster 8.3.x environment and then migrate the data and the configuration.

The MetroCluster 8.3.x environment must be set up from scratch and be configured on a clean system. The `metrocluster configure` command returns an error if data exists on the nodes. Existing NetApp clustered Data ONTAP environments cannot be converted in place to operate as a MetroCluster configuration. To transition to MetroCluster 8.3.x, there are two basic possibilities, depending on whether you intend to reuse existing equipment in the new MetroCluster environment:

- If the existing hardware is not supported with MetroCluster 8.3.x or if it is time to refresh the hardware, purchase a new MetroCluster configuration and copy over the data.
- If some or all of the existing hardware is supported with MetroCluster 8.3.x, purchase sufficient additional equipment (including switches, bridges, controllers, and storage) to create a four-node MetroCluster environment. As always, see the Interoperability Matrix Tool to verify that existing equipment is supported in the new release. Then use one of the following methods to make the transition:
  - Copy all data from the existing environment to a temporary location. Set up MetroCluster with the combined existing and new equipment and restore the data.
  - If suitable temporary equipment is available, you can stage it into the current environment, freeing up the equipment to create the new MetroCluster environment. Alternatively, you can use it to create the new MetroCluster environment in conjunction with the new equipment. These scenarios are explored in sections 5.1 and 5.2.
  - You must add the correct cards to reused controllers if necessary to support MetroCluster, in particular, a 16Gb FC-VI card and four FC initiator ports per controller.

In all cases, the data must be copied over to the new MetroCluster environment that runs clustered Data ONTAP 8.3.x.

5.1 Copying Data into MetroCluster 8.3.x

This section presents guidance on various scenarios for migrating to a MetroCluster environment. Because a number of variations are possible, NetApp highly recommends that you use professional services to develop detailed plans. The method for copying data and migrating the configuration depends on the source system.

Migrating from Clustered Data ONTAP

Host- or application-based methods or NetApp SnapMirror can be used to copy data from clustered Data ONTAP to a MetroCluster configuration. SnapMirror can copy only 64-bit source data, including all NetApp Snapshot copy data. For more information, see TR-3978: 64-Bit Aggregates – Overview and Best Practices.

Hardware Reuse Scenarios

If you plan to reuse hardware after evacuation, first verify that the hardware is supported with MetroCluster 8.3.x. Temporary nodes and storage are required.

If the temporary nodes are supported in the existing cluster as per the cluster mixing rules, the following process can be used. This approach is a good strategy if the available temporary nodes are not the same model as the existing nodes, because all nodes in the MetroCluster environment must be of the same type:

1. Add two temporary nodes and sufficient storage to the current clustered Data ONTAP cluster.
2. Use volume move and LIF migrate to nondisruptively evacuate the data from the original nodes to the temporary nodes.
3. Remove the evacuated nodes from the cluster.
4. Wipe clean the evacuated nodes and reinstall and configure them in the new MetroCluster environment. Add nodes and storage as necessary to make up the DR group as well as the FC or FCIP switches and FibreBridges.
5. Migrate the data from the temporary nodes in the clustered Data ONTAP cluster to the new MetroCluster environment.
6. After cutover, you can return the temporary nodes.

Following is an alternative method for migrating to MetroCluster from clustered Data ONTAP by using temporary equipment:

1. Install the MetroCluster system by using temporary nodes or a combination of temporary and purchased nodes. All nodes in an operational MetroCluster configuration must be the same controller model. For example, suppose that four existing FAS8080s are to be redeployed as two clusters for the MetroCluster configuration. In that case, four temporary controllers are required for the initial MetroCluster configuration. They do not have to match the existing controller type (FAS8080 in this case), but all four must be the same model (for example, four FAS8060s).

   A more common scenario is when an existing two-node clustered Data ONTAP cluster is to be transitioned to MetroCluster with the purchase of two new nodes to be installed at the second site. In this case, assuming that FAS8080 controllers are used, the MetroCluster configuration is initially deployed with temporary FAS8080s at one site and two purchased FAS8080s at the other site. The temporary equipment must match the existing model type.
2. Migrate the data from clustered Data ONTAP to the new MetroCluster environment.
3. Replace the temporary controllers in the MetroCluster environment with the evacuated original clustered Data ONTAP nodes. If the existing and temporary nodes are the same model, use takeover and giveback. If the nodes are different models, use the aggregate relocation method.
4. If temporary storage is being used in the MetroCluster environment, attach the original storage to the nodes and use volume move and nondisruptive shelf removal to evacuate the temporary storage.

A third method is the following:

1. Use aggregate relocation or takeover and giveback to upgrade or replace existing clustered Data ONTAP nodes with temporary nodes.
2. Install the MetroCluster system with the freed-up nodes and new or temporary storage.
3. Migrate the data from clustered Data ONTAP to the new MetroCluster environment.
4. If temporary storage is being used in the MetroCluster environment, attach the original storage to the nodes and use vol move and nondisruptive shelf removal to evacuate the temporary storage.

Migrating from Data ONTAP Operating in 7-Mode

Regardless of whether or not the 7-Mode system is running fabric MetroCluster (FMC) or stretch MetroCluster (SMC), the same methods are available for migrating or transitioning data to the MetroCluster environment.

Starting with 7-Mode Transition Tool (7MTT) 2.0, the 7MTT copy-based transition method supports transition from 7-Mode (both fabric and stretch MetroCluster) to MetroCluster 8.3.x. Data is copied from the 7-Mode controllers to the MetroCluster 8.3.x destination cluster. Copying a volume into MetroCluster 8.3.x automatically syncs the data to the remote plex. Because MetroCluster 8.3.x supports only mirrored aggregates, the 7MTT prechecks advise that any volumes in unmirrored aggregates be transitioned to mirrored aggregates. The cutover requires a brief application outage. 7MTT copy-free transition is not supported for data migration to MetroCluster operating on clustered Data ONTAP.

Application-, client-, and host-level methods, such as rsync or VMware Storage vMotion, can also be used for data migration. For more information, consult TR-4052: Successfully Transitioning to Clustered Data ONTAP and TR-4336: Enterprise Application Transition to Clustered Data ONTAP 8.3.

Clustered Data ONTAP 8.3.x supports only 64-bit aggregates, volumes, and Snapshot copies. If you use 7MTT or SnapMirror, you must convert all data to 64-bit before you copy it into the MetroCluster environment. Other methods that are not based on SnapMirror are not subject to this requirement. For more information, see TR-3978: 64-Bit Aggregates – Overview and Best Practices. If you use 7MTT, you must first upgrade the 7-Mode environment to a release that supports migration to clustered Data ONTAP 8.3.x. The transition collateral referred to in this section has more information.

Brocade 6510 switches that are currently deployed in a fabric MetroCluster environment, and their connecting ISLs, can be shared for transition purposes. During transition from one FMC environment, a single MetroCluster 8.3.x configuration can be connected to the same Brocade 6510 switches that share the existing ISLs. Sufficient bandwidth must be available to support the resultant traffic, and the ISLs must be at a speed that is supported with MetroCluster 8.3.x (minimum 4Gbps for each ISL).

The load imposed by the data copy is in addition to the existing application workload on the 7-Mode system, so you should be careful about the amount of bandwidth used. If the ISLs show signs of saturation (for example, if application throughput is dropping or client and host latencies are increasing), throttling can be used dynamically with 7MTT to slow the pace of replication. To reduce the overall demand on the ISLs, NetApp recommends performing the initial baseline data transfer at less busy times.

Another technique to spread out the traffic is to disable remote plexes in the MetroCluster 8.3.x environment by powering off the disks while the initial data transfer is taking place. When the disks are later powered on (after data transfer from the 7-Mode environment is complete), the aggregates automatically resync over the ISLs.

The MetroCluster Installation and Configuration Guide describes how to share Brocade 6510 switches between the two environments. After the transition is complete, the fabric MetroCluster configuration is disconnected from the fabric. The MetroCluster configuration can remain on its current switchports. If you want to move the MetroCluster 8.3.x configuration to the ports that FMC originally used, a disruption is required.

Only Brocade 6510 switches are supported for fabric sharing during transition; for any other switch model used with FMC, separate switches and ISLs are required for the MetroCluster 8.3.x configuration. **Only one** FMC can share the switches with one MetroCluster 8.3.x configuration.

Note: In FMC, 8Gb SFPs are used with the Brocade 6510 switch. MetroCluster 8.3.x supports only 16Gb SFPs for the 6510, and all components must connect to switchports with 16Gb SFPs, including the FC initiators, FC-VI ports, and ATTO bridges. The only exception to the 16Gb requirement is for the switchports that attach to the ISLs. Therefore, the requisite number of 16Gb SFPs must be purchased to attach the MetroCluster 8.3.x components to a 6510. No extra license is required to enable support of 16Gb on the switch. Ports 23 and above are used to attach the MetroCluster 8.3.x components, as documented in the MetroCluster Installation and Configuration Guide. If these ports were not yet enabled, the requisite Ports on Demand (PoD) license is required to activate them.

5.2 Hardware Reuse Transition Scenarios

Similarly to migration from a clustered Data ONTAP environment, different migration methods are available if you plan to reuse hardware after evacuation. As always, verify that the hardware is supported with MetroCluster 8.3.x. Temporary nodes and storage are required.

Following is one method for transitioning from a 7-Mode MetroCluster configuration. This method is appropriate when the temporary equipment that is available is not the same controller model as for the original nodes:

1. Commission a second 7-Mode MetroCluster environment with temporary equipment.
2. Copy the data from the original 7-Mode MetroCluster configuration to the new MetroCluster environment by using SnapMirror or other utilities (7-Mode to 7-Mode). Depending on the switches that you use, it might be possible to share the same switches between the 7-Mode environments.
3. Wipe clean the evacuated nodes, reinstall them, and configure them in the new MetroCluster 8.3.x environment. Add nodes and storage as necessary to make up the DR group as well as the FC or FCIP switches and FibreBridges. The switches and bridges can be reused if they are available and supported in MetroCluster 8.3.x. Brocade 6510 switches can be shared between the 7-Mode and the clustered Data ONTAP MetroCluster systems during transition.
4. Migrate the data to MetroCluster 8.3.x and cut over.

The following is an alternative method when there is an immediate need to upgrade an existing 7-Mode MetroCluster configuration and transition to MetroCluster 8.3.x at a later time. This method is also for situations in which the existing FMC or SMC was recently upgraded to FAS80x0 nodes that can be reused in MetroCluster 8.3.x:

1. Purchase a FAS80x0 and upgrade the existing FMC or SMC controllers. A period of time can elapse before you decide to move to MetroCluster 8.3.x; however, it must still be possible to purchase matching controller models because four identical nodes are required.
2. When you are ready to move to MetroCluster 8.3.x, purchase an additional FAS80x0 + chassis for each site.
3. Deploy MetroCluster 8.3.x with two temporary FAS80x0 nodes in one cluster and two purchased FAS80x0 nodes in the other cluster. FibreBridges are also required, and you need additional switches, unless you intend to share Brocade 6510 switches from the FMC environment.
4. Transition the data and cut over to MetroCluster, decommissioning the 7-Mode environment.
5. Replace the temporary FAS80x0 controllers with the freed-up equipment from the 7-Mode environment by using takeover and giveback.

   Note: All new and temporary nodes must be the same model.

When you plan to reuse equipment from 7-Mode in MetroCluster 8.3.x, verify that all reused hardware components (controllers, cards, storage, switches, ISL speed) are supported with clustered Data ONTAP 8.3.x. Additional or replacement parts might be required. All nodes must be the same controller model. Only the 16Gb FC-VI card (X1928) is supported with MetroCluster 8.3.x; the 8Gb FC-VI card (X1927) is not supported. You must have 16Gb SFPs for all ports on Brocade 6510 switches, apart from ISL ports. The Cisco 9148 switch is 8Gb only, so SFPs do not need to be replaced. The appropriate switch licenses must also be available. As always, check with the Interoperability Matrix Tool to verify supported configurations, including ISL speeds.

Additional Resources

- NetApp MetroCluster Product Documentation
  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=30022
- Hardware Universe
  https://hwu.netapp.com/Controller/Index
- MetroCluster Service Guide
  https://library.netapp.com/ecm/ecm_download_file/ECMP1650547
- MetroCluster Management and Disaster Recovery Guide
  https://library.netapp.com/ecm/ecm_download_file/ECMP12458277
- MetroCluster Installation and Configuration Guide
  https://library.netapp.com/ecm/ecm_download_file/ECMP12454947
- Tiebreaker Software Installation and Configuration Guide
  https://library.netapp.com/ecm/ecm_download_file/ECMP12007400
- NetApp Interoperability Matrix Tool
  https://mysupport.netapp.com/matrix/
- TR-3548: MetroCluster Version 8.2.1 Best Practices for Implementation (7-Mode)
  http://www.netapp.com/us/media/tr-3548.pdf
- TR-3982: NetApp Clustered Data ONTAP 8.3.x and 8.2.x: An Introduction
  http://www.netapp.com/us/media/tr-3982.pdf
- Cisco Storage Interoperability Matrix
  http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/interoperability/matrix/intmatrx/Matrix1.html

Contact Us

Let us know how we can improve this technical report.

Contact us at docfeedback@netapp.com.

Include TECHNICAL REPORT 4375 in the subject line.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

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.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.277-7103 (October 1988) and FAR 52-227-19 (June 1987).

Trademark Information