



Technical Report

# NetApp MetroCluster FC

Cheryl George, NetApp  
October 2021 | TR-4375

## Abstract

This document provides technical information about NetApp® MetroCluster FC software in a system that is run by NetApp ONTAP® data management software.

## TABLE OF CONTENTS

<b>Introduction to NetApp MetroCluster .....</b>	<b>4</b>
Features .....	4
New features in MetroCluster in ONTAP 9 .....	5
Architecture and supported configurations .....	5
MetroCluster replication.....	15
<b>Initial MetroCluster setup.....</b>	<b>17</b>
Hardware and software requirements.....	17
Summary of installation and setup procedure .....	23
Post setup configuration and administration.....	24
<b>Resiliency for planned and unplanned events .....</b>	<b>28</b>
MetroCluster with unplanned and planned operations.....	28
Performing planned (negotiated) switchover .....	31
Performing switchback .....	35
Performing forced switchover .....	39
Protecting volumes after forced switchover .....	40
Recovery from a forced switchover, including complete site disaster.....	40
<b>Interoperability .....</b>	<b>41</b>
ONTAP System Manager .....	41
Active IQ Unified Manager and health monitors .....	42
AutoSupport.....	44
MetroCluster Tiebreaker software .....	44
Config Advisor .....	46
Quality of service .....	46
SnapMirror Asynchronous data replication .....	47
SVM DR .....	47
SnapLock .....	48
Volume move.....	48
Volume rehost .....	48
FlexGroup.....	48
FlexCache .....	49
Flash Pool .....	49
NetApp AFF A-Series arrays .....	49
<b>Where to find additional information .....</b>	<b>49</b>

<b>Contact Us .....</b>	<b>50</b>
<b>Version History .....</b>	<b>50</b>

## LIST OF TABLES

Table 1) Required hardware components. ....	11
Table 2) Recommended shelf numbering schema. ....	13
Table 3) Disk ownership changes. ....	14
Table 4) Example of disk ownership changes. ....	14
Table 5) ATTO 7600N shelf count. ....	20
Table 6) ATTO 7500N shelf count. ....	20
Table 7) Unplanned operations and MetroCluster response and recovery methods. ....	28
Table 8) Planned operations with MetroCluster. ....	30

## LIST OF FIGURES

Figure 1) Four-node MetroCluster FC configuration. ....	6
Figure 2) Two-node stretch MetroCluster configuration. ....	6
Figure 3) Two-node MetroCluster stretch-bridge configuration. ....	8
Figure 4) Two-node MetroCluster fabric-attached configuration. ....	9
Figure 5) Four-node MetroCluster fabric-attached configuration. ....	9
Figure 6) Eight-node MetroCluster fabric-attached configuration. ....	10
Figure 7) MetroCluster FC IP configuration in ONTAP 9.x. ....	11
Figure 8) MetroCluster four-node configuration local and remote pool layout. ....	12
Figure 9) MetroCluster pool-to-shelf assignment. ....	22
Figure 10) Cluster A: The <code>sync_source</code> SVM is unlocked, and the <code>sync_destination</code> SVM is locked. ....	41
Figure 11) Cluster B: The <code>sync_source</code> SVM is unlocked and the <code>sync_destination</code> SVM is locked. ....	42
Figure 12) SVMs after switchover: All SVMs are unlocked. ....	42
Figure 13) Active IQ Unified Manager device and link monitoring. ....	43
Figure 14) Active IQ Unified Manager replication monitoring. ....	44
Figure 15) MetroCluster Tiebreaker software operation. ....	45
Figure 16) Config Advisor sample output. ....	46
Figure 17) SVM DR with MetroCluster. ....	47

# Introduction to NetApp MetroCluster

NetApp MetroCluster provides continuous data availability across geographically separated data centers for mission-critical applications. MetroCluster continuous availability and disaster recovery software runs on ONTAP data management software. Fabric-attached and stretch MetroCluster configurations are 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 with FC for ONTAP 9, specifically fabric-attached MetroCluster and stretch MetroCluster deployments. Unless otherwise stated, the term MetroCluster in this paper refers to MetroCluster with FC in ONTAP 9.0 and later.

This document assumes that you are familiar with the ONTAP architecture and its capabilities. [The ONTAP documentation center](#) is a good starting point, and the [NetApp Field Portal](#) contains a large selection of technical reports to help you with in-depth information about specific ONTAP features.

## Features

In today's enterprise, the IT department must meet increasing service-level demands while maintaining cost and operational efficiency. As data volume explodes and as 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 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:
  - ONTAP storage software provides nondisruptive operations within the data center. It withstands component, node, and network failures and enables 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).
- Combined array-based clustering with synchronous mirroring to deliver zero data loss:
  - MetroCluster provides a recovery point objective (RPO); the maximum amount of acceptable data loss) of zero.
  - MetroCluster provides a recovery time objective (RTO) of 120 seconds or less for planned switchover and switchback. The RTO is the maximum acceptable time that is required to make storage and associated data available in the correct operational state after a switchover to the other data center.
- Reduced administrative overhead:
  - After initial setup, subsequent changes on one cluster are automatically replicated to the second cluster.
  - Ongoing management and administration are almost identical to an ONTAP environment by using NetApp ONTAP® System Manager and Active IQ® Unified Manager.
  - 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:** Note that SMB applications, including SMB3 with continuous availability shares, must reconnect after a switchover or a switchback. This need is a limitation of the SMB protocol.

- Features that complement the full power of 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 capabilities are integrated directly into ONTAP and require no additional licenses.
- 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.

## New features in MetroCluster in ONTAP 9

This feature list encompasses MetroCluster 9.0 through 9.7, and the latest release supports all previously introduced features:

ONTAP 9.8 includes the following features:

- Non-disruptively upgrade controllers (head upgrade) with simpler operations that reduce the possibility of error.
- Nondisruptive transition from a four-node fabric-attached configuration using FC switches and ATTO bridges to a four-node MetroCluster IP configuration.
- **ONTAP 9.7.** New platforms include AFF A400, FAS8300, and NetApp FlexCache® support.
- **ONTAP 9.6.** ATTO 7600N bridge, NetApp FlexGroup support, in-band monitoring of bridges.
- **ONTAP 9.5.** SVM-DR with MetroCluster as a source.
- **ONTAP 9.4.** ATTO 7500 bridge firmware update capability from ONTAP, additional platforms and features for MetroCluster IP.
- **ONTAP 9.3.** Introduction of MetroCluster IP (see [TR-4689 MetroCluster IP](#)) and MetroCluster Tiebreaker enhancements.
- **ONTAP 9.2.** Eight-node SAN support and a 500-volume count. One thousand volumes with five aggregates are supported with a Feature Policy Variance Request (FPVR).
- **ONTAP 9.1.** NetApp AFF and FAS system currencies and FC Inter-Switch Link (ISL) Out of Order Delivery.
- **ONTAP 9.0.** Eight-node MetroCluster NAS and unmirrored aggregates.

## Architecture and supported configurations

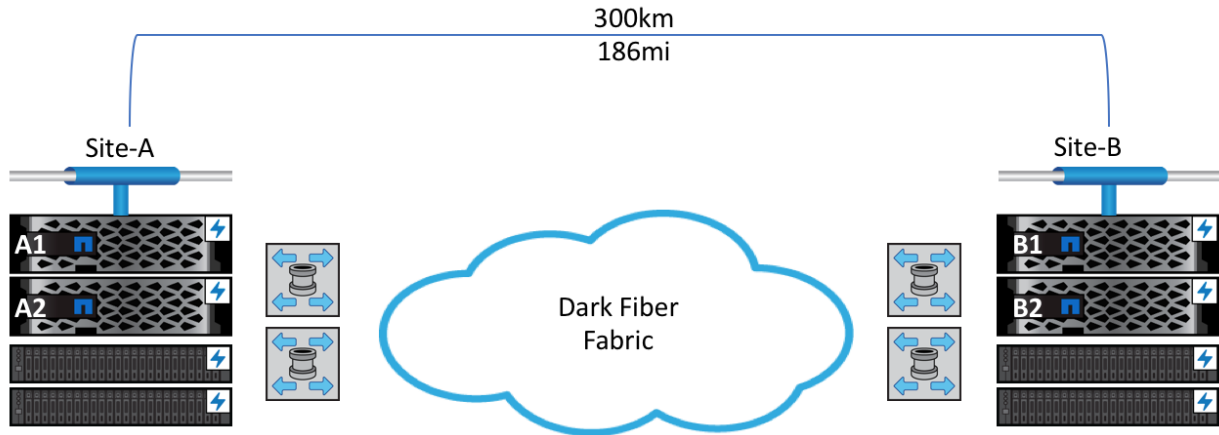
The architectural details described in this document are specific to MetroCluster with FC. References for MetroCluster IP and additional MetroCluster information are noted in the following section.

### Hardware configuration

MetroCluster installations require a fully redundant configuration with identical hardware present at each site. Figure 1 shows the core components and connections for a typical four-node configuration. For details about currently supported hardware components, consult the [Interoperability Matrix Tool for ONTAP 9.5 and older](#). For ONTAP 9.6 and newer, this information is located in the [Hardware Universe](#). This section details the deployment options for stretch, stretch-bridged, and fabric configurations.

**Note:** For simplicity, the technical diagrams in this document depict HA systems as two different controllers. HA systems, however, are a single chassis with redundant components.

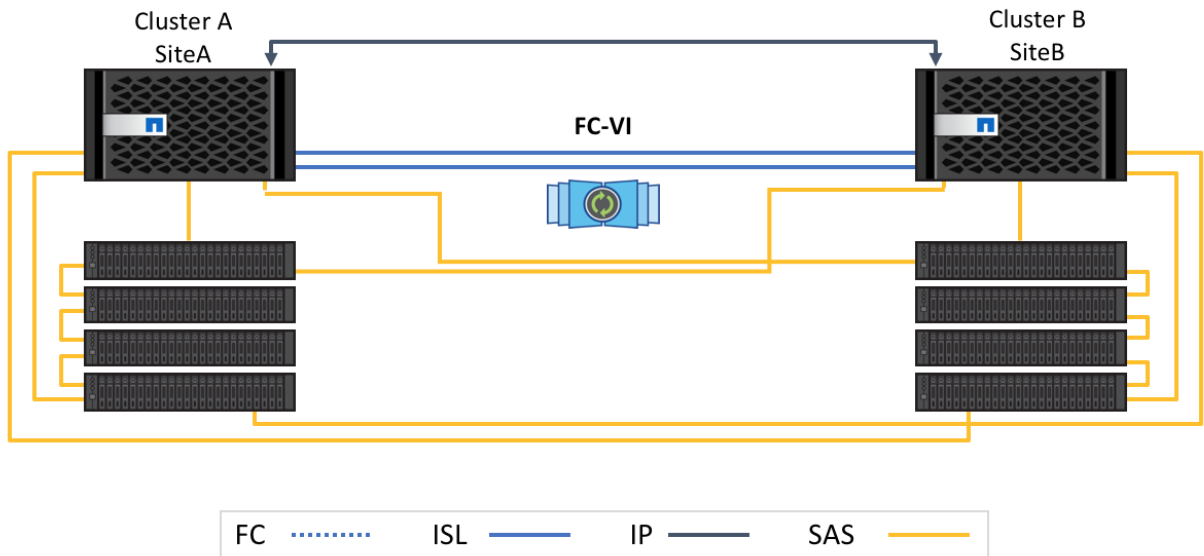
**Figure 1) Four-node MetroCluster FC configuration.**



### MetroCluster stretch configuration

A two-node, stretch, direct-attached configuration is intended for a rack-to-rack installation or between data halls in a data center. The distance between the nodes is limited to 100m with SAS-3 cabling over multimode fiber. A stretch deployment does not require FC or FC over IP (FCIP) switches or SAS-to-FC bridges. All connectivity to storage is by extended optical SAS or optical patch panel cables. All nodes in both clusters have visibility to all the storage. See Figure 2 for a two-node stretch MetroCluster configuration.

**Figure 2) Two-node stretch MetroCluster configuration.**



You should also consider the following issues regarding a stretch MetroCluster configuration:

- It requires a single node at each location.

- The stretch distance varies depending on the hardware, SAS, and disk shelves. For proper hardware provisioning and the maximum distance that is supported, see the [Interoperability Matrix Tool and Hardware Universe](#).
- FAS9000 and AFF A700 controllers require four virtual interface over Fibre Channel (FC-VI) ports per node, for a total of eight ISLs between the nodes. Note that both the FAS9000 and the AFF A700 require a minimum of six ISLs per fabric for NVRAM and NetApp SyncMirror® technology.
- For the FC-VI port connectivity, the distance for the SMC direct/bridge for any platform depends on the SFP type. You must determine the correct cable type. For example, if you use 8Gb SFP, OM3 cable supports 150m. For 8g long-wave (LW) SFP, 500m is supported. LW SFP in FMC is not supported.
- The stretch bridge is based on how the LR/SR SFP distance is determined. For example, ATTO FibreBridge 6500N supports only 8Gb short wave (SW) SFPs, but the ATTO FibreBridge 7500N supports 16Gb SW and LW SFPs. The ATTO FibreBridge 7600N supports 32Gb short-wave and long-wave SFPs.

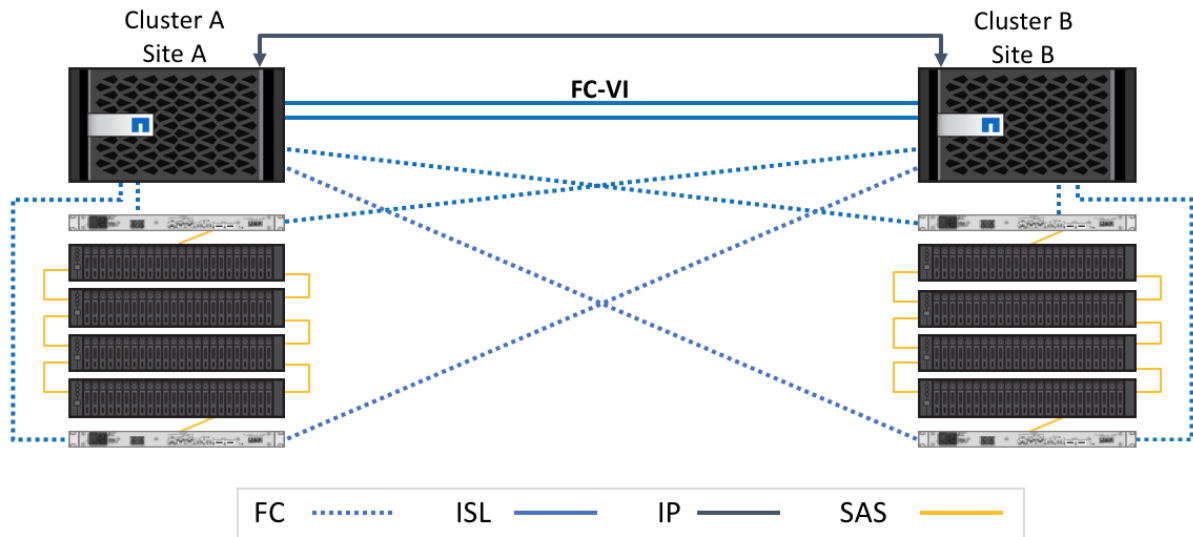
The following examples of SAS connectivity depend on the hardware and the optics that you use. For supported configurations, always see the IMT and [Hardware Universe](#):

- FAS9000 X92071A mini-SAS HD port—mini-SAS HD X66047A/X66048A LC—LC Multimode 100m cable LC—LC X66047A/ X66048A Mini-SAS HD port on DS212C, DS224C, or DS460C shelves.
- FAS8200 onboard mini-SAS HD port—mini-SAS HD X66047A/X66048A LC—LC Multimode 100m cable LC—LC X66047A/ X66048A mini-SAS HD port on DS212C, DS224C, or DS460C shelves.
- FAS8200 with X2069-R6 QSFP port—QSFP X66014A-R6 LC—LC Single Mode 500m cable LC—LC X66014A-R6 QSPF β> QSFP port on DS2246, DS4243, or DS4246 shelves.
- FAS80xx with X2069-R6 QSFP port—QSFP X66014A-R6 LC—LC Single Mode 500m cable LC—LC X66014A-R6 QSPF—QSFP port on DS2246, DS4243, or DS4246 shelves.
- FAS80xx onboard SAS QSFP port—QSFP X66014A-R6 LC—LC Single Mode 500m cable LC—LC X66014A-R6 QSPF—QSFP port on DS2246, DS4243, or DS4246 shelves.

### **MetroCluster stretch-bridge configuration**

A two-node stretch bridge-attached configuration, using SAS-to-FC FibreBridges, provides connectivity to the nodes that stretch beyond SAS distance capabilities. This design provides greater flexibility for deploying MetroCluster FC between buildings on a campus or on floors in the same building where connectivity beyond 100m is required. With this configuration, FC or FCIP switches are not required. All connectivity to the storage is with FC cables. All nodes in both clusters have visibility to all the storage. Figure 3 depicts the design for a two-node, stretch-bridged MetroCluster system.

Figure 3) Two-node MetroCluster stretch-bridge configuration.



If you are considering a MetroCluster stretch-bridge configuration, you should be aware of these additional details:

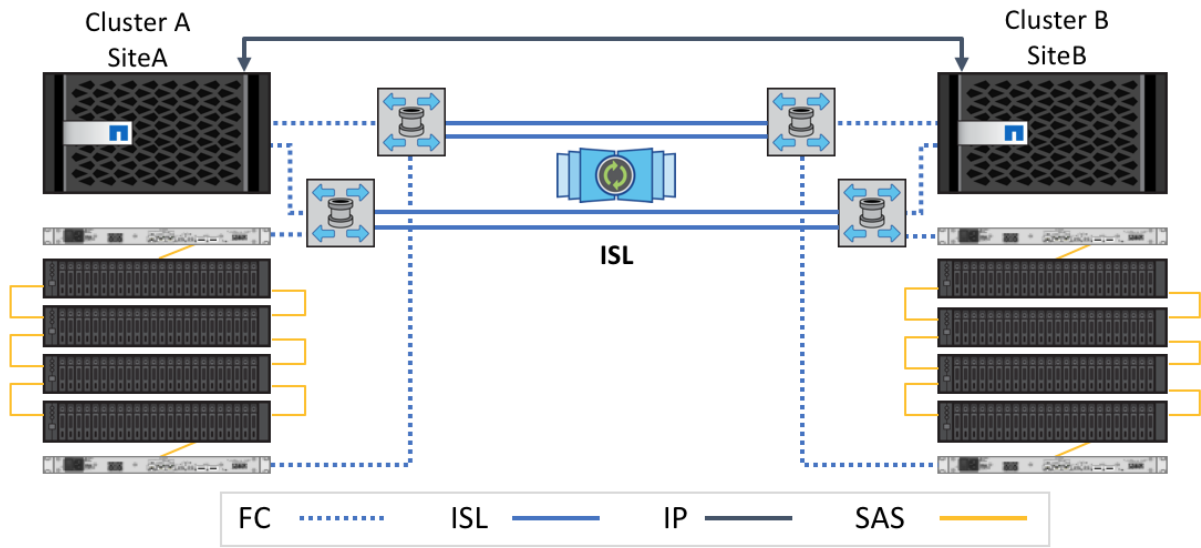
- A stretch configuration with ATTO 6500N can reach up to 270m.
- A stretch configuration with ATTO 7500N or 7600N can reach up to 500m.
- Fibre switches are not required for this design.
- FAS9000 and AFF A700 controllers require four FC-VI interfaces per node, for a total of four ISLs per fabric between the nodes.

### Two-node fabric-attached configuration

A two-node, fabric-attached configuration (Figure 4) with four FC or FCIP switches (two at each site) connects to the nodes through FC initiators and through FC-VI connections. This configuration also connects to storage through SAS-to-FC bridges. With this connectivity in place, all nodes in both clusters have visibility to all the storage. When using FC switches, this configuration has a cluster-to-cluster range of 185 miles (300km). For an FC-IP deployment, see the section “FCIP MetroCluster configuration.”



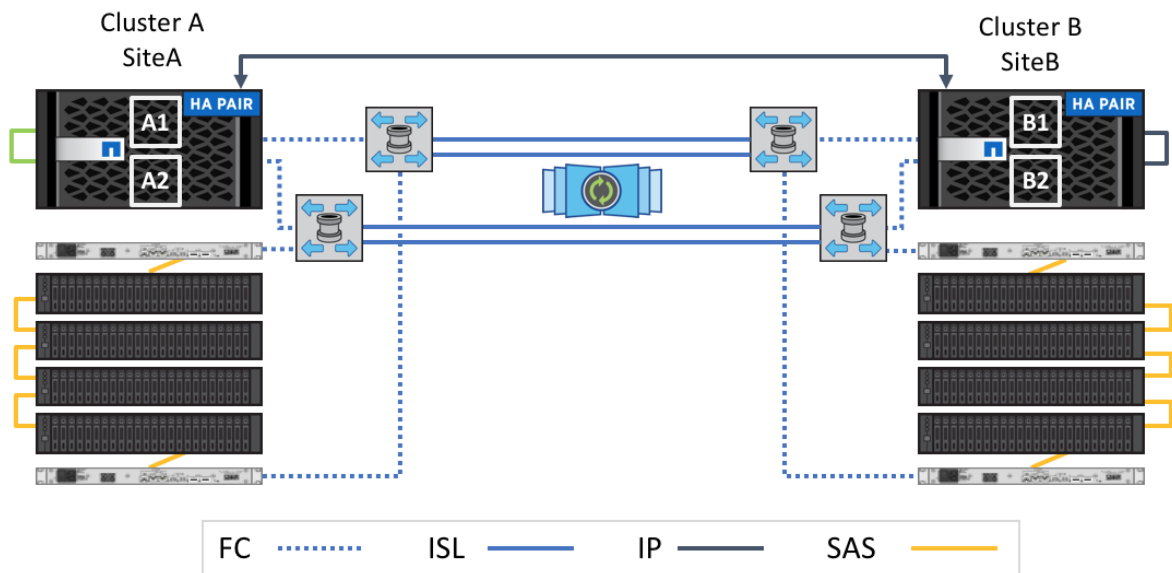
**Figure 4) Two-node MetroCluster fabric-attached configuration.**



**Four-node MetroCluster fabric-attached configuration**

In a four-node configuration, each cluster includes the standard NetApp ONTAP cluster interconnects. Typically, the configuration is a switchless or switched 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 to the storage through SAS-to-FC bridges. With this connectivity in place, all nodes in both clusters have visibility to all the storage. See Figure 5 for the configuration of a four-node fabric-attached MetroCluster system.

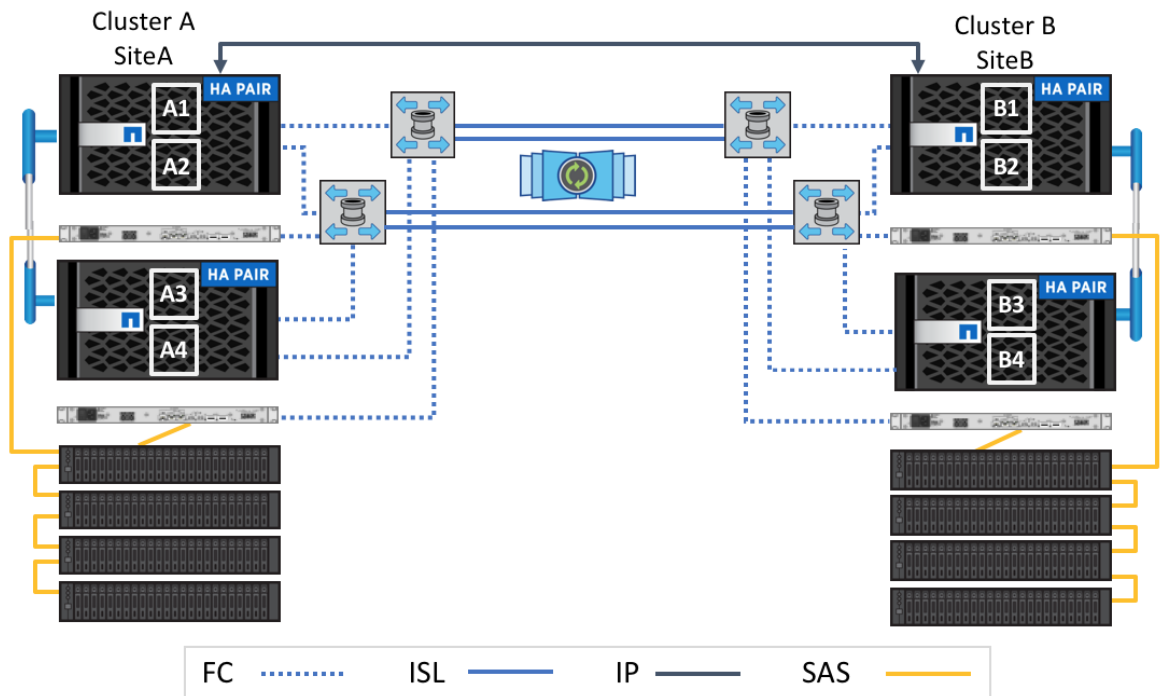
**Figure 5) Four-node MetroCluster fabric-attached configuration.**



## Eight-node MetroCluster fabric-attached configuration

An eight-node MetroCluster configuration scales to two HA pairs at each site, creating two logical disaster recovery (DR) groups. Each DR group must have identical hardware within the group, but hardware can be different between the DR sites. This approach allows greater flexibility in mixing AFF and FAS controllers in the same MetroCluster cluster. See Figure 6) for the configuration of an eight-node fabric-attached MetroCluster system.

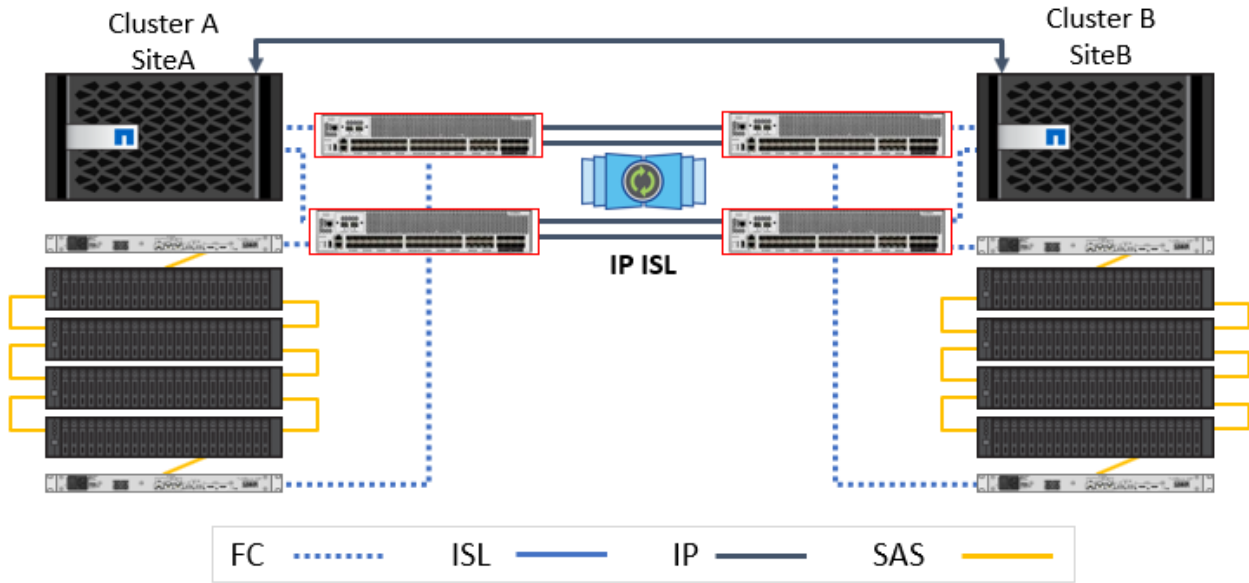
Figure 6) Eight-node MetroCluster fabric-attached configuration.



## FCIP MetroCluster configuration

In an FCIP configuration, MetroCluster uses an IP ISL to connect to the remote MetroCluster cluster. This configuration uses four Cisco MDS 9250i, Brocade 7840 or 7810 FCIP switches, two at each site, to connect to the nodes through both FC initiators and FC-VI connections (see Figure 7). These switches are also used 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 7) MetroCluster FC IP configuration in ONTAP 9.x.



For updates to the maximum supported distances for FCIP MetroCluster configurations, see the [Interoperability Matrix Tool for ONTAP 9.5 and older](#). For ONTAP 9.6 and newer refer to the [Hardware Universe](#).

Table 1 describes the individual components in more detail.

Table 1) Required hardware components.

Component	Description
Two ONTAP clusters: <ul style="list-style-type: none"> <li>Four-node: four controllers</li> <li>Two-node: two controllers</li> </ul>	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 16Gb FC-VI card (two ports, with one connection to each local switch) and four FC initiators (8Gb or 16Gb, with two connections to each local switch). FAS and FlexArray controllers are supported.
Four FC switches (supported Brocade or Cisco models): <ul style="list-style-type: none"> <li>Not required for two-node, direct-attached or bridge-attached configurations</li> </ul>	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 FC-to-SAS bridges (ATTO 6500N/7500N/7600N 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 because only SAS shelves are supported, they bridge the protocol from SAS to FC. The FibreBridge is used only to attach NetApp disk shelves; storage arrays connect directly to the switch.
Recommended minimum SAS disk shelves per site (or equivalent storage array disks [array LUNs]): <ul style="list-style-type: none"> <li>Four-node: four disk shelves</li> <li>Two-node: two disk shelves</li> </ul>	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

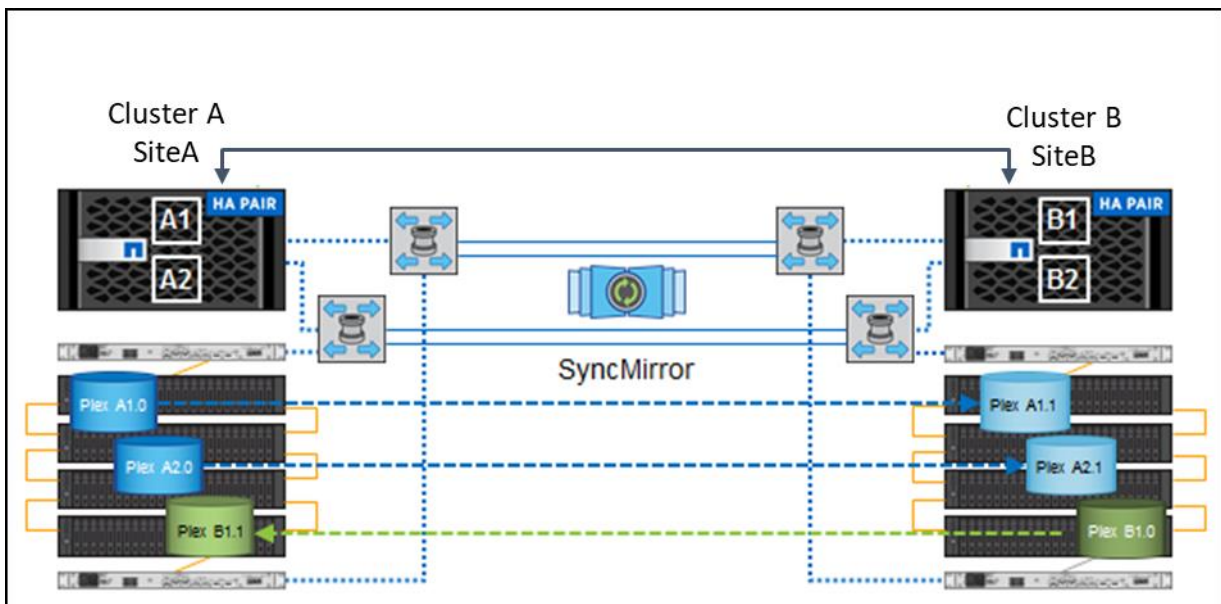
Component	Description
	configuration) at each site is supported, but NetApp does not recommend it. See the <a href="#">Interoperability Matrix Tool and Hardware Universe</a> for supported storage, number of shelves supported in a stack, and storage type mixing rules. All storage in the MetroCluster system 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. For more information about how aggregates are assigned to pools and across the shelves, see the section "Initial MetroCluster setup."

In a four-node MetroCluster configuration, there are a total of eight pools: a local pool (pool10) and a remote pool (pool11) for each of the four nodes, as shown in Figure 8. Cluster A local pools and cluster B remote pools are at site A. Cluster B local pools and cluster A remote pools are 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 during original setup and automatic ownership of any failed disk replacements. If shelves are not dedicated to pools, you must manually assign disk ownership during initial installation and for any subsequent failed disk replacements. NetApp recommends that you provide each shelf in the MetroCluster configuration (across both sites) with a unique shelf ID. Table 2 shows the shelf assignments that NetApp recommends.

**Table 2) 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. Before the clusters are created, you should verify this assignment and adjust it for the desired node-to-disk layout in maintenance mode so that the correct DR partner is chosen for each node. For more information, see section 0, Summary of installation and setup procedure.

Disk ownership is updated temporarily during an HA failover or DR switchover. ONTAP software must track which controller owns a particular disk and must 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 it identifies a disk in an aggregate that has been switched over from the partner cluster. Table 3 shows how the fields change during the different events.

**Table 3) Disk ownership changes.**

Field	Value during:			
	Normal Operation (All Nodes Up)	Local HA Failover (Four-Node Configuration)	MetroCluster Switchover	HA Failover After Switchover
<code>owner</code>	Name of the node that has access to the disk	Name of the HA partner that has temporary access to the disk	Name of the DR partner that has temporary access to the disk	Name of the DR partner's HA partner that has temporary access to the disk while in switchover
<code>home</code>	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
<code>dr-home</code>	Unassigned	Unassigned	Name of the original owning node	Name of the original owning node

Table 4 shows the ownership changes of a disk in the A1 remote pool, `pool1`. The disk is physically on site B but is owned in normal operation by node A1.

**Table 4) Example of disk ownership changes.**

MetroCluster state	Value of ownership fields			Notes
	<code>owner</code>	<code>home</code>	<code>dr-home</code>	
Normal operation: all nodes up	A1	A1	Unassigned	–
Local HA failover: A1<>2	A2	A1	–	<b>Note:</b> A2 takes over A1, so it has temporary ownership of A1's disks. After giveback to A1, disk ownership returns to normal operation.
Site switchover: A1/A2<>B1/B2	B1	B1	A1	<b>Note:</b> The original owning node name is saved in <code>dr-home</code> . B1 now owns A1's resources.  After switchback to site A, disk ownership returns to normal operation.

MetroCluster state	Value of ownership fields			Notes
	owner	home	dr-home	
Site switchover followed by HA takeover: A1/A2<>B1/B2 B1<>B2	B2	B1	A1	<p><b>Note:</b> As the sole surviving node, B2 now owns A1's resources.</p> <p>Recovery from this scenario is a two-step process:</p> <ol style="list-style-type: none"> <li>1. B2 gives back to B1 and restores <code>owner</code> to B1 (same state as the site switchover line).</li> <li>2. Site B switches back to site A. Ownership returns to the normal operation state with <code>dr-home</code> once again unassigned.</li> </ol>

## MetroCluster replication

Two types of replication are possible with MetroCluster: aggregate replication with SyncMirror and configuration replication with the configuration replication service. This section explains and demonstrates how aggregates, disks, and configurations are handled with MetroCluster. The following command shows the plex number spanning both local and remote nodes, aggregate assignment, and disk allocation.

```
tme-mcc-A::> storage aggr show -fields diskcount, plexe
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
```

## Plex read behavior

By default, all reads are from the local plex. You can set a RAID option so that read operations alternate between the local and the remote plexes. Some workloads might experience a performance increase when reading from both plexes, particularly if the sites are only a short distance apart. You can change the RAID option 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 this option have the same setting on both nodes of an HA pair. There might be workload environments in which different read behavior on each node gives optimal performance. If the setting is different between the nodes, the settings are not propagated to the other node if an HA failover occurs.

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: the HA partner is B2, and the DR partner is A1.

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

```
tme-mcc-B::> metrocluster interconnect show
```

Node	Partner Name	Partner Type	Mirror Admin Status	Mirror Oper 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: the HA partner is A2, and the DR partner is B1.
- Node A2: the HA partner is A1, and the DR partner is B2.

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

## Configuration replication service

The NetApp ONTAP configuration replication service (CRS) protects the MetroCluster configuration by synchronously replicating the local node configuration to the DR partner in the partner cluster. This replication is carried out over the cluster peering network. The information that is replicated includes the cluster configuration and the storage virtual machine (SVM, called “Vserver” in the CLI) configuration.

To verify the metadata volume (MDV), run the following command:

```
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	-	-	-

## Verify configuration replication services

Created objects are automatically propagated to the other cluster over the cluster peering network by the CRS. User-created job schedules are automatically replicated between clusters in a MetroCluster configuration. If you create, modify, or delete a job schedule on a cluster, the same schedule is automatically created on the partner cluster by the CRS.



System-created schedules are not replicated, and you must manually perform the same operation on the partner cluster so that job schedules on both clusters are identical.

```
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.

## 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 details about currently supported hardware components, consult the [Interoperability Matrix Tool for ONTAP 9.5 and older](#). For ONTAP 9.6 and newer the information is located in the [Hardware Universe](#). On the software side, MetroCluster is completely integrated into the NetApp ONTAP software. 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.

The [Fabric-Attached 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 that are contained in this guide.

## Hardware and software requirements

### Requirements for NetApp ONTAP

The following information applies to four-node MetroCluster configurations:

- Four nodes are required, with two nodes at each site. The four nodes are known collectively as a DR group. All four nodes must be the same FAS or AFF model (for example, four FAS9000 systems or four AFF A300 systems). FAS and FlexArray 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 ONTAP license. Protocols and other features such as NetApp SnapMirror® technology require licenses if they are used in the cluster. Licenses must be symmetrical across both sites. For example, if you use SMB, it 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.
- NetApp ONTAP FlexGroup volumes are supported in a MetroCluster configuration starting with ONTAP 9.6.
- 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. ONTAP systems can attach to NetApp E-Series storage arrays with full disk encryption (FDE) drives. The root aggregate can reside on encrypted drives.
- NetApp Volume Encryption (NVE) and NetApp Aggregate Encryption (NAE) are both supported.

The following information applies to two-node MetroCluster configurations:

- Two nodes are required, with one node at each site. The two nodes are configured as two separate clusters and are known as a DR group. Both nodes must be the same FAS, AFF or FlexArray model (for example, two FAS8200 systems or two FAS9000 systems). FAS and FlexArray 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 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, if you use SMB, it 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.
- Advanced Disk Partitioning is not supported.
- NSE drives are not supported in a MetroCluster configuration. ONTAP systems can attach to NetApp E-Series storage arrays with FDE drives. The root aggregate can reside on encrypted drives.

## Requirements for MetroCluster configurations with array LUNs

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

The platforms and storage array must be listed as supported for MetroCluster configurations.

**Note:** The Interoperability Matrix Tool contains details for MetroCluster configurations that use array LUNs. It includes information about supported storage arrays, switches, NetApp controllers, and ONTAP software 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 FlexArray Virtualization for Fabric MetroCluster as the storage solution. For details about currently supported hardware components, consult the [Interoperability Matrix Tool](#) for ONTAP 9.5 and older. For ONTAP 9.6 and newer the information is located in the [Hardware Universe](#).

- All 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 you mix 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 symmetrical, 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, solid-state drive (SSD), or SAS) that you use 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 requirements are for setting up a MetroCluster configuration with FC switches:

- The switches and switch firmware must be identified 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-attached configurations and two-node direct-attached configurations do not require FC switches. An eight-node fabric configuration uses two switches at each location in a manner that is like four-node deployment. Nodes in each cluster connect to the same switches at each location.
- 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 switches that are used for MetroCluster.
- To provide redundancy if device and path failures occur, each ONTAP system must be connected to storage with redundant components.
- ONTAP supports one to eight ISLs per fabric depending on the switches that you use. 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 setting that NetApp recommends, regardless of the number of ISLs per fabric.

**Note:** For more information about basic switch configuration, ISL settings, and FC-VI configurations, see the [Fabric-Attached 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 FCIP switches and an IP ISL. For FCIP MetroCluster deployments, NetApp supports Cisco MDS 9250i, Brocade 7840, and Brocade 7810 FCIP switches for IP ISL connectivity. The cabling configuration remains the same. The FCIP configuration requires a dedicated IP network for the IP ISL and can be directly attached. Intermediate switches are not supported.

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

- The switch firmware must be identified 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-attached and two-node direct-attached configurations do not require FCIP switches.
- All switches in the configuration must be Cisco MDS 9250i, Brocade 7840, or Brocade 7810 switches, 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 switches that are used for MetroCluster.
- To provide redundancy if device and path failures occur, each ONTAP system must be connected to storage with redundant components.
- To verify ISL fabric maximums and trunking requirements.

**Note:** For more information about basic switch configuration, ISL settings, and FC-VI configurations, see the [Fabric-Attached MetroCluster Installation and Configuration Guide](#).

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

- The FCIP switches 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.
- The FCIP ISLs between the MetroCluster clusters cannot be shared.

## Requirements for FibreBridges

All four-node configurations with NetApp FAS and AFF shelves require FibreBridges. The two-node fabric and two-node bridge-attached 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. A FibreBridge is not used for a MetroCluster configuration that uses only array LUNs.

Starting with ONTAP 9.6, the new ATTO 7600N bridge is available and replaces the 7500N bridge. For expansion of existing configurations using the ATTO 7500N, the 7600N offers the same functionality.

The key advantages of the ATTO 7600N include the following:

- Dual power supplies
- Increased throughput performance
- Two 32Gb FC ports; both can be enabled
- Four 12Gb SAS ports; all four can be enabled
- In-band monitoring capability starting in ONTAP 9.5

The ATTO FibreBridge 7500N started shipping with ONTAP 8.3.2. The ATTO 7500N provided improved performance and lower latency compared to the ATTO FibreBridge 6500N. In ONTAP 9.5 and later, the ATTO 7500N also supports in-band monitoring. This allows for increased security compared to using SNMP and Ethernet for monitoring.

In version earlier than ONTAP 8.3.2, the ATTO FibreBridge 6500N is used for connecting storage.

A configuration requires 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 an FC port to a switch and through a SAS port to the SAS disk stack. Up to 10 shelves per stack are supported when only HDDs are used. Mixing shelf-module types (for example, IOM3 and IOM6) in the stack is supported as described the [storage subsystem requirements](#). See Table 6 for the maximum number of ATTO 7500N disk shelves.

**Table 5) ATTO 7600N shelf count.**

Disk configuration	Disk type	Total shelves per bridge pair			
All SSDs	SSD	4			
All HDDs	HDD	10			
Mixed SSDs and HDDs	SSD	1	2	3	4
	HDD	9	8	7	6

**Table 6) ATTO 7500N shelf count.**

Disk configuration	Disk type	Total shelves per bridge pair
All SSDs	SSD	4
All HDDs	HDD	10

Disk configuration	Disk type	Total shelves per bridge pair			
Mixed SSDs and HDDs	SSD	1	2	3	4
	HDD	9	8	7	6

If SSDs are in the FibreBridge stack, the maximum supported stack depths are shown in previous tables. 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 such a configuration, 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 only contain SSDs, with no more than two SSD shelves in the stack. Optimal performance is achieved with one SSD shelf per stack.

**Note:** For details on current support for platforms, disk shelves, and disk devices, see the [Hardware Universe](#) for ONTAP 9.6 and newer. For ONTAP 9.5 and older, see the [Interoperability Matrix Tool](#).

## Requirements for zoning

The following requirements are 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 worldwide port name (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 requirements are for the SyncMirror storage:

- The disk configuration must be identical between the two sites for mirrored aggregates only. This requirement includes NetApp Flash Pool configurations. If Flash Pool is required for a node's workload, the same capacity for Flash Pool intelligent caching must exist in the mirrored plex for each Flash Pool aggregate.
- RAID 4 and NetApp RAID DP<sup>®</sup> 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 pool-shelf isolation is preserved. Although the use of fewer than four shelves per site is supported, it makes disk assignment more complex and also significantly limits the amount of usable storage that is available.

Figure 9 shows a sample layout of pool-to-shelf assignment with shelf IDs assigned as recommended in Table 2) 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 automatic assignment is disabled, and the scope of impact of a shelf failure is also increased.

Figure 9) MetroCluster pool-to-shelf assignment.



If SSDs are included in FibreBridge-attached stacks, see the [Interoperability Matrix Tool](#) and [Hardware Universe](#) 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.

Furthermore, for optimal performance, NetApp recommends that you place SSD shelves either in their own dedicated stack (no HDD shelves) or at the top or at 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, you can add storage one shelf at a time to each cluster; there is no requirement to add multiple shelves. However, to preserve mirroring symmetry, you must add the same storage at both sites.

### 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 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.”

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 must 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 you attach the cabling this way, even if you have deviated from the recommended port assignments.

**Note:** All nodes must use either on-board FC-VI ports or adapters with FC-VI ports. Mixing on-board and FC-VI adapter ports is not supported.

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

## Summary of installation and setup procedure

The [Fabric-Attached MetroCluster Installation and Configuration Guide](#) provides worksheets and instructions to help you configure your MetroCluster components.

For reference, see the following 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, and then install and connect interdevice and intersite cabling (ISLs and cluster peering connections). This step includes the ONTAP nodes (FAS or FlexArray controllers, the cluster management switch, or the 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 switches that are supported by MetroCluster. If manual configuration is required, follow the steps in the [Fabric-Attached MetroCluster Installation and Configuration Guide](#).
3. Do not connect the ISL links until the product documentation directs you to do so. If the links are already up, local HA cluster setup does not work.
4. In maintenance mode, complete the following steps:
  - a. Verify disk ownership. Equipment that you receive from manufacturing should have disk ownership preassigned equally across the nodes. However, if the assignment does not match your requirements, manually reassign the disks before you create the clusters. See the [Fabric-Attached 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 that are 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. Also verify that the two nodes with system IDs have matching disk configurations. 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 [Fabric-Attached 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, complete the following steps:
  - 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 require a dedicated aggregate.

- e. Install any additional ONTAP feature licenses that are required. To achieve correct client and host access after switchover, licenses must be symmetrical on both clusters. 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 Active IQ Unified Manager (for more details, see the section “Active IQ Unified Manager and health monitors”).
  - i. Verify the MetroCluster configuration with the `metrocluster check run` command. Follow the additional verification steps in the [Fabric-Attached MetroCluster Installation and Configuration Guide](#).
6. Install Active IQ 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 that are 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 to check site switchover, healing, and giveback, follow the steps in the [Fabric-Attached MetroCluster Installation and Configuration Guide](#).

## Post setup configuration and administration

After MetroCluster configuration is complete, ongoing administration is almost identical to administration of an 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. These objects are automatically replicated to the other cluster over the cluster peering network. You can configure SVMs by using the CLI, Active IQ System Manager, or NetApp OnCommand Workflow Automation.

## Aggregate resynchronization and Snapshot copies

Aggregate NetApp Snapshot™ copies are created at regular intervals for SyncMirror operation. The default interval is 60 minutes. NetApp recommends that you reduce this interval to 15 minutes. Also, if you use Flash Pool, you should reduce this interval to 5 minutes. See the section “Flash Pool,” for the command that you need 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 when you use OnCommand System Manager.

## Viewing aggregates

In normal operation, each cluster sees only its own aggregates, as is 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,



aggr0_tme_A2	1.38TB	741.9GB	48%	online	1	tme-mcc-A2	normal	raid_dp, mirrored, normal
aggr1_tme_A1	2.07TB	2.06TB	1%	online	4	tme-mcc-A1	normal	raid_dp, mirrored, normal
aggr1_tme_A2	2.07TB	2.06TB	0%	online	1	tme-mcc-A2	normal	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 ONTAP 9.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, `svml_mccA`, and an SVM on cluster B, `svml_mccB`.

The cluster A output shows `svml_mccA`, with the type `sync-source`. The entry SVM `svml_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
```

Vserver	Type	Subtype	Admin State	Operational State	Root Volume	Aggregate
svml_mccA	data	sync-source	running	running	svml_mccA_root	aggr1_tme_A1
svml_mccB-mc	data	sync-destination	running	stopped	svml_mccB_root	aggr1_tme_B1

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

```
tme-mcc-B::> vserver show -type data
```

Vserver	Type	Subtype	Admin State	Operational State	Root Volume	Aggregate
svml_mccA-mc	data	sync-destination	running	stopped	svml_mccA_root	aggr1_tme_A1
svml_mccB	data	sync-source	running	running	svml_mccB_root	aggr1_tme_B1

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 only the local cluster. Note that on cluster A, only SVM `svml_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%
svml_mccA	svml_mccA_lun1_vol	aggr1_tme_A1	online	RW	1.24GB	248.2MB	80%
svml_mccA	svml_mccA_root	aggr1_tme_A1	online	RW	1GB	972.5MB	5%
svml_mccA	vol1	aggr1_tme_A1	online	RW	2GB	1.85GB	7%
svml_mccB-mc	svml_mccB_root	aggr1_tme_B1	-	RW	-	-	-
svml_mccB-mc	vol1	aggr1_tme_B2	-	RW	-	-	-
svml_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%
svml_mccA-mc	svml_mccA_lun1_vol	aggr1_tme_A1	-	RW	-	-	-
svml_mccA-mc	svml_mccA_root	aggr1_tme_A1	-	RW	-	-	-
svml_mccA-mc	vol1	aggr1_tme_A1	-	RW	-	-	-
svml_mccB	svml_mccB_root	aggr1_tme_B1	online	RW	1GB	972.5MB	5%
svml_mccB	vol1	aggr1_tme_B2	online	RW	2GB	1.85GB	7%
svml_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
svml_mccA	/vol/svml_mccA_lun1_vol/svml_mccA_lun1				

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 that is 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 Viewing LIFs in the section "Performing planned (negotiated) switchover." 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 be assigned correctly on the partner cluster, and switchover is not possible.

```
tme-mcc-A::> network interface show -vserver svm*
Vserver      Logical      Status      Network      Current      Current      Is
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 that are replicated from cluster A are in an operational state of `down`.

```
tme-mcc-B::> network interface show -vserver svm*
Vserver      Logical      Status      Network      Current      Current      Is
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
```

# Resiliency for planned and unplanned events

NetApp MetroCluster enhances the high availability (HA) and the nondisruptive operations of NetApp hardware and ONTAP configurations, providing site-wide 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 if failure occurs 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. MetroCluster uses 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 configuration means that operations continue seamlessly if a single component fails, and the systems return automatically to redundant operation when the failed component is repaired or replaced.

HA takeover and giveback functionality is inherent in all NetApp 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 a 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 an ONTAP upgrade on a node) or unplanned (such as for a node hardware or software failure). During 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 [ONTAP High-Availability Configuration 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.

## 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.
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.

Unplanned operation	Recovery method
	<p>The aggregate is placed in degraded mode because only one plex is active.</p> <p>If the failure is due to power loss on the shelf, when power is restored, the affected aggregate automatically resynchronizes itself to catch up with any changes.</p> <p>If disks must be replaced, the administrator deletes the failed plex (<code>storage aggregate plex delete</code> command) and then remirrors the affected aggregate (<code>storage aggregate mirror</code> command). This action begins the automatic resynchronization process.</p> <p>After resynchronization, the aggregate returns automatically to normal mirrored mode.</p>
Switch fails	<p>Data continues to be served on the surviving path; all plexes remain available. Because there are two fabrics, one fabric can completely fail and the operation will be preserved.</p>
FibreBridge fails	<p>All traffic continues to the affected stack by the surviving bridge; all plexes remain available.</p>
Single node fails	<p><b>Four-node configuration.</b> 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.</p> <p><b>Two-node configuration.</b> 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.</p>
ISL loss between sites	<p>If one or more ISLs fail, I/O continues through the remaining links. If <b>all</b> 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 is extremely low, given that two independent fabric connections are required (which can use independent network providers) that and up to four separate ISLs per fabric can be configured. After a minimum of one ISL is restored, the plexes resynchronize automatically.</p> <p>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.</p> <p>In this case, manual intervention is required for recovery after the switchover; see the section “Protecting volumes after forced switchover” for more details. If it is likely that no ISLs are 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 occurs 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.</p>
Peered cluster link (CRS) fails	<p>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.</p>

Unplanned operation	Recovery method
	<p>These changes are kept in the local CRS metadata volumes and are automatically propagated to the other cluster after the restoration of the peered cluster link.</p> <p>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.</p>
<p>All nodes at a site fail, or the complete site is destroyed</p>	<p>The administrator performs a forced switchover to resume services of the failed nodes on the surviving site. Forced switchover is a manual operation; see the section “MetroCluster Tiebreaker software” for more information. After the failed nodes or sites are restored, a switchback operation is performed to restore steady-state operation of the configuration.</p> <p>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.</p>

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
<p>Upgrading ONTAP software</p>	<p>For four-node configurations, this operation is performed by using nondisruptive upgrade (NDU: failover and giveback within the HA pair) as for any ONTAP upgrade, through simpler operations with lower possibility of error in ONTAP 9.8. For a minor update, 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 ONTAP versions. However, for steady-state operation, NetApp recommends that you finish the upgrade process as soon as possible. A planned switchover should be delayed until all nodes are upgraded to the same ONTAP version.</p> <p>Major updates 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 ONTAP versions.</p>
<p>Upgrading controller hardware by using aggregate relocation (ARL)</p>	<p>You can use ARL to nondisruptively upgrade controller models in place: for example, to upgrade FAS8040 to FAS8200. 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 the following sequence:</p> <ol style="list-style-type: none"> <li>1. Disable NVRAM replication on all nodes.</li> <li>2. Upgrade the nodes in the first cluster.</li> <li>3. Upgrade the nodes in the second cluster.</li> <li>4. Enable NVRAM replication on all nodes.</li> </ol>

Planned operation	Nondisruptive process
	<p>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 synchronized. 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.</p> <p>The command to disable or to enable NVRAM mirroring is:</p> <pre>metrocluster interconnect modify -node &lt;nodename&gt; -partner-type DR -mirror-status OFFLINE ONLINE</pre> <p>The ARL process is lengthy and has many steps. For the detailed steps for this process, see the <a href="#">controller upgrade guide</a>. Follow the steps closely.</p>

## 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 when 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 give backs in progress.
- An 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 must 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 the section “AutoSupport.”
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 by using the command `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. By using the same rules as in the `metrocluster switchover -simulate` command, it checks the configuration for conditions that can be vetoed.
  - 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 the NetApp WAFL® file system, 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 resume.
  - 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 that you fence the storage only if necessary: for example, if power is cut to the data center or if 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 resynchronization 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](#). All NetApp 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 the section “Performing switchback.”

Following are the equivalent commands and checks from the section “Post setup configuration and administration,” but executed after switchover. These commands show how the object view changes, and they 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 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, `svml_mccA-mc`, has changed to a running state.

```
tme-mcc-B::> vservers show -type data
-----
Vserver      Type      Subtype      Admin      Operational  Root
State      State      Volume      Aggregate
-----
svml_mccA-mc
  data      sync-destination      running      svml_mccA_root      aggr1_tme_A1
svml_mccB   data      sync-source      running      svml_mccB_root      aggr1_tme_B1
```

## 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%
-----
svml_mccA-mc
  svml_mccA_lun1_vol
    aggr1_tme_A1 online      RW      1.24GB      248.2MB      80%
svml_mccA-mc
  svml_mccA_root
    aggr1_tme_A1 online      RW      1GB      972.5MB      5%
svml_mccA-mc
  vol1      aggr1_tme_A1 online      RW      2GB      1.85GB      7%
svml_mccB   svml_mccB_root
  aggr1_tme_B1 online      RW      1GB      972.5MB      5%
svml_mccB   vol1      aggr1_tme_B2 online      RW      2GB      1.85GB      7%
svml_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
-----
svml_mccA-mc
  /vol/svml_mccA_lun1_vol/svml_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 the LIF IP address, the 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*
```

Vserver	Logical Interface	Status Admin/Oper	Network Address/Mask	Current Node	Current Port	Is Home
-----						
svml_mccA-mc						
	svml_mccA_iscsi_lifA1_1	up/up	10.228.22.68/24	tme-mcc-B1	e0a	true
	svml_mccA_iscsi_lifA1_2	up/up	10.228.22.69/24	tme-mcc-B1	e0b	true
	svml_mccA_iscsi_lifA2_1	up/up	10.228.22.97/24	tme-mcc-B2	e0a	true
	svml_mccA_iscsi_lifA2_2	up/up	10.228.22.98/24	tme-mcc-B2	e0b	true
	svml_mccA_nas_A1_1	up/up	10.228.22.62/24	tme-mcc-B1	e0b	true
svml_mccB						
	svml_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 they are 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 "SnapMirror Asynchronous data replication," for important considerations.

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

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 the switchback is performed.

It is possible to extend an aggregate from either cluster while in switched-over 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 the aggregate 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.

## 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 shut-down site was offline during the switchover, resynchronization is automatically initiated to propagate any I/O that occurred during switchover. Wait for the resynchronization to be complete; the next phase of switchback cannot execute until the data aggregates are resynchronized. To monitor the status, use the following command:

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

Aggregate	Resyncing Plex	Complete 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 resynchronization is complete, verify the aggregate status again. All aggregates should show a RAID status of 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

- 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.
- 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
```

- 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
```

- 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
```

- 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
```

- 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. By using the same rules as in the `metrocluster switchback -simulate` command, it checks the configuration for conditions that cause switchback to be vetoed.
- 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 for maintenance.

## Performing forced switchover

A forced switchover differs from a negotiated switchover in that the nodes and the 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 “Fencing Off the Disaster Site” section in the [MetroCluster Management and Disaster Recovery Guide](#)).

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.
2. Manually reestablish the SVM peering (if the destination of the relationship is in the MetroCluster configuration).

## 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 to applications after taking appropriate action against 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 the `volume create` or the `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 the NVFAIL state, the root recovery procedure must be used.

NetApp recommends that you make 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](#).

## 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.



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 to recover from any subsequent failures as quickly as possible. Do not perform a major 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 issues that you must consider when you perform the eventual switchback. For example, you must upgrade the down nodes at the disaster site to the same ONTAP release before you perform the switchback.

Recovery steps after a forced switchover or a disaster are documented in the [MetroCluster Management and Disaster Recovery Guide](#). If you must replace hardware, 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.”

## Interoperability

Several management tools are available to monitor the NetApp MetroCluster configuration, including NetApp ONTAP System Manager, Active IQ Unified Manager, AutoSupport, MetroCluster Tiebreaker software, and Config Advisor. This section discusses their use with MetroCluster.

To highlight their compatibility and interactions with MetroCluster configurations, several standard NetApp ONTAP features, including SnapMirror, SnapVault, QoS, 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, ONTAP System Manager, the CLI, or ONTAP REST APIs.

### ONTAP System Manager

ONTAP System Manager is a web-based, on-box application in the ONTAP 9 software 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 to configure the SVMs and their associated objects. Of course, the CLI and ONTAP REST APIs 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. Figure 10 shows SVM `svm1_mccA` as the `sync_source` SVM on cluster A and shows its state as running and its configuration state as unlocked. SVM `svm1_mccB-mc` is the `sync_destination` copy of SVM `svm1_mccB`, and Figure 12 shows the state as stopped and the configuration state as locked on cluster B.

**Figure 10) Cluster A: The `sync_source` SVM is unlocked, and the `sync_destination` SVM is locked.**

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

In Figure 13, 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 is in the configuration state of unlocked (subtype `sync_source`).

**Figure 11) Cluster B: The `sync_source` SVM is unlocked and the `sync_destination` SVM is locked.**

The screenshot shows the System Manager interface for Cluster B. On the left, a tree view shows 'Storage Virtual Machines' expanded to 'tme-mcc-3250-B', which contains 'svm1\_mccA-mc' and 'svm1\_mccB'. The main pane displays a table with the following data:

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 to update export policies. Figure 12 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 12) SVMs after switchover: All SVMs are unlocked.**

The screenshot shows the System Manager interface for Cluster A (tme-mcc-3250-A). On the left, a tree view shows 'Storage Virtual Machines' expanded to 'tme-mcc-3250-A', which contains 'svm1\_mccA' and 'svm1\_mccB-mc'. The main pane displays a table with the following data:

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

## Active IQ Unified Manager and health monitors

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

- SNMP or in-band monitoring of the 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 Unified Manager for monitoring every MetroCluster installation. It can be used in conjunction with enterprise monitoring tools.

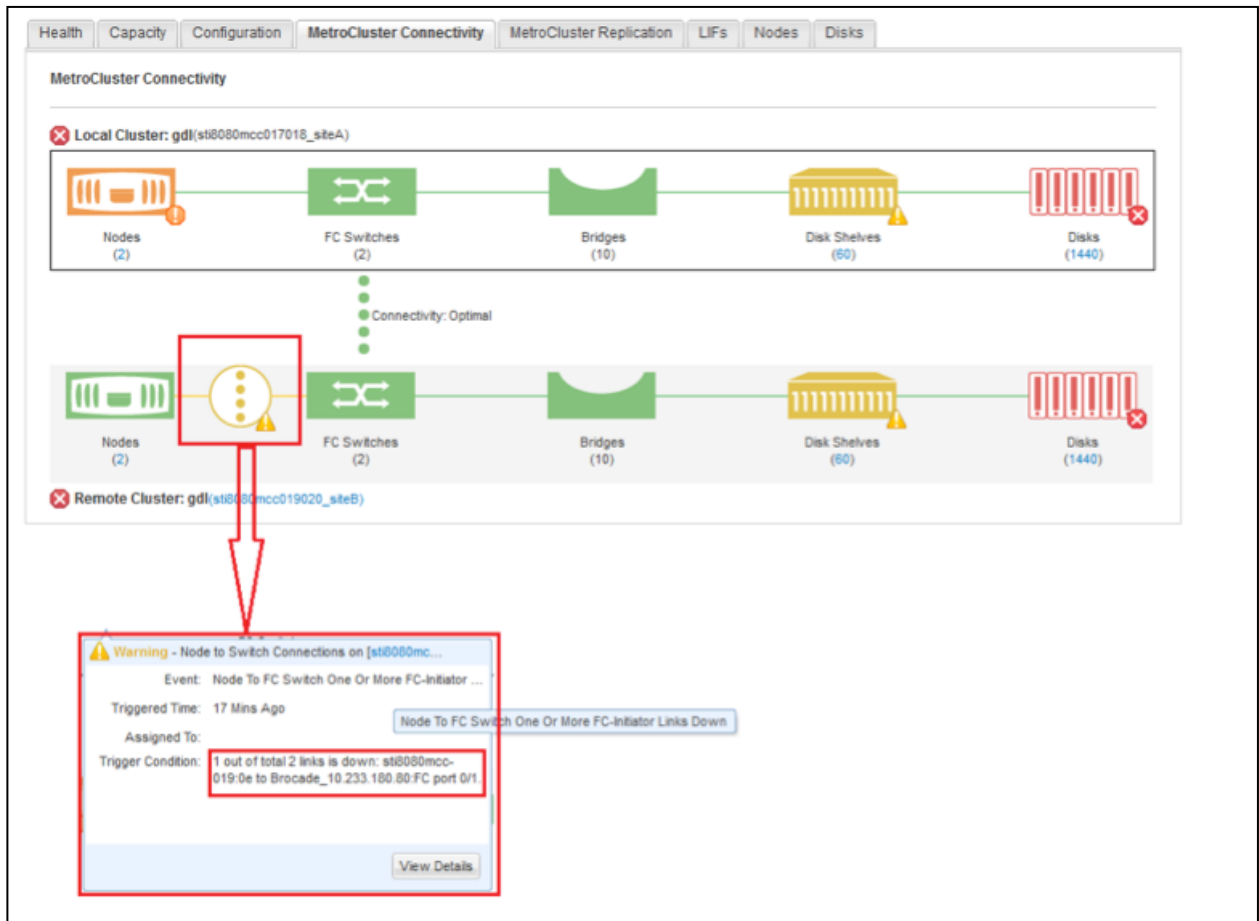
Active IQ Unified Manager uses the cluster topology information to build an end-to-end graphical representation of the MetroCluster configuration. This representation is built automatically when Unified Manager discovers the clusters that form a MetroCluster system. Unified Manager then periodically polls the health monitors for detected faults and translates them to Unified Manager alerts. An alert that is raised by Unified Manager includes information about the likely cause and suggests remediation actions. It can be assigned to system administrators for appropriate handling. In Active IQ Unified Manager 7.2 with ONTAP 9, 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.

Unified Manager uses the information that is collected by the MetroCluster health monitors to gather information about the configuration and to collect events that are related to the components. The health monitors use SNMP to monitor the switches and bridges. For current MetroCluster versions in-band monitoring is available for bridges (7500N or later models) for configurations where SNMP is not desired.

Active IQ 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 Unified Manager interface.

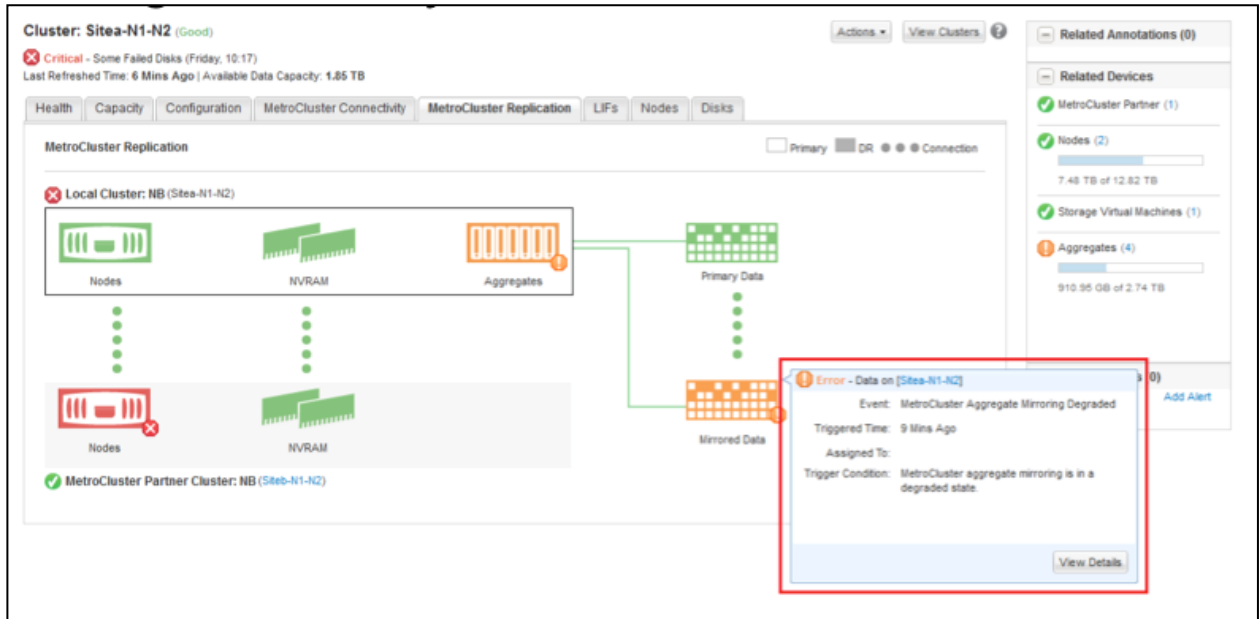
With connectivity monitoring, Unified Manager monitors and reviews the health of the hardware in the MetroCluster configuration and raises alerts for issues that are related to the physical connectivity between devices. Alerts show the likely cause and the impact of the issue and suggest remediation.

**Figure 13) Active IQ Unified Manager device and link monitoring.**



Replication monitoring (Figure 16) displays the health of the synchronous relationships in MetroCluster: SyncMirror for aggregates and the NVRAM replication.

Figure 14) Active IQ Unified Manager replication monitoring.



## AutoSupport

NetApp AutoSupport messages that are specific to MetroCluster configurations are also sent automatically. 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 enables NetApp Support to respond quickly and proactively.

If you perform MetroCluster operations solely for testing purposes or for 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 Knowledgebase article [1015155](#) (logging in to the NetApp Support site is required) has more information about using AutoSupport for this purpose.

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

## 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 could 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."

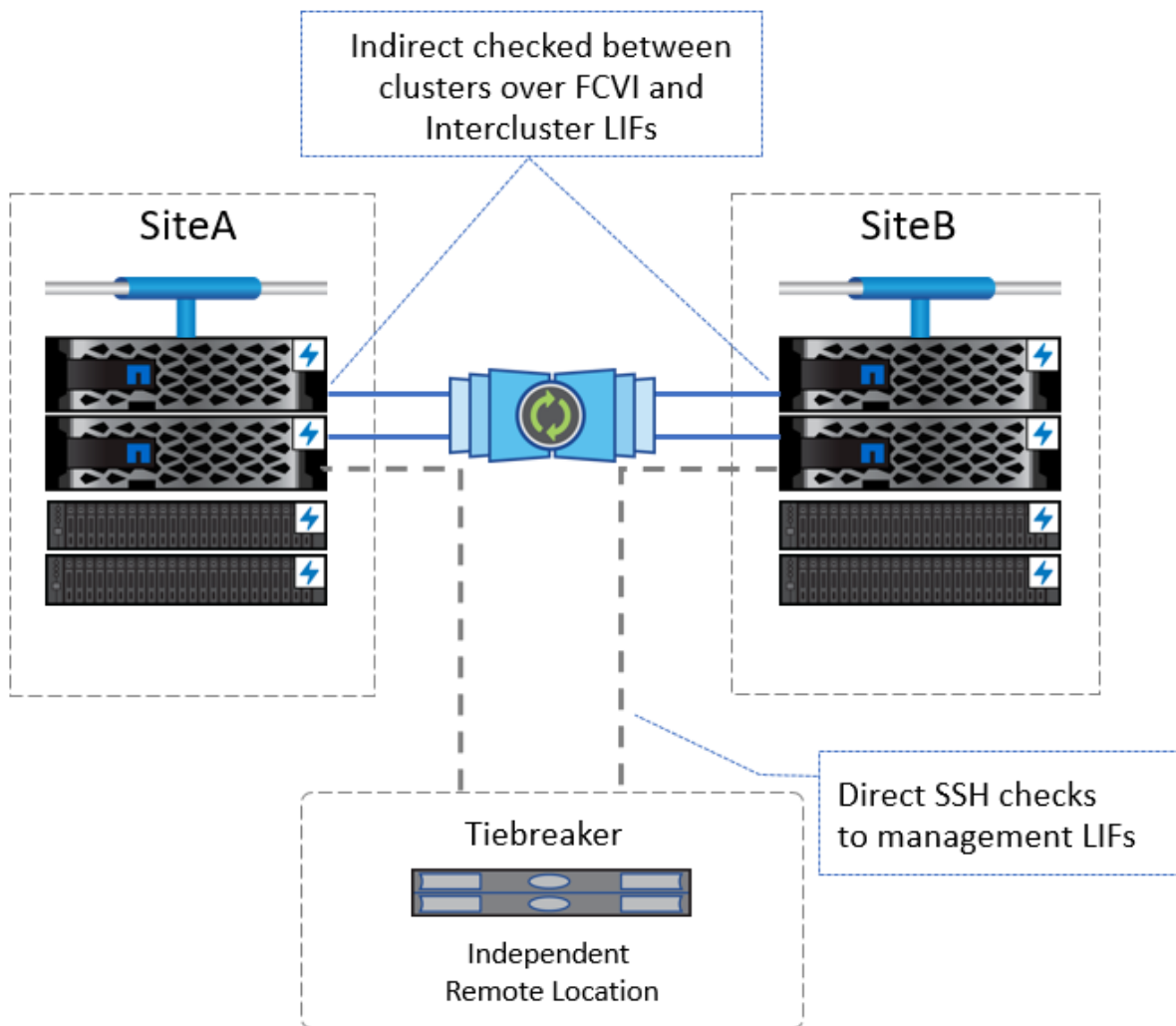
The need for switchover, then, can be either human determined, or application led. NetApp provides a fully supported capability, MetroCluster Tiebreaker software, which is installed at a third site with independent connections to each of the two clusters. The purpose of the Tiebreaker software is to monitor and to detect both individual site failures and intersite link failures. MetroCluster Tiebreaker software can raise an SNMP trap if a site disaster occurs. 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. Tiebreaker software can be configured to automatically issue the command for switchover if a disaster occurs.

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 link and cluster condition state. The update indicates whether Tiebreaker detected an HA takeover event, a site failure, or failure of all intersite links.

A direct link is through Secure Shell (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 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. In this scenario, MetroCluster continues to operate. For more information, see Table 7.

**Figure 15) MetroCluster Tiebreaker software operation.**



MetroCluster Tiebreaker software is distributed through 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](#) and select MetroCluster Tiebreaker. The installation and configuration guide and the system prerequisites are also available at this link. No additional ONTAP license is required for the Tiebreaker software.

## Config Advisor

Config Advisor is a configuration validation and health-check tool for NetApp systems. Config Advisor 5.2 and later versions support MetroCluster configurations, with a pool of 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. Figure 16 shows an output example when Config Advisor is run against a MetroCluster system.

**Figure 16) Config Advisor sample output.**

Impact Level	Category	Rule Target	Risk / Description	Details
Pass	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
Pass	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
Pass	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
Pass	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
Pass	SwitchGeneric_SwUniqueName	tme-mcc-B-fab2	Verify that all switches have a unique name.	Switch names are unique within this DRGroup.
Pass	SwitchGeneric_SwUniqueName	tme-mcc-B-fab1	Verify that all switches have a unique name.	Switch names are unique within this DRGroup.
Pass	SwitchGeneric_SwUniqueName	tme-mcc-A-fab1	Verify that all switches have a unique name.	Switch names are unique within this DRGroup.
Pass	SwitchGeneric_SwUniqueName	tme-mcc-A-fab2	Verify that all switches have a unique name.	Switch names are unique within this DRGroup.
Pass	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
Pass	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
Pass	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
Pass	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
Fail	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
Fail	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

## Quality of service

QoS can be used in MetroCluster configurations to extend its typical use cases in an ONTAP cluster. QoS policies can be dynamically applied and modified as necessary. See the following examples for use of QoS in MetroCluster environments:

- 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 that is applied to the system sizing, the reduction in available resources could affect client and application workloads. To provide more resource availability to critical workloads, QoS policies can be configured to apply a ceiling (IOPS or throughput) to noncritical workloads. The policies can be disabled after switchover when normal operation has resumed.

For more information about the use of QoS, see the white paper [ONTAP Performance Management Power Guide](#). See also the section “Managing Workload Performance by Using Storage QoS” in the Data System Administration Guide for Cluster Administrators in the ONTAP product documentation.

## SnapMirror Asynchronous data replication

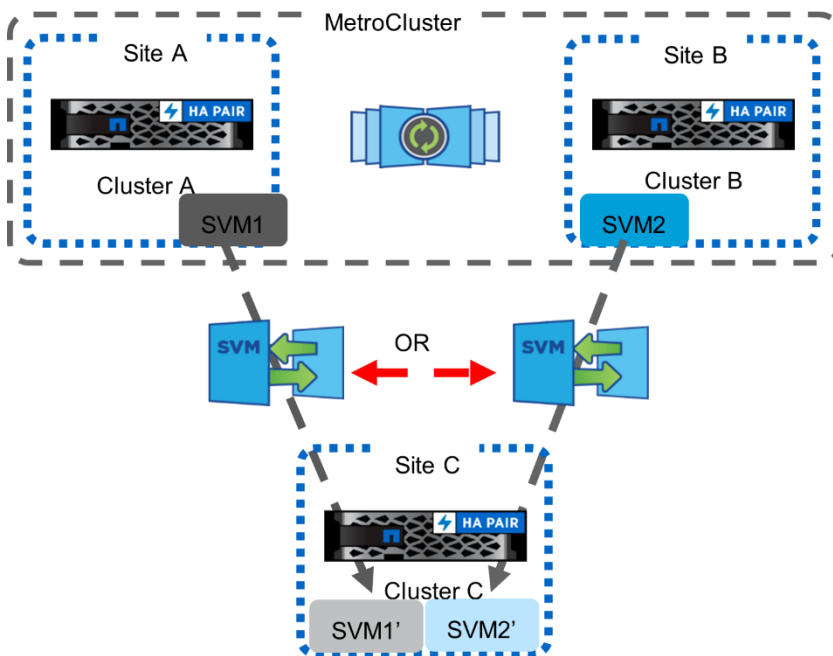
SnapMirror unified replication relationships can be created by using MetroCluster protected volumes as either the source or the destination. The data protection relationships can be within the MetroCluster environment (in the same cluster or in the other cluster) or to and from other ONTAP clusters without 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.
- You can run SnapMirror operations only from the cluster that runs 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® technology, SnapMirror, or any other operation that requires the volume to be brought online. The other cluster accesses the volumes only when a switchover is performed.
- For MetroCluster volumes as a SnapMirror 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 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 the `snapmirror create` command. A SnapMirror re-baseline is not required. An [OnCommand WFA workflow](#), “Re-Create SnapMirror and SnapVault Protection After MetroCluster Switchover and Switchback,” is available to automate the re-creation of the relationships.

## SVM DR

MetroCluster in ONTAP 9.5 adds support for SVM DR to asynchronously mirror an SVM to a third cluster for an additional level of protection. MetroCluster can only be the source of an SVM DR relationship.

Figure 17) SVM DR with MetroCluster



For more information about SVM DR see the [“Managing SnapMirror SVM replication”](#) section of the Data Protection and Disaster Recovery guide in the ONTAP 9 documentation.

## SnapLock

MetroCluster in ONTAP 9 supports NetApp SnapLock® software.

## Volume move

Data mobility by using volume move (NetApp DataMotion™ for Volumes) is one of the core ONTAP nondisruptive operations. Nondisruptive movement of 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 event management system (EMS) is logged, and the original source volume must be manually deleted.

It is not possible to perform 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, and if the operation is confirmed, the volume move operation 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 doing a volume move of an MDV is that the aggregate that contains the MDV 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.

## Volume rehost

MetroCluster does not support volume rehost.

## FlexGroup

NetApp FlexGroup volumes are supported with MetroCluster starting with ONTAP 9.6. A FlexGroup volume is a scale-out NAS container that provides high performance along with automatic load distribution and scalability. For more information refer to the [Scalability and Performance Using FlexGroup Volumes Power Guide](#) in the ONTAP documentation.



## FlexCache

Starting with ONTAP 9.7, NetApp FlexCache software is supported with MetroCluster FC.

- [FlexCache Volumes for Faster Data Access Power Guide](#)
- [TR 4743: FlexCache in NetApp ONTAP](#)

## Flash Pool

NetApp Flash Pool caching 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 system 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 resynchronization 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 MetroCluster FC configurations.

## NetApp AFF A-Series arrays

NetApp AFF systems help you meet your enterprise storage requirements with industry-leading high performance, superior flexibility, and best-in-class data management and cloud integration. Combined with the industry's first end-to-end NVMe technologies and NetApp ONTAP data management software, AFF systems accelerate, manage, and protect your critical data. With an AFF system, you can make an easy and risk-free transition to flash for your digital transformation.

Designed specifically for flash, the AFF A-Series systems deliver industry-leading performance, capacity density, scalability, security, and network connectivity in dense form factors. As the industry's first all-flash arrays to provide both 100 Gigabit Ethernet (100GbE) and 32Gb FC connectivity together, AFF A-Series systems are designed to support high performance workloads. Having been a leader in supporting high-capacity 15TB SSDs and multistream write SSDs, AFF is leading again as the first all-flash system to support 30TB SSDs. You can further reduce your storage footprint with the high density of 2PB SSD storage in a 2U drive shelf and move toward an optimally efficient data center.

MetroCluster extends the capabilities of AFF to deliver high performance and low latency for critical business applications that need continuously available storage. AFF is transparent to MetroCluster and is supported on all two-node, four-node, and eight-node MetroCluster configurations.

## Where to find additional information

To learn more about the information that is described in this document, see the following documents and websites:

- TR-4689: NetApp MetroCluster IP  
<http://www.netapp.com/us/media/tr-4689.pdf>
- TR-3978: 64-Bit Aggregates: Overview and Best Practices  
<http://www.netapp.com/us/media/tr-3978.pdf>

- MetroCluster FC Technical FAQ (NetApp Field Portal; login required)  
<https://fieldportal.netapp.com/content/617080>
- MetroCluster IP and FC ISL Sizing Spreadsheet (NetApp Field Portal; login required)  
<https://fieldportal.netapp.com/content/699509>
- NetApp Interoperability Matrix Tool  
<http://mysupport.netapp.com/matrix/>
- NetApp MetroCluster Product Documentation  
<http://docs.netapp.com/ontap-9/topic/com.netapp.nav.mc/home.html>
- NetApp MetroCluster Resources page  
<http://mysupport.netapp.com/metrocluster/resources>
- NetApp ONTAP Resources page  
<http://mysupport.netapp.com/ontap/resources>
- NetApp Product Documentation  
<https://docs.netapp.com>

## Contact Us

To let us know how we can improve this technical report, contact us at [doccomments@netapp.com](mailto:doccomments@netapp.com). Include TECHNICAL REPORT 4375 in the subject line.

## Version History

Version	Date	Document version history
Version 1.9	October 2021	Update to Figure 17
Version 1.8	November 2020	ONTAP 9.8 updates Added Brocade switches Updated ONTAP System Manager and Active IQ Unified Manager nomenclature
Version 1.7	November 2019	ONTAP 9.7 updates
Version 1.6	May 2019	ONTAP 9.6 updates
Version 1.5	November 2018	ONTAP 9.4, 9.5 updates
Version 1.4	June 2018	ONTAP 9.3 updates; removed 7-Mode references
Version 1.3	August 2016	ONTAP 9.0 for 8-node MetroCluster, unmirrored aggregates
Version 1.2	February 2016	Clustered Data ONTAP 8.3.2 updates for FCIP configurations and new features
Version 1.1	September 2015	Clustered Data ONTAP 8.3.1 updates for two-node configurations and minor corrections
Version 1.0	April 2015	Initial release

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**

Copyright © 2021 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

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

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

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

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

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**

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

TR-4375-1021