Technical Report

NetApp MetroCluster

Solution Architecture and Design

Mike Braden, NetApp
November 2019 | TR-4705

Abstract

This document describes high-level architecture and design concepts for NetApp® MetroCluster™ features in NetApp ONTAP® 9.7 storage management software.

NetApp®

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 MetroCluster Overview

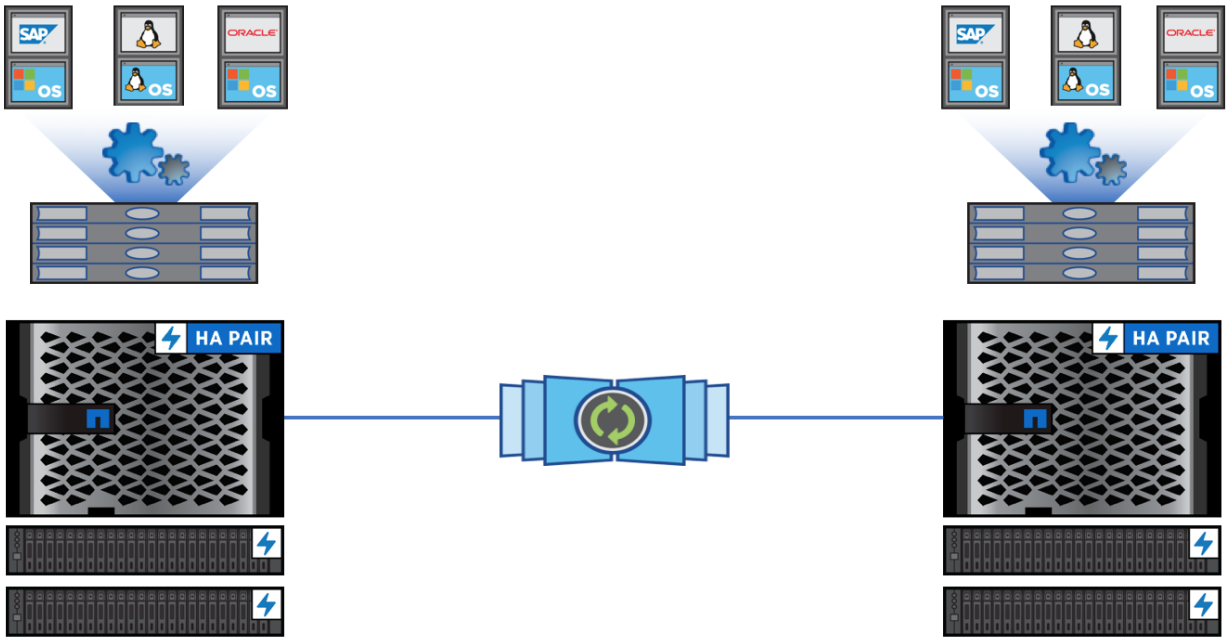
Enterprise-class customers must meet increasing service-level demands while maintaining cost and operational efficiency. As data volumes proliferate and more applications move to shared virtual infrastructures, the need for continuous availability for both mission-critical and other business applications dramatically increases.

In an environment with highly virtualized infrastructures running hundreds of business-critical applications, an enterprise would be severely affected if these applications became unavailable. Such a critical infrastructure requires zero data loss and system recovery in minutes rather than hours. This requirement is true for both private and public cloud infrastructures, as well as for the hybrid cloud infrastructures that bridge the two.

NetApp MetroCluster software is a solution that combines array-based clustering with synchronous replication to deliver continuous availability and zero data loss at the lowest cost. Administration of the array-based cluster is simpler because the dependencies and complexity normally associated with host-based clustering are eliminated. MetroCluster immediately duplicates all your mission-critical data on a transaction-by-transaction basis, providing uninterrupted access to your applications and data. And unlike standard data replication solutions, MetroCluster works seamlessly with your host environment to provide continuous data availability while eliminating the need to create and maintain complicated failover scripts. With MetroCluster, you can:

- Protect against hardware, network, or site failure with transparent switchover
- Eliminate planned and unplanned downtime and change management
- Upgrade hardware and software without disrupting operations
- Deploy without complex scripting, application, or operating system dependencies
- Achieve continuous availability for VMware, Microsoft, Oracle, SAP, or any critical application

Figure 1) MetroCluster.



NetApp MetroCluster enhances the built-in high-availability (HA) and nondisruptive operations of NetApp hardware and ONTAP storage software, providing an additional layer of protection for the entire storage and host environment. Whether your environment is composed of standalone servers, HA server clusters,

or virtualized servers, MetroCluster seamlessly maintains application availability in the face of a total storage outage. Such an outage could result from loss of power, cooling, or network connectivity; a storage array shutdown; or operational error.

MetroCluster is an array-based, active-active clustered solution that eliminates the need for complex failover scripts, server reboots, or application restarts. MetroCluster maintains its identity in the event of a failure and thus provides application transparency in switchover and switchback events. In fact, most MetroCluster customers report that their users experience no application interruption when a cluster recovery takes place. MetroCluster provides the utmost flexibility, integrating seamlessly into any environment with support for mixed protocols.

MetroCluster provides the following benefits:

- SAN and NAS host support
- Mixed controller deployments with AFF and FAS
- Integration with NetApp SnapMirror® technology to support asynchronous replication, distance, and SLA requirements
- Support for synchronous replication over FC or IP networks
- Zero RPO and near-zero RTO
- MetroCluster is a no-charge feature built into ONTAP
- Mirror only what you need
- Support for third-party storage with NetApp FlexArray® technology
- Data efficiencies include deduplication, compression, and compaction

1.1 Target Audience

The audience for this document includes, but is not limited to, sales engineers, field consultants, professional services personnel, IT managers, partner engineers, and customers who want to take advantage of an infrastructure built to deliver resiliency and simplicity.

1.2 Data Protection with NetApp SyncMirror Technology

At the simplest level, synchronous replication means any change must be made to both sides of mirrored storage. For example, an Oracle database commits a transaction, and data is written to a redo log on synchronously mirrored storage. The storage system must not acknowledge the write until it has been committed to nonvolatile media on both sites. Only then is it safe to proceed without the risk of data loss.

The use of synchronous replication technology is only the first step in designing and managing a synchronous replication solution. The most important consideration is to know exactly what happens during various planned and unplanned failure scenarios. Not all synchronous replication solutions offer the same capabilities. When a customer asks for a solution that delivers a recovery point objective (RPO) of zero (meaning zero data loss), we must think about failure scenarios. In particular, we must determine the expected result when replication is impossible due to loss of connectivity between sites.

1.3 True HA Data Center with MetroCluster

MetroCluster replication is based on NetApp SyncMirror® technology, which is designed to easily switch into and out of synchronous mode. This technology meets the requirements of a majority of customers who demand synchronous replication under normal conditions. In the case of a partial failure such as a backhoe event that severs all connectivity between sites, it is preferable to have the storage system continue operating but in a nonreplicated state.

NetApp MetroCluster is ideal for organizations that require 24/7 operation for critical business applications. By synchronously replicating data between NetApp AFF and/or FAS hybrid systems that are colocated in the same data center, between buildings, across a campus, or across regions, MetroCluster

transparently fits into any disaster recovery (DR) and business continuity strategy. In addition, third-party storage systems are also supported with the NetApp FlexArray feature.

1.4 Campus, Metro, and Regional Protection

NetApp MetroCluster can also significantly simplify the design, deployment, and maintenance of campus wide or metropolitanwide HA solutions, with distances of up to 700km between sites. During a total site disruption, data services are restored at the secondary site in a matter of seconds with an automated single command and no complex failover scripts or restart procedures.

1.5 Your Choice of Protection

Achieve new levels of flexibility and choice for business continuity. When deployed with ONTAP 9 software, MetroCluster allows you to scale from a two-node to an eight-node cluster (four nodes on each end of the replication), even with a mix of AFF and FAS hybrid controllers. You can even choose which storage pools or aggregates to replicate, so that you do not have to commit your full dataset to a synchronous DR relationship.

Note: Synchronous replication over an FC network is supported with two-node, four-node, and eight-node configurations. Synchronous replication over an IP network is currently only supported with a four-node configuration.

1.6 WAN-Based DR

If your business is geographically dispersed beyond metropolitan distances, you can add NetApp SnapMirror software to replicate data across your global network simply and reliably. NetApp SnapMirror software works with your MetroCluster solution to replicate data at high speeds over WAN connections, protecting your critical applications from regional disruptions.

1.7 Simplified Administration: Set It Once

Most array-based data replication solutions require duplicate efforts for storage system administration, configuration, and maintenance because the replication relationships between the primary and secondary storage arrays are managed separately. This duplication increases management overhead, and it can also expose you to greater risk if configuration inconsistencies arise between the primary and secondary storage arrays. Because MetroCluster is a true clustered storage solution, the active-active storage pair is managed as a single entity, eliminating duplicate administration work and maintaining configuration consistency.

1.8 Application Transparency

MetroCluster is designed to be transparent and agnostic to any front-end application environment, and few if any changes are required for applications, hosts, and clients. Connection paths are identical before and after switchover, and most applications, hosts, and clients (NFS and SAN) do not need to reconnect or rediscover their storage but instead automatically resume. SMB applications, including SMB 3 with continuous availability shares, need to reconnect after a switchover or a switchback. This need is a limitation of the SMB protocol.

2 Architecture

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

2.1 MetroCluster Physical Architecture

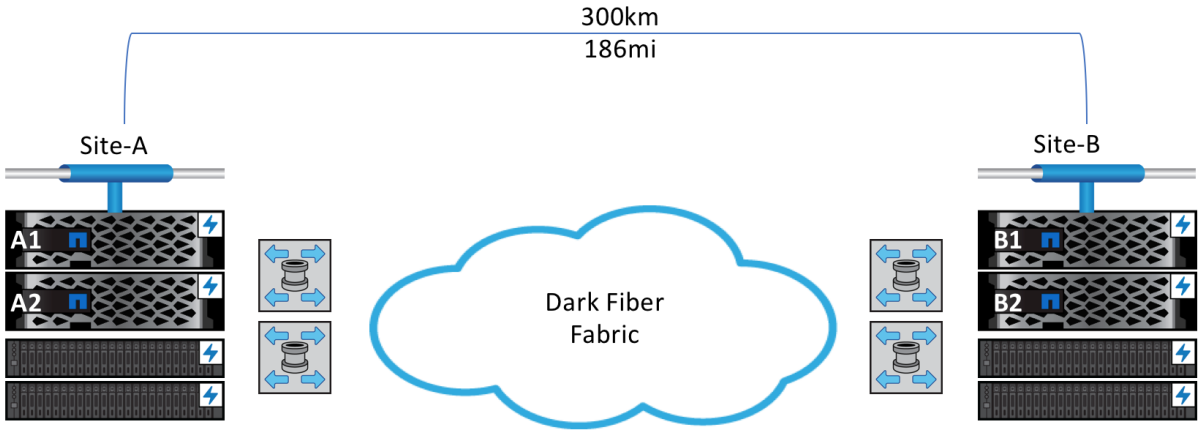
MetroCluster configurations protect data by using two distinct clusters that are separated by a distance of up to 700km. Each cluster synchronously mirrors the data and configuration information of the other. Effectively, all storage virtual machines (SVMs) and their associated configurations are replicated. Independent clusters provide isolation and resilience to logical errors.

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

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

Figure 2 shows the basic four-node MetroCluster configuration. The two data centers, A and B, are separated by a distance of up to 300km with Interswitch links (ISLs) running over dedicated FC links. If you are using a MetroCluster IP (MC-IP) deployment, the maximum distance is 700km. The cluster at each site consists of two nodes in an HA pair. We use this configuration and naming throughout this report. Review the section “Deployment Options” for the various deployment options.

Figure 2) Four-node MetroCluster deployment.



The two clusters and sites are connected by two separate networks that provide the replication transport. The cluster peering network is an IP network that is used to replicate cluster configuration information between the sites. The shared storage fabric is an FC connection that is used for storage and NVRAM synchronous replication between the two clusters. For MC-IP, the fabric is IP based, and replication uses both iWARP for NVRAM and iSCSI for disk replication. All storage is visible to all controllers through the shared storage fabric.

2.2 Comparing MetroCluster FC and MetroCluster IP

MetroCluster IP (MC-IP) was introduced with ONTAP 9.3 and uses Ethernet/IP ISLs for the fabric, unlike MetroCluster FC (MC-FC), which uses Fibre Channel ISLs. Additionally, MC-IP clusters use high-speed Ethernet for both NVRAM and SyncMirror replication.

MetroCluster IP has several features that offer reduced operational costs, including the ability to use site-to-site links that are shared with other non-MetroCluster traffic (shared layer-2). Starting in ONTAP 9.7, MetroCluster IP is offered without dedicated switches, allowing the use of existing switches as long as they are compliant with the requirements for MetroCluster IP. For more information see the MetroCluster IP Installation and Configuration Guide.

Table 1 summarizes the differences between these two configurations and indicates how data is replicated between the two MetroCluster sites. For deployment options and switchover behavior, see the section “Deployment Options” and the section 5, “Resiliency for Planned and Unplanned Events.”

Table 1) Compare MetroCluster FC and MetroCluster IP.

Function	MetroCluster FC	MetroCluster IP
MetroCluster fabric	FC ISLs	Ethernet/IP ISLs
Fabric fibre switches	Two per site	None
SAS bridges	Two per site	None
Fabric Ethernet switches	None	Two per site
FC-VI adapters	Yes Number of adapters depends on controller	None
Fabric 25G/40G/100G Ethernet adapters	None	One per node depending on platform Adapter is used to replication both iWARP and iSCSI
Intercluster	Switchless and switched	Switchless and switched
Shelves	Physically visible to both sites	Not visible to remote clusters
NVRAM replication	FC protocol	IP/iWARP
SyncMirror replication	FC protocol	IP/iSCSI
Configuration replication services	No changes	No changes
MetroCluster size	Two, four, and eight nodes	Four nodes only
MetroCluster stretch	Yes	No
Advanced Disk Partitioning	No	Yes, for AFF only

2.3 HA and Remote Replication

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

The four-node and eight-node architectures provides both local HA failover and remote DR switchover. Each node has an HA partner in the same local cluster and a DR partner in the remote cluster, as shown in Figure 3. A1 and A2 are HA partners, as are B1 and B2. Node A1 and B1 are DR partners, as are A2 and B2. NVRAM is replicated to both the HA and the DR partner, as explained further in the section “Campus, Metro, and Regional Protection.” The DR partner for a node is automatically selected when MetroCluster is configured, and the partner is chosen according to system ID (NVRAM ID) order.

Note: System ID is hardcoded and not changeable. You should note the system IDs before the cluster is configured to create proper partnerships between local and remote peers.

Figure 3) HA and DR groups.

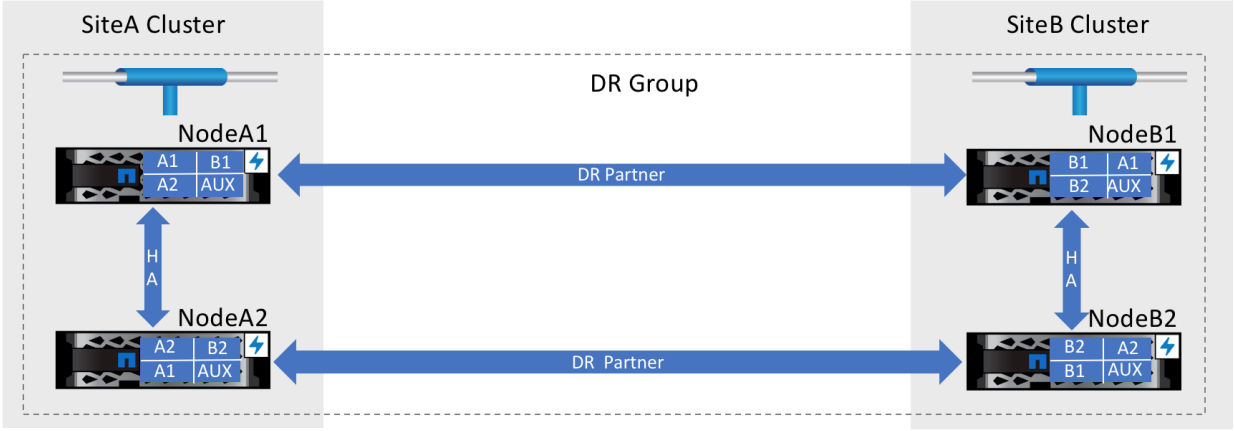
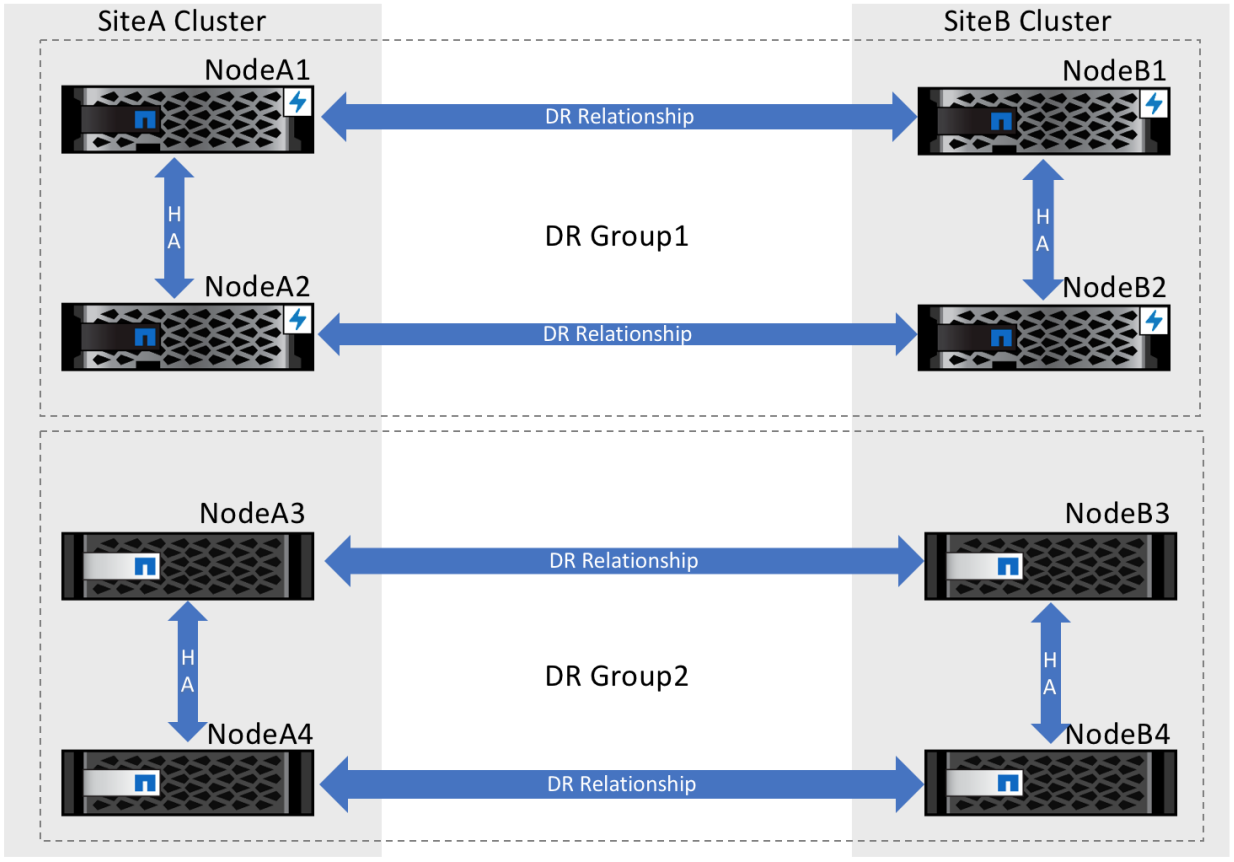


Figure 4 depicts an eight-node MetroCluster configuration and the DR group relationships. In an eight-node deployment, there are two independent DR groups. Each DR group can support different hardware as long as the hardware matches the local and remote clusters. For example, the hardware can be different in each DR group; Group 1 could use an AFF A700, and Group 2 could use an FAS8200.

Figure 4) Eight-node DR groups.



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

In a switchover, one cluster assumes the storage and services of the other while continuing to perform its own workloads. For example, if site A switches over to site B, the cluster B nodes take temporary control of the storage and services owned by cluster A. After switchover, the SVMs from cluster A are brought online and continue running on cluster B.

Switchover can be negotiated (planned), for example, to perform testing or site maintenance, or it can be forced (unplanned) in the event of a disaster that destroys one of the sites. Switchback is the process in which the surviving cluster sends the switched-over resources back to their original location to restore the steady operational state. Switchback is coordinated between the two clusters and is always a planned operation. Either site can switch over to the other.

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

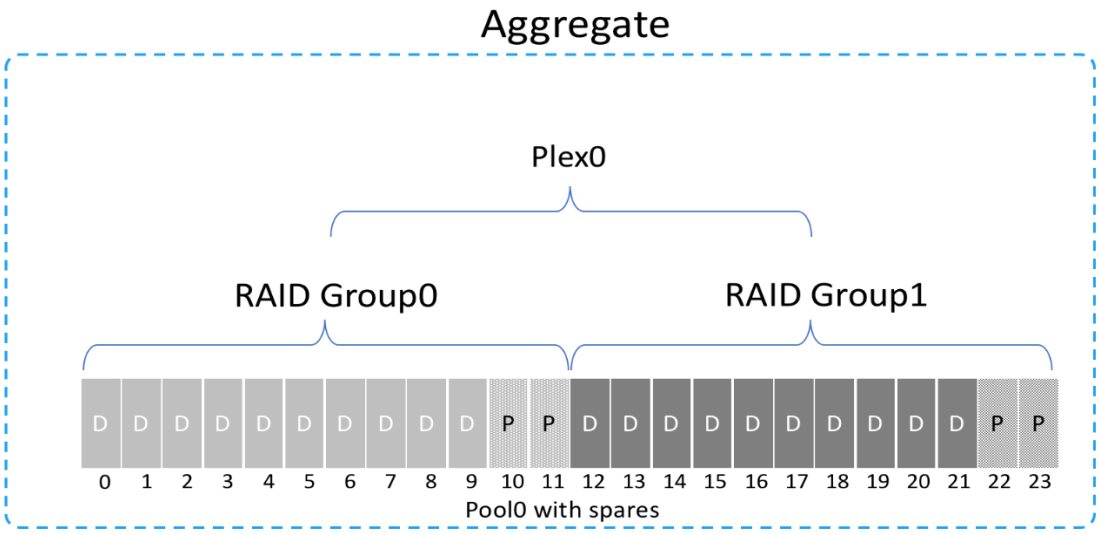
2.4 MetroCluster Data Replication

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

SyncMirror Storage Replication

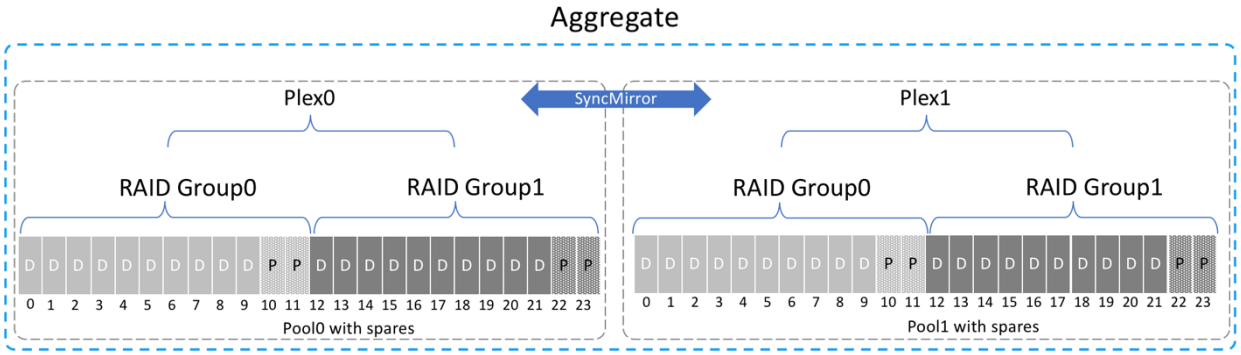
An ONTAP system stores data in NetApp FlexVol® volumes that are provisioned from aggregates. Each aggregate contains a NetApp WAFL® file system. In a configuration without MetroCluster, the disks in each aggregate consist of a single or multiple RAID groups, known as a plex (Figure 5). The plex resides in local storage attached to the controller.

Figure 5) Unmirrored aggregate: Plex0.



In a MetroCluster configuration, each aggregate consists of two plexes that are physically separated: a local plex and a remote plex (Figure 6). All storage is shared and is visible to all the controllers in the MetroCluster configuration. The local plex must contain only disks from the local pool (`pool0`), and the remote plex must contain only disks from the remote pool. The local plex is always `plex0`. Each remote plex has a number other than `0` to indicate that it is remote (for example, `plex1` or `plex2`).

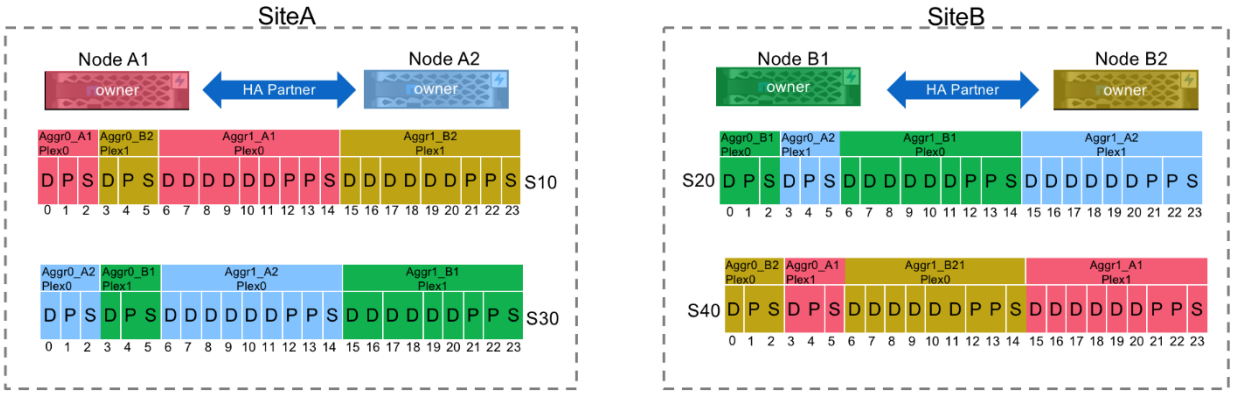
Figure 6) MetroCluster mirrored aggregate.



Both mirrored and unmirrored aggregates are supported with MetroCluster starting with 9.0. However, unmirrored aggregates are not currently supported with MC-IP. The `-mirror true` flag must therefore be used when creating aggregates after MetroCluster has been configured; if it is not specified, the `create` command fails. The number of disks that are specified by the `-diskcount` parameter is automatically halved. For example, to create an aggregate with six usable disks, 12 must be specified as the disk count. That way, the local plex is allocated six disks from the local pool, and the remote plex is allocated six disks from the remote pool. The same process applies when adding disks to an aggregate; twice the number of disks must be specified as are required for capacity.

The example in Figure 7 shows how the disks have been assigned to the aggregates. Each node has a root aggregate and one data aggregate. Each root aggregate contains six drives for each node, assuming two minimum shelves used per cluster, of which three are on the local plex and three are on the remote plex. Therefore, the available capacity of the aggregate is three drives. Similarly, each of the data aggregates contains 18 drives: nine local and nine remote. With MetroCluster and particularly with AFF, the root aggregate uses RAID 4, and data aggregates use RAID DP® or RAID-TEC™.

Figure 7) Root and data aggregates.



In normal MetroCluster operation, both plexes are updated simultaneously at the RAID level. All writes, whether from client and host I/O or cluster metadata, generate two physical write operations, one to the

local plex and one to the remote plex, using the ISL connection between the two clusters. By default, reads are fulfilled from the local plex.

Aggregate Snapshot Copies

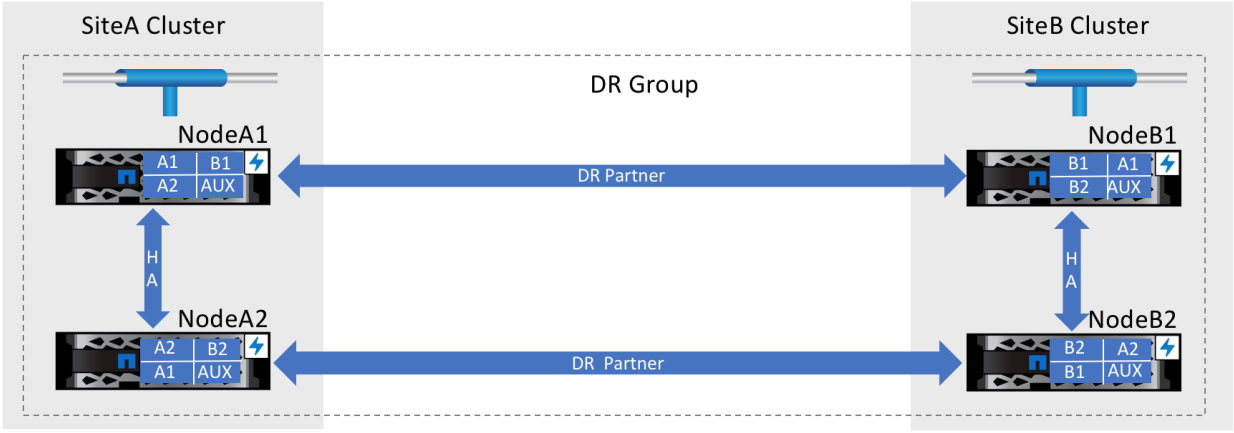
Automatic aggregate NetApp Snapshot™ copies are taken, and, by default, 5% of aggregate capacity is reserved for these Snapshot copies. These Snapshot copies are used as the baseline for resyncing the aggregates when necessary.

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

NVRAM Cache Mirroring

In an ONTAP HA pair, each node mirrors its NVRAM to the other node by using the HA interconnect. NVRAM is split into two segments, one for each node's NVRAM. MetroCluster provides additional mirroring; each node has a DR partner node on the other site, and the NVRAM is mirrored to the DR partner over the ISL connection. Therefore, in a four-node configuration, each node's NVRAM is mirrored twice—to the HA partner and the DR partner—and each node's NVRAM is split into four segments, as shown in Figure 8.

Figure 8) NVRAM allocation.



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

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

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

2.5 Configuration Replication Services

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

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

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

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

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

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

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

```
tme-mcc-A::> volume show -volume MDV*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
tme-mcc-A MDV_CRS_cd7628c7f1cc11e3840800a0985522b8_A
                       aggr1_tme_A1 online     RW         10GB     9.50GB    5%
tme-mcc-A MDV_CRS_cd7628c7f1cc11e3840800a0985522b8_B
                       aggr1_tme_A2 online     RW         10GB     9.50GB    5%
tme-mcc-A MDV_CRS_e8fef00df27311e387ad00a0985466e6_A
                       aggr1_tme_B1 -          RW            -          -     -
tme-mcc-A MDV_CRS_e8fef00df27311e387ad00a0985466e6_B
                       aggr1_tme_B2 -          RW            -          -     -
```

2.6 Active-Active and Active-Passive Configurations

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

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

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

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

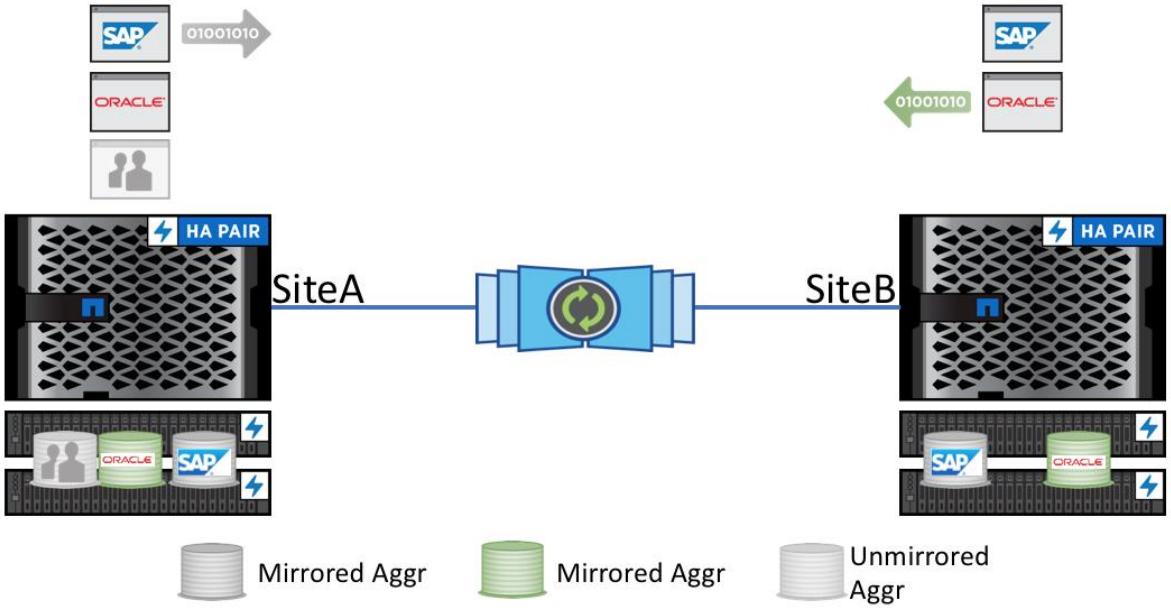
2.7 Unmirrored Aggregates

Beginning with ONTAP 9, MetroCluster configurations support unmirrored data aggregates for data that does not require the redundant mirroring provided by MetroCluster configurations. Unmirrored aggregates are not protected in the event of a site disaster and write I/O to these aggregates must be counted for when sizing the ISLs.

Unmirrored aggregates are only supported on MetroCluster FC configurations. Unmirrored aggregates are not supported on MetroCluster IP.

Figure 9 depicts the granular control of mirroring aggregates: SAP is mirrored to the Site B cluster, and Oracle is mirrored to its Site A cluster. The Home User directory on Site A is not a critical aggregate, and it is not mirrored to the remote cluster. In the event of a failure on Site A, this aggregate is not available.

Figure 9) Unmirrored aggregates in MetroCluster.



When considering unmirrored aggregates in MetroCluster FC, keep in mind the following issues:

- In MetroCluster FC configurations, the unmirrored aggregates are only online after a switchover if the remote disks in the aggregate are accessible. If the ISLs fail, the local node might be unable to access the data in the unmirrored remote disks. The failure of an aggregate can lead to a reboot of the local node.
- Drives and array LUNs are owned by a specific node. When you create an aggregate, all drives in that aggregate must be owned by the same node, which becomes the home node for that aggregate.
- Aggregate names should conform to the naming scheme you determined when you planned your MetroCluster configuration.
- The ONTAP Data Protection Guide contains more information about mirroring aggregates.

3 Deployment Options

MetroCluster is a fully redundant configuration with identical hardware required at each site. Additionally, MetroCluster offers the flexibility of both stretch and fabric-attached configurations. Table 2 depicts the different deployment options at a high level and presents the supported switchover features.

Table 2) Hardware requirements.

Feature	IP Configuration	Fabric-Attached Configuration		Stretch Configuration	
		Four-Node or Eight-Node	Two-Node	Two-Node Bridge-Attached	Two-Node Direct-Attached
Number of controllers	Four	Four or eight	Two	Two	Two
FC switch storage fabric	No	Yes	Yes	No	No

IP switch storage fabric	Yes	No	No	No	No
FC-to-SAS bridges	No	Yes	Yes	Yes	Yes
Direct-attached storage	Yes (local attached only)	No	No	No	Yes
Supports local HA	Yes	Yes	No	No	No
Supports automatic switchover	Yes (with mediator)	Yes	Yes	Yes	Yes
Supports unmirrored aggregates	No	Yes	Yes	Yes	Yes
Supports array LUNs	No	Yes	Yes	Yes	Yes

3.1 Stretch and Stretch-Bridged Configurations

A two-node MetroCluster stretch configuration uses only one node per site. Two-node supports both stretch using SAS cables (Figure 10) and stretch bridged using SAS bridges to extend the distance inside the data center, for example, between data halls, as noted in Figure 11. Stretch and stretch-bridged deployments are ideal for data center deployments with total flexibility. Both deployments have reduced infrastructure demands in terms of cabling, FC switches, and rack space.

Figure 10) Two-node stretch configuration.

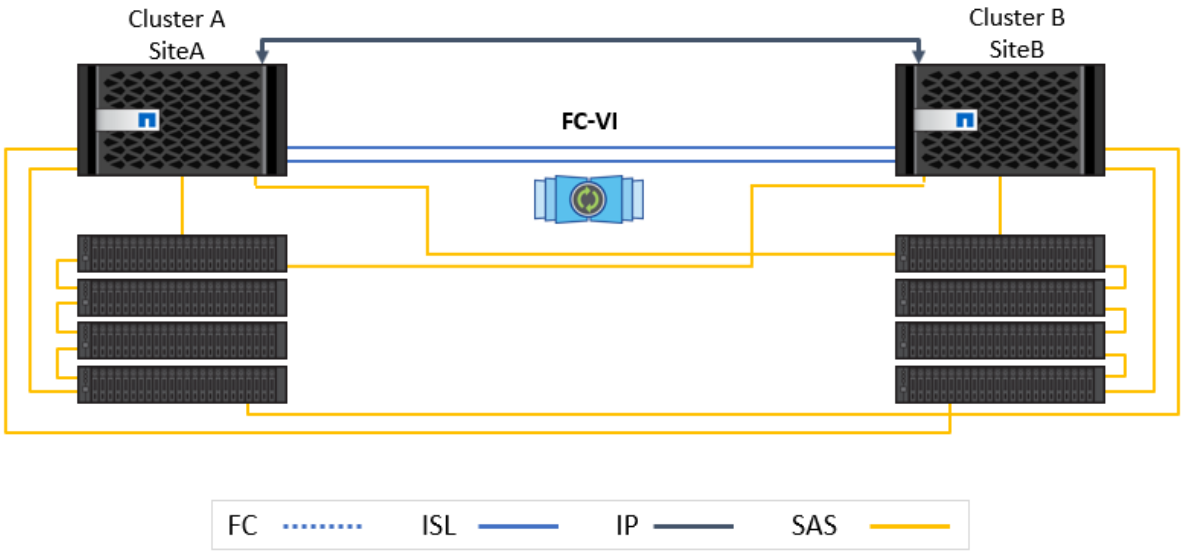
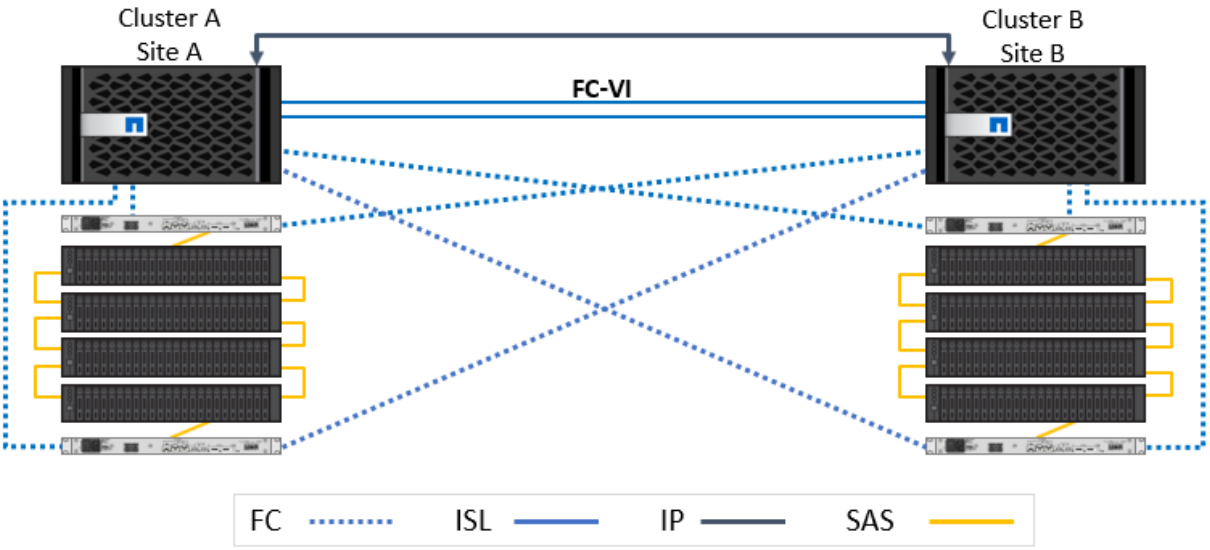


Figure 11) Two-node stretch-bridge configuration.



3.2 Fabric-Attached FC Configuration

Fabric-attached MetroCluster deployments provide the most flexibility to meet distance and controller requirements, and this deployment method supports both AFF and FAS with distances up to 300km.

With a fabric-attached configuration, it is possible to deploy two-, four-, and eight-node MetroCluster systems with any combination of hardware supported by MetroCluster. Table 2 highlights the supported hardware for ONTAP 9.3. However, NetApp recommends that you review the NetApp Interoperability Matrix Tool (IMT) for the latest supported hardware and software.

Note: Starting with ONTAP 9.6, the NetApp Hardware Universe contains maximums and platform compatibility information. Older releases remain in the Interoperability Matrix Tool.

The following figures depict the different deployment options.

Figure 12 depicts a two-node MetroCluster deployment. The number of ISLs and switch types depend on the distance between locations and the amount of write I/O required. Refer to the IMT and the ISL sizing tool listed in the references for more information.

Figure 12) Two-node fabric-attached deployment.

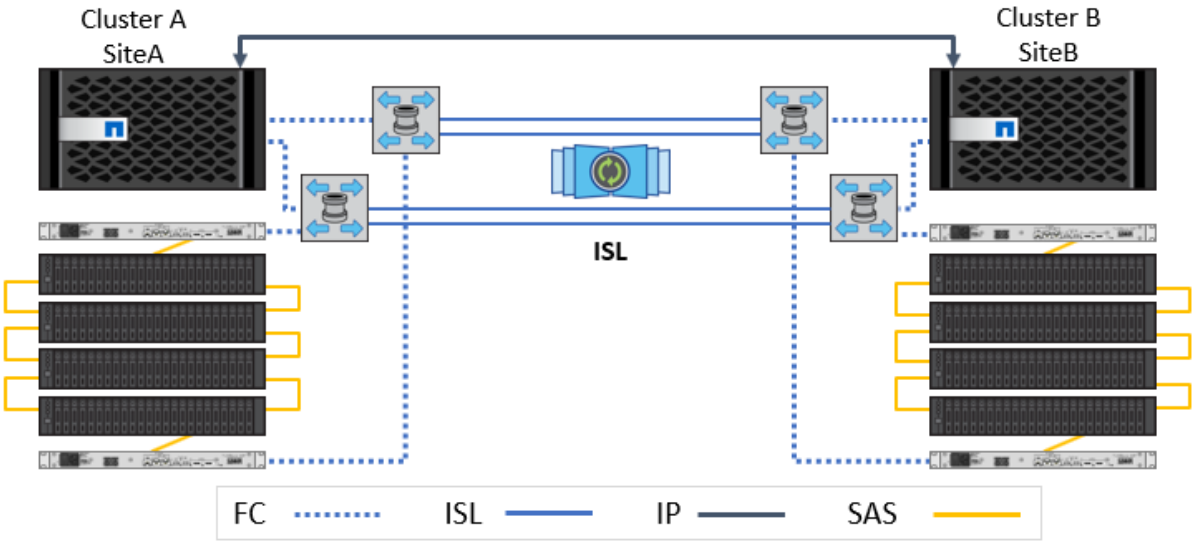


Figure 13 and Figure 14 depict the four-node and eight-node deployment options. For specific hardware and ISL requirements, consult the IMT.

Figure 13) Four-node fabric-attached deployment.

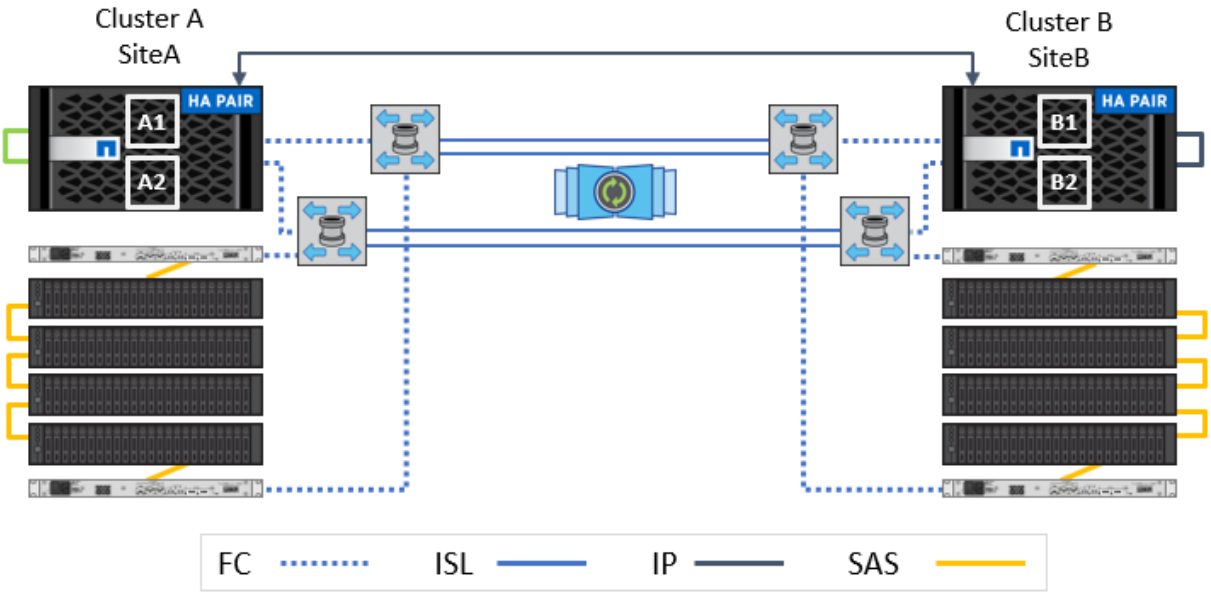


Figure 14) Eight-node fabric-attached deployment.

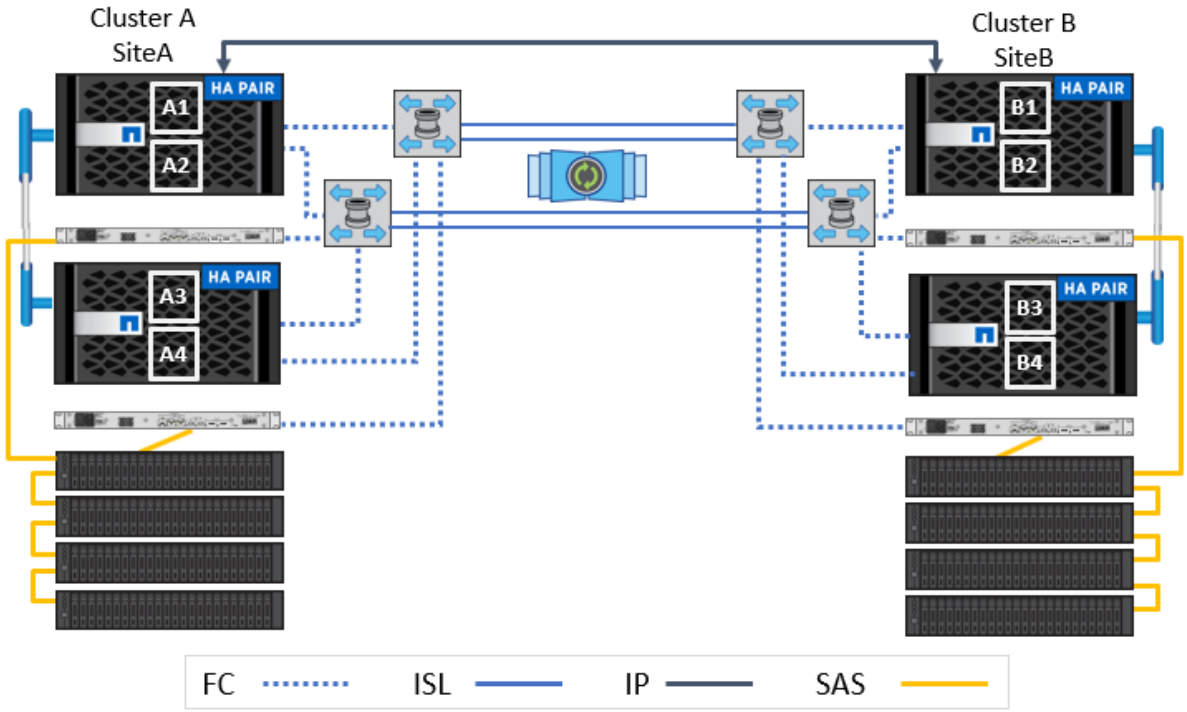
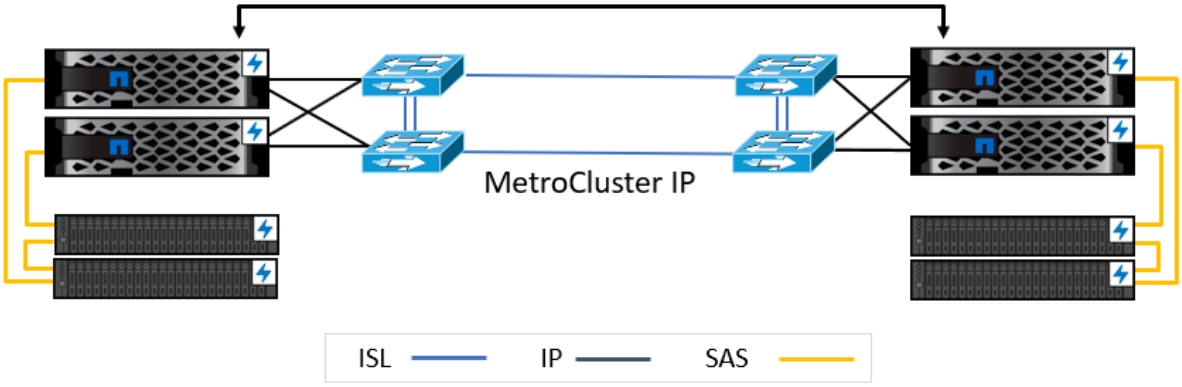


Figure 15 depicts the four-node MetroCluster fabric-attached IP deployment option.

Figure 15) Four-node MetroCluster IP.



4 Technology Requirements

This section covers the technology requirements for the MetroCluster FC and IP solution.

4.1 Hardware Requirements

The following tables list the hardware components that are required to implement the solution. The hardware components that are used in any implementation of the solution might vary based on customer deployment and whether MetroCluster FC or IP is desired.

Table 3) MetroCluster FC and Stretch Hardware

Controllers	Switches	FC Bridged
FAS8200	• Cisco: FC, and FCIP switches • Brocade: FC and FCIP switches	ATTO 6500, ATTO 7500, and ATTO 7600
FAS8300		
FAS9000		
AFF A300		
AFF A400		
AFF A700		

Table 4) MetroCluster IP Hardware

Controllers	Switches
FAS2750	• Cisco: Ethernet • Broadcom: Ethernet (optionally without switches, except A220/FAS2750)
FAS8200	
FAS9000	
AFF A220	
AFF A300	
AFF A320	
AFF A700	
AFF A800	

4.2 Software Requirements

ONTAP is the only software component required to implement the solution. MetroCluster is a standard ONTAP feature that does not require a separate license. Standard ONTAP licensing is used for client and host side protocols or the additional capabilities for SnapMirror to protect data using an asynchronous mirror or XDP to replicate data to a third cluster for backup data protection.

5 Resiliency for Planned and Unplanned Events

This section covers the different types of failures and disasters and how MetroCluster configuration maintains availability, data protection, and remediation.

5.1 Single-Node Failure

Consider a scenario in which a single component in the local HA pair fails. In a four-node MetroCluster configuration, this failure might lead to an automatic or a negotiated takeover of the impaired node, depending on the component that failed. Data recovery is described in the ONTAP High-Availability Configuration Guide. In a two-node MetroCluster configuration, this failure leads to an automatic unplanned switchover (AUSO).

5.2 Sitewide Controller Failure

Consider a scenario in which all controller modules fail at a site because of a loss of power, the replacement of equipment, or a disaster. Typically, MetroCluster configurations cannot differentiate between failures and disasters. However, witness software, such as the MetroCluster Tiebreaker software, can differentiate between these two possibilities. A sitewide controller failure condition can lead to an automatic switchover if ISLs and switches are up and the storage is accessible.

The ONTAP High-Availability Configuration Guide has more information about how to recover from sitewide controller failures that do not include controller failures, as well as failures that include one or more controllers.

5.3 ISL Failure

Consider a scenario in which the links between the sites fail. In this situation, the MetroCluster configuration takes no action. Each node continues to serve data normally, but the mirrors are not written to the respective DR sites because access to them is lost.

5.4 Multiple Sequential Failures

Consider a scenario in which multiple components fail in sequence. For example, a controller module, a switch fabric, and a shelf fail in a sequence and result in a storage failover, fabric redundancy, and SyncMirror sequentially protecting against downtime and data loss.

Table describes failure types and the corresponding DR mechanism and recovery method.

Note: AUSO is only supported on MetroCluster IP configurations when using ONTAP Mediator and ONTAP 9.7 or later.

Table 5) Failure types and recovery methods.

Failure Type	DR Mechanism		Summary of Recovery Methods	
	Four-Node Configuration	Two-Node Configuration	Four-Node Configuration	Two-Node Configuration
Single-node failure	Local HA failure	AUSO	Not required if automatic failover and giveback are enabled.	After the node is restored, manual healing and switchback by using the `metrocluster heal -phase aggregates`, `metrocluster heal -phase root-aggregates`, and `metrocluster switchback` commands are required.
Site failure	MetroCluster switchover		After the node is restored, manual healing and switchback using the `metrocluster healing` and `metrocluster switchback` commands are required.	

Failure Type	DR Mechanism		Summary of Recovery Methods	
	Four-Node Configuration	Two-Node Configuration	Four-Node Configuration	Two-Node Configuration
Sitewide controller failure	AUSO Only if the storage at the disaster site is accessible.	AUSO Same as single-node failure.	After the node is restored, manual healing and switchback using the `metrocluster healing` and `metrocluster switchback` commands are required.	
ISL failure	No MetroCluster switchover. The two clusters independently serve their data.		Not required for this type of failure. After you restore connectivity, the storage resynchronizes automatically.	
Multiple sequential failures	Local HA failover followed by MetroCluster forced switchover using the `metrocluster switchover -forced-ondisaster` command. **Note:** Depending on the component that failed, a forced switchover might not be required.	MetroCluster forced switchover using the `metrocluster switchover -forced-ondisaster` command.	After the node is restored, manual healing and switchback using the `metrocluster healing` and `metrocluster switchback` commands are required.	

5.5 Eight-Node and Four-Node Nondisruptive Operations

In the case of an issue limited to a single node, failover and giveback in the local HA pair provide continued nondisruptive operation. In this case, the MetroCluster configuration does not require a switchover to the remote site.

Because the eight-node or four-node MetroCluster configuration consists of one or more HA pairs at each site, each site can withstand local failures and perform nondisruptive operations without requiring a switchover to the partner site. The operation of the HA pair is the same as HA pairs in configurations other than MetroCluster.

For four-node and eight-node MetroCluster configurations, node failures due to panic or power loss can cause an automatic switchover.

If a second failure occurs after a local failover, the MetroCluster switchover event provides continued nondisruptive operations. Similarly, after a switchover operation in the event of a second failure in one of the surviving nodes, a local failover event provides continued nondisruptive operations. In this case, the single surviving node serves data for the other three nodes in the DR group.

Consequences of Local Failover After Switchover

If a MetroCluster switchover occurs, and an issue then arises at the surviving site, a local failover can provide continued, nondisruptive operation. However, the system is at risk because it is no longer in a redundant configuration.

If a local failover occurs after a switchover has occurred, a single controller serves data for all storage systems in the MetroCluster configuration, leading to possible resource issues. The surviving controller is also vulnerable to additional failures.

5.6 Two-Node Nondisruptive Operations

If one of the two sites suffer a panic failure, MetroCluster switchover provides continued nondisruptive operation. If the power loss affects both the node and the storage, then the switchover is not automatic, and there is a disruption until the `metrocluster switchover` command is issued.

Because all storage is mirrored, a switchover operation can be used to provide nondisruptive resiliency in case of a site failure similar to that found in a storage failover in an HA pair for a node failure.

For two-node configurations, the same events that trigger an automatic storage failover in an HA pair trigger AUSO. This fact means that a two-node MetroCluster configuration has the same level of protection as an HA pair.

5.7 Overview of the Switchover Process

The MetroCluster switchover operation enables immediate resumption of services following a disaster by moving storage and client access from the source cluster to the remote site. You must be aware of what changes to expect and which actions you need to perform if a switchover occurs.

During a switchover operation, the system takes the following actions:

- Ownership of the disks that belong to the disaster site is changed to the DR partner. This situation is similar to the case of a local failover in an HA pair in which ownership of the disks belonging to the down partner is changed to the healthy partner.
- The surviving plexes that are located on the surviving site but belong to the nodes in the disaster cluster are brought online on the cluster at the surviving site.
- The sync source SVM that belongs to the disaster site is brought down only during a negotiated switchover.

  Note: This approach is applicable only to a negotiated switchover.

- The sync destination SVM belonging to the disaster site is brought up.

While being switched over, the root aggregates of the DR partner are not brought online.

The `metrocluster switchover` command switches over the nodes in all DR groups in the MetroCluster configuration. For example, in an eight-node MetroCluster configuration, this command switches over the nodes in both DR groups.

If you are only switching over services to the remote site, you should perform a negotiated switchover without fencing the site. If storage or equipment is unreliable, you should fence the disaster site and then perform an unplanned switchover. Fencing prevents RAID reconstructions when the disks power up in a staggered manner.

Note: This procedure should be only used if the other site is stable and you do not intend to take it offline.

5.8 Difference Between MetroCluster FC and IP Switchover

In MetroCluster IP configurations, the remote disks are accessed through the remote DR partner nodes acting as iSCSI targets. Therefore, the remote disks are not accessible when the remote nodes are taken down in a switchover operation. This approach results in differences with MetroCluster FC configurations:

- Mirrored aggregates that are owned by the local cluster become degraded.
- Mirrored aggregates that were switched over from the remote cluster become degraded.

MetroCluster 9.5 introduces a new feature called Auto Heal for MetroCluster IP. This functionality combines healing root and data aggregates in a simplified process when performing a planned switchover and switchback such as DR testing.

5.9 ONTAP Mediator

ONTAP 9.7 includes a new MetroCluster IP solution for handling failures called ONTAP Mediator. Additional functionality has been added to ONTAP, including the use of ONTAP Mediator to provide AUSO capability for MetroCluster IP. ONTAP Mediator is installed on a Red Hat Enterprise Linux or CentOS Linux physical or virtual server located in a separate (third) failure domain from the MetroCluster nodes.

For more information about the requirements for ONTAP Mediator and details about failures, see to the MetroCluster IP Installation and Configuration Guide.

Note: Managing the same MetroCluster configuration with both Tiebreaker and ONTAP Mediator is not supported. Only one of the products can be used to manage a MetroCluster configuration.

5.10 NetApp Tiebreaker

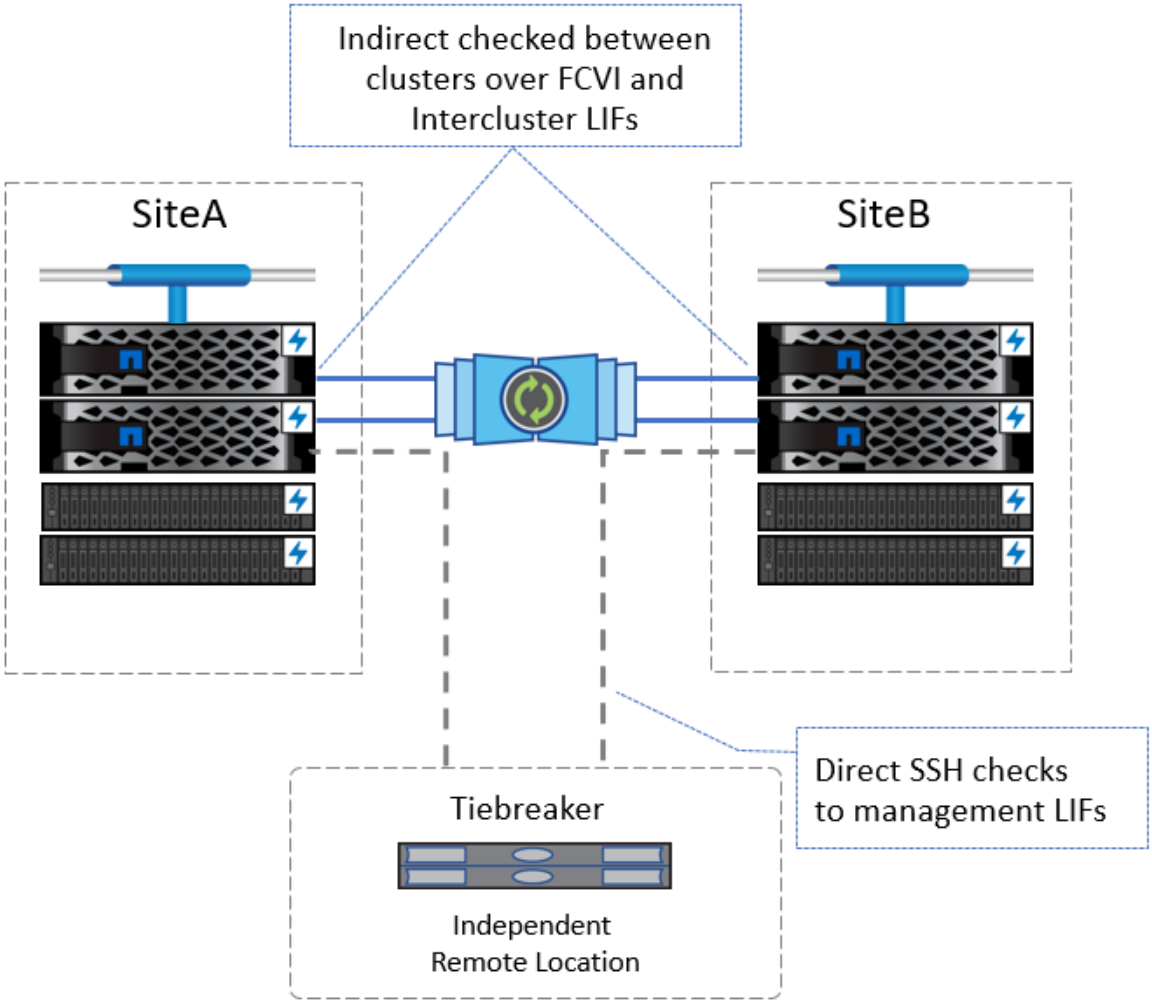
The MetroCluster Tiebreaker software alerts you if all connectivity between the sites is lost. The MetroCluster Tiebreaker software supports all the MetroCluster configurations that are supported in ONTAP 8.3 and 9.0 through 9.7.

The Tiebreaker software resides on a Linux host. You need Tiebreaker software only if you want to monitor two clusters and the connectivity status between them from a third site. Doing so enables each partner in a cluster to distinguish between an ISL failure, when intersite links are down, from a site failure.

Note: You should only have one MetroCluster Tiebreaker monitor per MetroCluster configuration to avoid any conflict between multiple Tiebreaker monitors.

The NetApp MetroCluster Tiebreaker software checks the reachability of the nodes in a MetroCluster configuration and the cluster to determine whether a site failure has occurred. The Tiebreaker software also triggers an alert under certain conditions. MetroCluster Tiebreaker detects direct and indirect failures, as shown in Figure 16, so that the Tiebreaker doesn’t initiate a switchover if the fabric is intact.

Figure 16) MetroCluster Tiebreaker checks.



Detecting Failures with MetroCluster Tiebreaker

The Tiebreaker software resides on a Linux host. You need the Tiebreaker software only if you want to monitor two clusters and the connectivity status between them from a third site. Doing so enables each partner in a cluster to distinguish between an ISL failure, when intersite links are down, from a site failure.

After you install the Tiebreaker software on a Linux host, you can configure the clusters in a MetroCluster configuration to monitor for disaster conditions.

Detecting Intersite Connectivity Failures

The MetroCluster Tiebreaker software alerts you if all connectivity between the sites is lost. The following types of network paths are used by MetroCluster and monitored by MetroCluster Tiebreaker:

- **FC networks.** This type of network is composed of two redundant FC switch fabrics. Each switch fabric has two FC switches, with one switch of each switch fabric co-located with a cluster. Each cluster has two FC switches, one from each switch fabric. All the nodes have FC (NV interconnect and FCP initiator) connectivity to each of the co-located FC switches. Data is replicated from cluster to cluster over the ISL.

- **Intercluster peering networks.** This type of network is composed of a redundant IP network path between the two clusters. The cluster peering network provides the connectivity that is required to mirror the SVM configuration. The configuration of all the SVMs on one cluster is mirrored by the partner cluster.
- **IP network.** This type of network is composed of two redundant IP switch networks. Each network has two IP switches, with one switch of each switch fabric co-located with a cluster. Each cluster has two IP switches, one from each switch fabric. All the nodes have connectivity to each of the co-located FC switches. Data is replicated from cluster to cluster over the ISL.

Monitoring Intersite Connectivity

The Tiebreaker software regularly retrieves the status of intersite connectivity from the nodes. If NV interconnect connectivity is lost and the intercluster peering does not respond to pings, then the clusters assume that the sites are isolated, and the Tiebreaker software triggers an “AllLinksSevered” alert. If a cluster identifies the “AllLinksSevered” status and the other cluster is not reachable through the network, then the Tiebreaker software triggers a “disaster” alert.

Components Monitored by Tiebreaker

The Tiebreaker software monitors each controller in the MetroCluster configuration by establishing redundant connections through multiple paths to a node management LIF and to the cluster management LIF, both hosted on the IP network.

The Tiebreaker software monitors the following components in the MetroCluster configuration:

- Nodes through local node interfaces
- The cluster through the cluster-designated interfaces
- The surviving cluster to evaluate whether it has connectivity to the disaster site (NV interconnect, storage, and intercluster peering)

When there is a loss of connection between the Tiebreaker software and all the nodes in the cluster and to the cluster itself, the cluster is declared to be “not reachable” by the Tiebreaker software. It takes around three to five seconds to detect a connection failure. If a cluster is unreachable from the Tiebreaker software, the surviving cluster (the cluster that is still reachable) must indicate that all of the links to the partner cluster are severed before the Tiebreaker software triggers an alert.

Note: All the links are severed if the surviving cluster can no longer communicate with the cluster at the disaster site through FC (NV interconnect and storage) and intercluster peering.

Tiebreaker Failure Scenarios

The Tiebreaker software triggers an alert when the cluster (all the nodes) at the disaster site is down or unreachable and the cluster at the surviving site indicates the “AllLinksSevered” status.

The Tiebreaker software does not trigger an alert (or the alert is vetoed) in any of the following scenarios:

- In an eight-node MetroCluster configuration, if one HA pair at the disaster site is down.
- In a cluster with all the nodes at the disaster site down, one HA pair at the surviving site down, and the cluster at the surviving site indicates the “AllLinksSevered” status. The Tiebreaker software triggers an alert, but ONTAP vetoes that alert. In this situation, a manual switchover is also vetoed.
- Any scenario in which either the Tiebreaker software can reach at least one node or the cluster interface at the disaster site or the surviving site can still reach either node at the disaster site through either FC (NV interconnect and storage) or intercluster peering.

6 Conclusion

The various deployment options for MetroCluster, including support for both FC and IP fabrics, provide the most flexibility, a high level of data protection, and seamless front-end integration for all protocols, applications, and virtualized environments.

Where to Find Additional Information

To learn more about the information described in this document, refer to the following documents and/or websites:

- TR-4375: NetApp MetroCluster FC
  http://www.netapp.com/us/media/tr-4375.pdf
- 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 IP 40Gb Switch Technical (NetApp Field Portal; login required)
  https://fieldportal.netapp.com/content/729700
- MetroCluster IP 100Gb Switch Technical (NetApp Field Portal; login required)
  https://fieldportal.netapp.com/content/757495
- MetroCluster FC Technical FAQ (NetApp Field Portal; login required)
  https://fieldportal.netapp.com/content/617080
- MetroCluster IP Technical FAQ (NetApp Field Portal; login required)
  https://fieldportal.netapp.com/content/748972
- 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 Hardware Universe
  https://hwu.netapp.com/
- NetApp MetroCluster Product Documentation
  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=30022
- 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

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

Trademark Information

TR-4705-1119