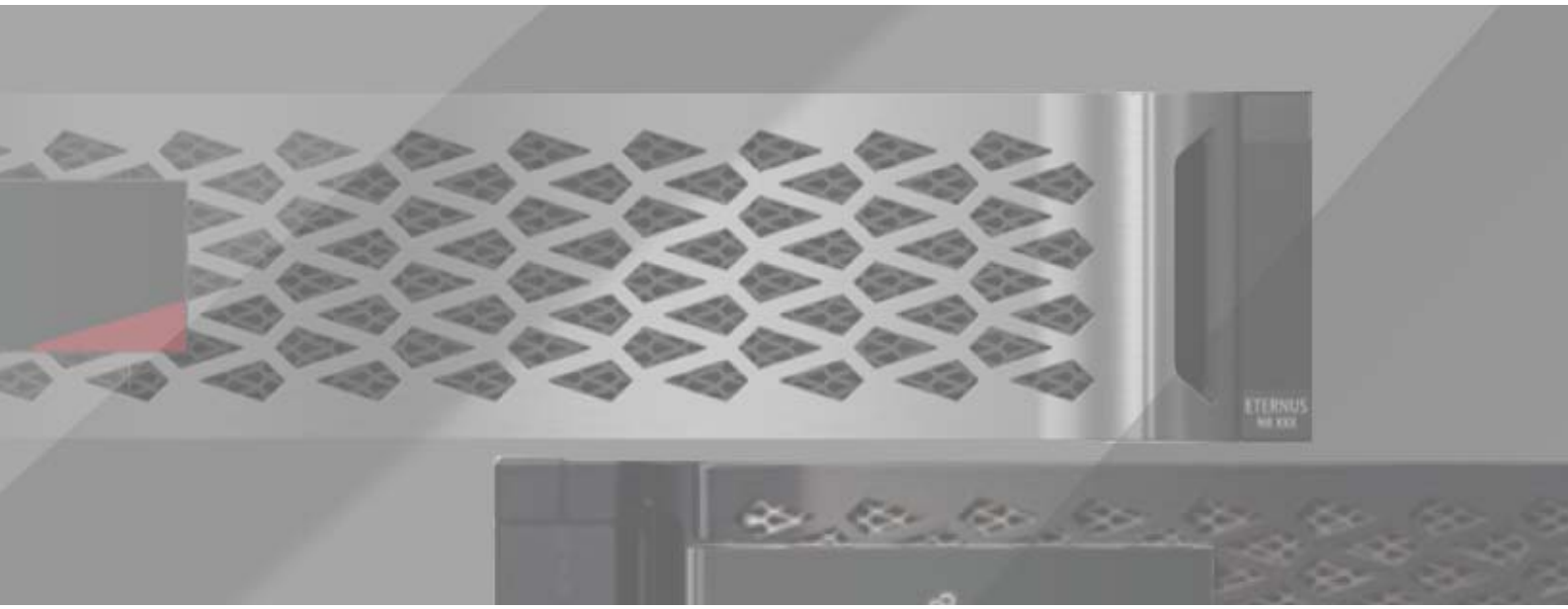
C140-0043-01ENZ3

Fujitsu Storage
ETERNUS AX series All-Flash Arrays,
ETERNUS HX series Hybrid Arrays

MetroCluster
Solution Architecture and Design



FUJITSU

Table of Contents

List of Figures

List of Tables

Preface

This document describes high-level architecture and design concepts for MetroCluster features in ONTAP 9.12.1 storage management software.

First Edition
June 2023

Trademarks

Third-party trademark information related to this product is available at:
https://www.fujitsu.com/global/products/computing/storage/eternus/trademarks.html

Trademark symbols such as ™ and ® are omitted in this document.

About This Manual

Intended Audience

This manual is intended for system administrators who configure and manage operations of the ETERNUS AX/HX, or field engineers who perform maintenance. Refer to this manual as required.

Related Information and Documents

The latest information for the ETERNUS AX/HX is available at:
https://www.fujitsu.com/global/support/products/computing/storage/manuals-list.html

Document Conventions

Notice Symbols

The following notice symbols are used in this manual:

Caution	Indicates information that you need to observe when using the ETERNUS AX/HX. Make sure to read the information.
Note	Indicates information and suggestions that supplement the descriptions included in this manual.

1. MetroCluster Overview

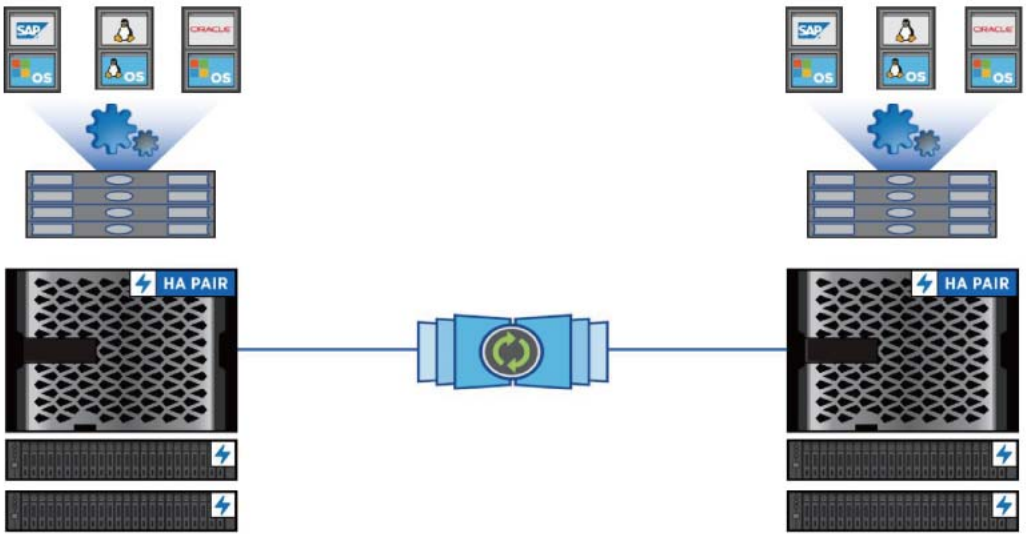
Enterprise-class customers must meet increasing service-level demands while maintaining cost and operational efficiency. As data volumes proliferate and more applications move to shared virtual infrastructures, the need for continuous availability for both mission-critical and other business applications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

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.

MetroCluster software is a solution that combines array-based clustering with synchronous replication to deliver continuous availability and zero data loss. Administration of the array-based cluster is simpler because the dependencies and complexities 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 traditional 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 a 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



MetroCluster enhances the built-in high-availability (HA) and nondisruptive operations of 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 site 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:

- Strictest SLAs (RPO(recovery point objective)=0 and RTO(recovery time objective)<2 minutes) achieved through synchronous replication and seamless storage promotion to applications.
- Multiprotocol support for a wide range of SAN and NAS client and host protocols.
- Synchronous replication supported over IP networks.
- No charge for MetroCluster functionality.
- Mirror only critical data: support for mirrored and unmirrored aggregates.
- Easy import of third party storage with Foreign LUN Import (FLI).
- Storage and network efficiencies achieved from deduplication, compression, and compaction.
- Integration with SnapMirror technology to support asynchronous replication, distance, and SLA requirements.

Data Protection with 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 operation has completed 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 an RPO of zero (meaning zero data loss), we must think about failure scenarios. We must determine the expected result when replication is impossible due to loss of connectivity between sites.

True HA Data Center with MetroCluster

MetroCluster replication is based on SyncMirror technology which provides synchronous mirroring of data, implemented at the RAID level. You can use SyncMirror to create aggregates that consist of two copies of the same WAFL file system. The two copies, known as plexes, are simultaneously updated and are always identical. This technology meets the requirements of most customers who demand synchronous replication under normal conditions.

Note

In cases of a partial failure that severs all connectivity between sites, the storage system can continue operating but in a nonreplicated state.

MetroCluster is ideal for organizations that require 24/7 operation for critical business applications. By synchronously replicating data between ETERNUS AX/HX series 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.

Campus, Metro, and Regional Protection

MetroCluster can also significantly simplify the design, deployment, and maintenance of campus wide or metropolitan wide HA solutions, with validated 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.

Your Choice of Protection

Achieve new levels of flexibility and choice for business continuity. When deployed with ONTAP 9 software, MetroCluster can scale from a four-node to an eight-node cluster (four nodes on each end of the replication), even with a mix of ETERNUS AX/HX series systems. Scaling up from a four-node to eight-node configuration is a non-disruptive process. 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.

Synchronous replication over an IP network is supported with four-node and eight-node configurations.

WAN-based DR

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

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.

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 (NAS 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

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.

MetroCluster Physical Architecture

MetroCluster configurations protect data by using two distinct clusters that are separated by a distance 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 ONTAP, a MetroCluster four-node configuration consists of a two-node HA pair at each site. This configuration allows most 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 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 IP connection that is used for storage and NVRAM synchronous replication between the two clusters. For MetroCluster IP, replication uses both iWARP for NVRAM and iSCSI for drive replication. All storage is visible to all controllers through the shared storage fabric.

> **Note**
>
> iWARP (Internet Wide Area RDMA Protocol) is a networking protocol that enables Remote Direct Memory Access (RDMA) over Ethernet Networks. It allows for high-speed, low-latency data transfers between servers, storage systems, and other networked devices, while reducing overhead associated with traditional network communication protocols.

MetroCluster IP Functions

MetroCluster IP uses Ethernet/IP ISLs for the fabric. Additionally, MetroCluster 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 (Layer 2 - shared VLAN, Layer 3 - VIP/BGP). Starting in ONTAP 9.7, MetroCluster IP is offered without dedicated switches, allowing the use of existing switches if they are compliant with the requirements for MetroCluster IP. For more information see the "ETERNUS AX/HX series MetroCluster IP Installation and Configuration Guide".

Table 1 indicates how data is replicated between the two MetroCluster sites.

Table 1 MetroCluster IP functions

Function	Details
MetroCluster fabric	Ethernet/IP ISLs
Fabric fibre switches	None
SAS bridges	None
Fabric Ethernet switches	Two per site
Fabric 25G/40G/100G Ethernet adapters	One per node depending on platform. Adapter is used to replication both iWARP and iSCSI
Intercluster	Switched
Shelves	Not visible to remote clusters
NVRAM replication	IP/iWARP
SyncMirror replication	IP/iSCSI
Configuration replication services	No changes
MetroCluster size	Four and eight nodes
MetroCluster stretch	No
Advanced Drive Partitioning	Yes, for the ETRENUS AX series only

Local Failover (HA) and Remote Switchover (DR)

In a two-node architecture, both HA failover and remote DR are accomplished by using MetroCluster switchover and switchback functionality. Each node acts as both the HA partner and a DR partner for its peer. NVRAM is replicated to the remote partner, like a four-node configuration.

The four-node and eight-node architectures provide 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 2. A1 and A2 are HA partners, as are B1 and B2. Node A1 and B1 are DR partners, as are A2 and B2. NVRAM is replicated to both the HA and the DR partner, as explained further in "NVRAM Replication" (page 17). The DR partner for a node is automatically selected when MetroCluster is configured, and the partner is chosen according to a system ID (NVRAM ID) order.

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 2 HA and DR groups

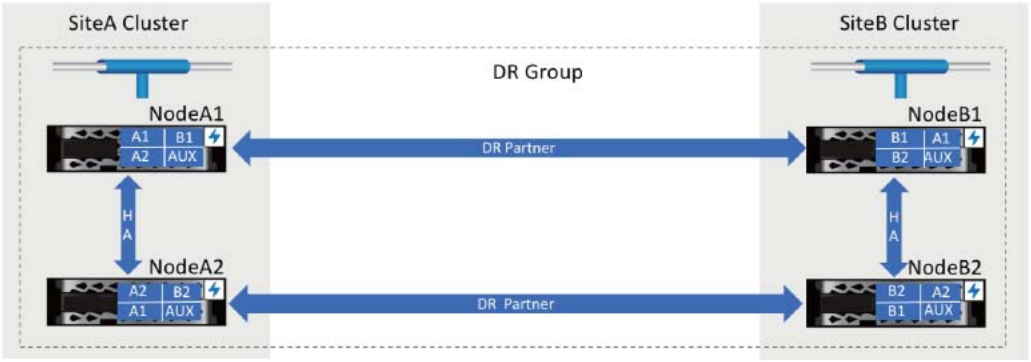
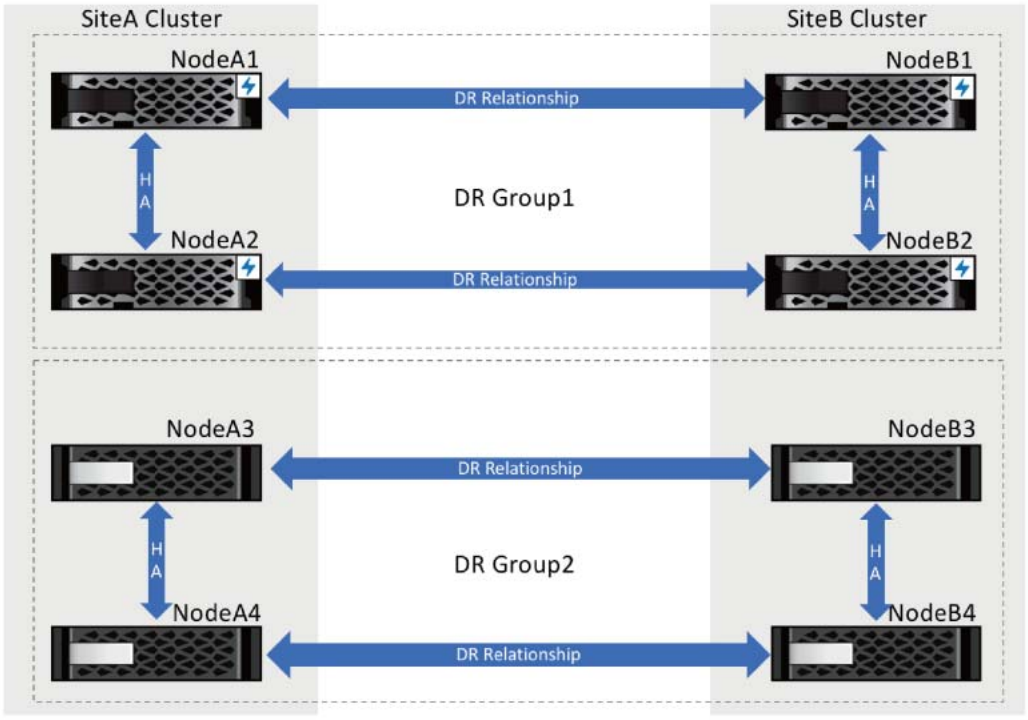


Figure 3 depicts an eight-node MetroCluster configuration and the DR group relationships. In an eight-node deployment, there are two independent DR groups. The hardware within a DR group must be the same in Site A and Site B. However, the hardware in DR Group 1 does not have to match the hardware in DR Group 2. For example, the hardware can be different in each DR group.

Figure 3 8-node DR group



In a local HA failover, one of the nodes in the HA pair temporarily takes over the shared 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.

During a switchover, the peer cluster takes over the storage and services of the other cluster while still performing its own workloads. For instance, when site A switches over to site B, the nodes of cluster B temporarily assume control of the storage and services previously owned by cluster A. Once the switchover is completed, the SVMs from cluster A are brought back online and can continue to run 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.

MetroCluster Replication

MetroCluster IP leverages direct attached storage, which eliminates the need for external serial-attached SCSI (SAS) bridges to connect drives to the storage fabric. Each node in the disaster recovery group acts as a storage proxy or iSCSI target that exports its drives to the other nodes in the group. iSCSI (SCSI over TCP/IP) is the storage transport protocol for the IP fabric that allows the iSCSI initiator and targets to communicate over a TCP/IP fabric. Each node in the disaster recovery group accesses its remote storage through an iSCSI initiator that establishes an iSCSI session with a remote disaster recovery partner iSCSI target.

The use of iSCSI and direct-attached storage also enables the use of systems that have internal drives. iSCSI allows the nodes to provide the disaster recovery partner node access to internal storage in addition to storage devices located in external drive shelves.

MetroCluster has three planes of replication:

(1) Configuration replication

(2) NVRAM replication

(3) Storage replication

Configuration Replication

MetroCluster configurations consists of two ONTAP clusters, each with its own replicated database (RDB) that contains its own metadata or configuration information. When a switchover occurs, the stopped cluster's metadata objects are activated on the surviving cluster, which requires the transfer of these objects from the owning cluster to the other cluster. The transfer mechanism has three components: cluster peering, configuration replication service (CRS), and a volume that contains metadata.

- **Cluster peering** is a method of establishing a customer-supplied TCP/IP connection between two ONTAP clusters using intercluster logical interfaces (LIFs). It enables the replication of configuration objects between the clusters and is used in MetroCluster and ONTAP SnapMirror software. The cluster peering network is typically the front-end or host-side network, and it transfers objects such as storage virtual machines (SVMs), LIFs, volumes, aggregates, and LUNs. The peering network for MetroCluster is the same as a regular ONTAP cluster, and it can also be the same front-end network used by hosts to access storage. The replication is conducted over the peering network, which is a customer-supplied IP network with intercluster LIFs.
- The **Configuration Replication Service (CRS)** is a component of a MetroCluster configuration that runs on each cluster and is responsible for replicating the required metadata objects from the owning cluster to the peered cluster's replicated database (RDB). This service replicates configuration objects (e.g., SVMs, LIFs, volumes, aggregates, LUNs) and protocol objects (e.g., CIFS, NFS, SAN) between the clusters using the peering network. If there is an interruption in the cluster peering network that affects CRS, replication catches up automatically after the connection is re-established. The CRS requires a small volume on a data aggregate to store metadata referred to as the metadata volume.
- Volumes that contain metadata are staging volumes used for cluster metadata information in a MetroCluster configuration. When MetroCluster is configured, two volumes, each 10GB in size, are created on each cluster. These volumes must be created on separate non-root aggregates, so at least two data aggregates are recommended on each cluster before configuring MetroCluster. The volumes that contain metadata provide resiliency, and updates are logged in them whenever an object is created or updated. Changes are not committed to the local RDB until the logging is complete. Updates are propagated 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 to the configuration of one cluster automatically propagate to the other cluster so that switchover is achieved with zero data or configuration loss. The update is automatic, and almost no ongoing administration is required that is specific to a MetroCluster configuration. If changes cannot be propagated due to temporary errors in the configuration replication network, the changes are automatically sent to the other cluster after connectivity is restored. To promote resiliency, redundant networks are recommended for the cluster configuration network.

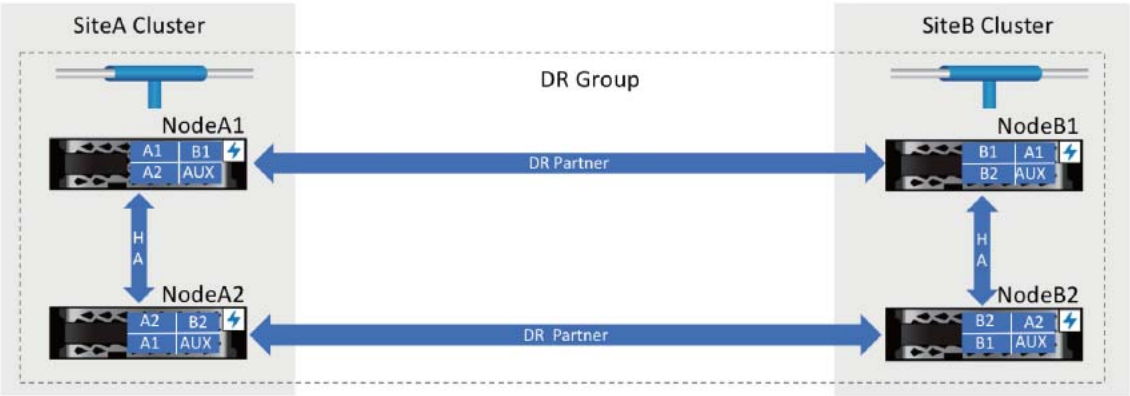
In the example below, you can see that the MDVs have been assigned system-assigned names and are visible on each cluster. The first two volumes listed are the local MDVs with the state of "online" because the command was issued from cluster A. The remaining two MDVs belong to cluster B, as indicated by their hosting aggregate, and are currently 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            -          -     -
```

NVRAM Replication

NVRAM replication is a process of copying the local node's NVRAM to the NVRAM of the remote disaster recovery node to protect against data loss in the event of a failover or switchover. In an ONTAP HA pair, each node mirrors its NVRAM to the other node via the HA interconnect. The NVRAM is divided into two segments, one for each node's NVRAM. Figure 4 shows MetroCluster provides additional mirroring by having a DR partner node on the other site, and the NVRAM is mirrored to the DR partner via the
Inter-Switch Link (ISL) connection. In a four-node configuration, each node's NVRAM is mirrored twice, once to the HA partner and once to the DR partner, and each node's NVRAM is split into four segments.

Figure 4 NVRAM allocation



Write operations are first staged to nonvolatile memory (NVRAM) before being written to drive, and acknowledgement is sent to the issuing host or application only after all NVRAM segments have been updated. The NVRAM on each storage controller is mirrored both locally to a local high-availability (HA) partner and remotely to a disaster recovery (DR) partner on the partner site. In a four-node configuration, the nonvolatile cache is split into four partitions for the local, HA partner, DR partner, and DR auxiliary partner. In the event of a local HA takeover, DR mirroring can continue by automatically switching to the DR auxiliary partner. Once a successful giveback is completed, mirroring will automatically return to the DR partner. To illustrate, if NodeB1 fails and is taken over by NodeB2, the local cache of NodeA1 cannot be mirrored to NodeB1, and as a result, mirroring will switch to the DR auxiliary partner, NodeB2.

Updates to the DR partner's NVRAM are transmitted over the ISL using the iWARP protocol for MetroCluster IP. For MetroCluster IP, iWARP is offloaded to hardware using RDMA-capable network adapters to minimize latency from being affected by the IP stack. Switch quality of service (QoS) is used to prioritize iWARP traffic over storage replication.
However, if the ISL latency increases, write performance might be affected, as it takes longer to acknowledge the write to the DR partner's NVRAM. To allow continued local operation in the event of temporary site isolation (e.g., all ISLs down, remote node not responding), writes are acknowledged after a system timeout. The remote NVRAM mirror resynchronizes automatically when at least one ISL becomes available.

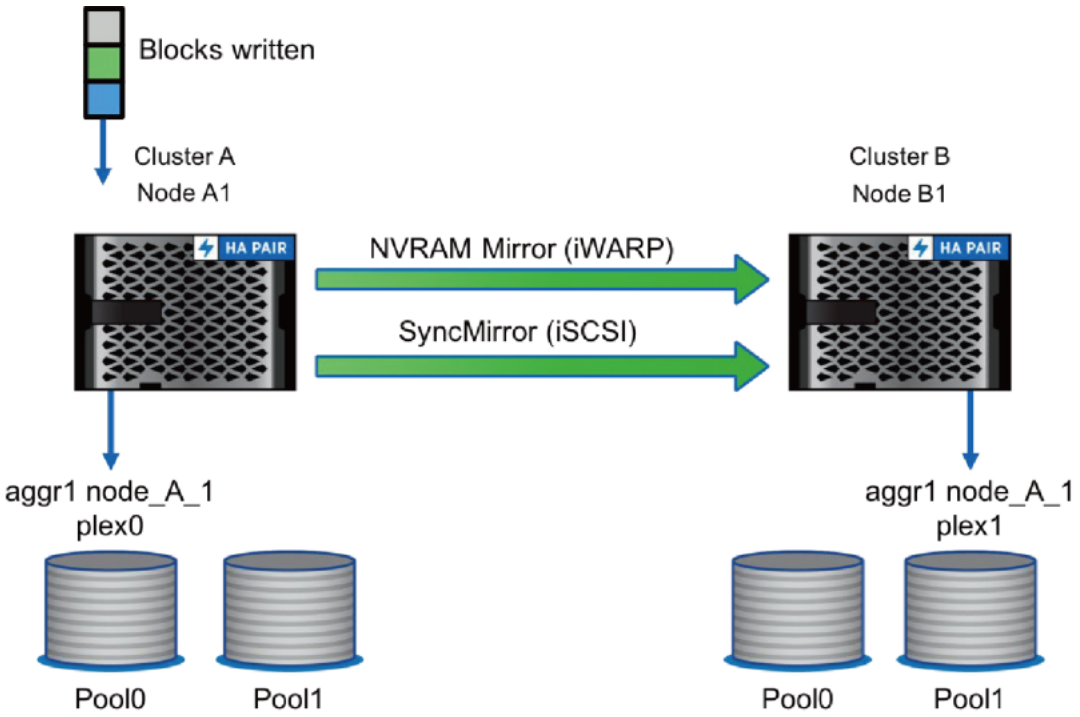
To prevent data loss, NVRAM transactions are committed to drive through a consistency point at least once every 10 seconds. Upon controller boot, WAFL uses the most recent consistency point on drive, eliminating the need for lengthy file system checks after a power loss or system failure. The storage 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.

Storage Replication

Storage replication mirrors the local and remote back-end drives using RAID SyncMirror (RSM). MetroCluster IP presents the back-end storage as logically shared by making each node in a disaster recovery group serve as a remote iSCSI target. For a node to access its remote back-end drives, it goes through its remote disaster recovery partner node to access the remote drives that are served through an iSCSI target.

Figure 5 illustrates the MetroCluster IP planes of replication for NVRAM and storage. NodeB1 exports its locally attached drives to remote partner nodes in the disaster recovery group through an iSCSI target.
NodeA1 pool0 drives are locally attached to NodeA1, whereas pool1 remote drives are exported through the iSCSI target hosted by B1. The aggregate `aggr1 node_A_1` local `plex 0` consists of locally attached drives from pool0. The aggregate `aggr1 node_A_1` remote `plex 1` consists of drives directly attached to B1 and exported to A1 through the iSCSI target hosted in B1.

Figure 5 Mirroring write data blocks



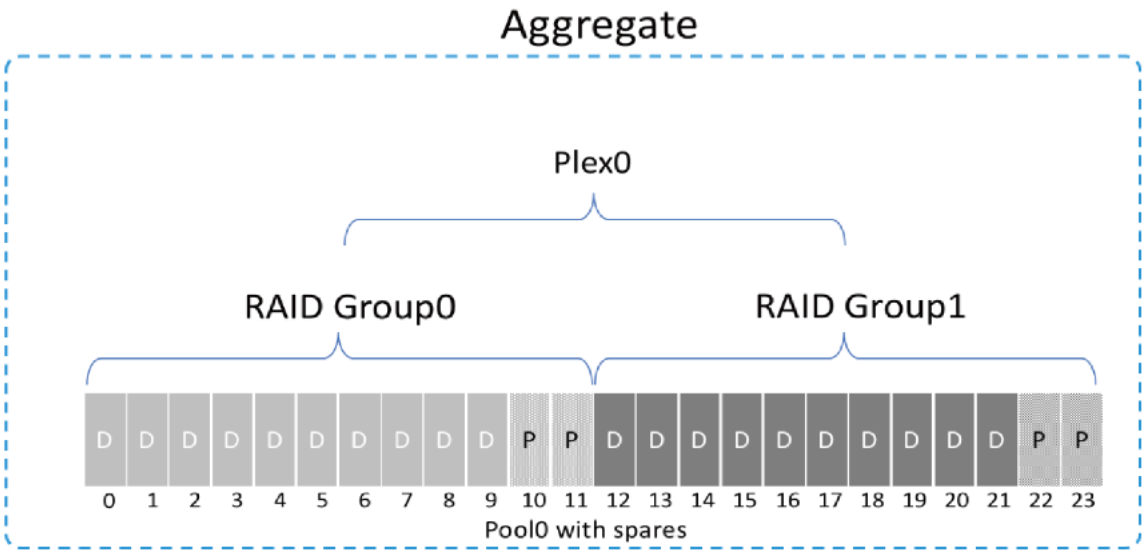
Blocks are written to the paired nodes at each site with both NVRAM (or NVMEM) and SyncMirror. SyncMirror writes data to two plexes for each mirrored aggregate, one local plex and one remote plex. SyncMirror writes occur in the RAID layer, which means that any storage efficiencies such as deduplication and compression reduce the data written by the SyncMirror operations.

Blocks are read from the local storage and do not affect performance or use of the ISLs for read operations.

SyncMirror Storage Replication

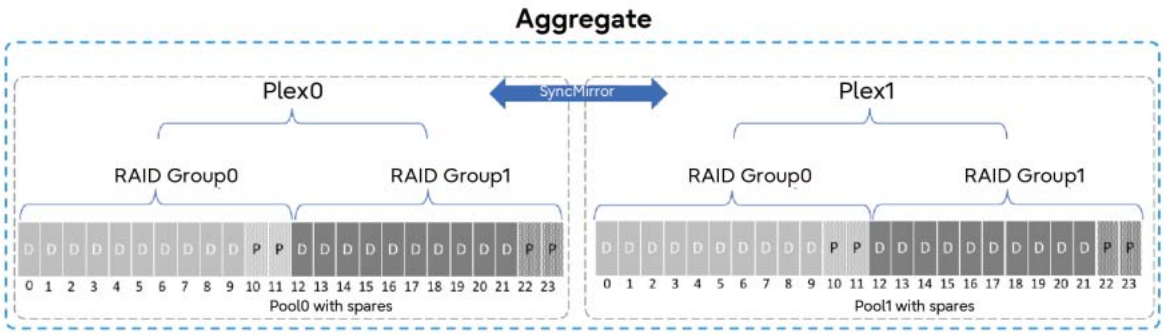
An ONTAP system stores data in FlexVol volumes that are provisioned from aggregates. Each aggregate contains a WAFL file system. In a configuration without MetroCluster, the drives in each aggregate consist of a single or multiple RAID group, known as a plex (Figure 6). The plex resides in local storage attached to the controller.

Figure 6 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 7). All storage is shared and is visible to all the controllers in the MetroCluster configuration. The local plex must contain only drives from the local pool (pool0), and the remote plex must contain only drives 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).

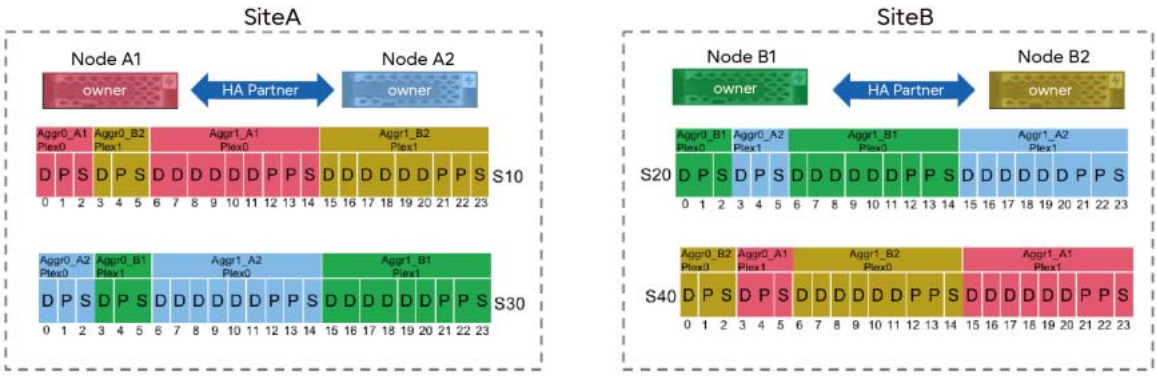
Figure 7 MetroCluster mirrored aggregate



Both mirrored and unmirrored aggregates are supported with MetroCluster. Unmirrored aggregates are supported with MetroCluster IP starting with ONTAP 9.8. The `-mirror true` flag must be used when creating aggregates after MetroCluster has been configured; if it is not specified, the `create` command fails. The number of drives that are specified by the `-diskcount` parameter is automatically halved. For example, to create an aggregate with six usable drives, 12 must be specified as the drive count. That way, the local plex is allocated six drives from the local pool, and the remote plex is allocated six drives from the remote pool. The same process applies when adding drives to an aggregate; twice the number of drives must be specified as are required for capacity.

The example in Figure 8 shows how the drives 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 drives and nine remote drives. With MetroCluster and particularly with the ETRENUS AX series, the root aggregate uses RAID 4, and data aggregates use RAID DP or RAID-TEC.

Figure 8 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 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.

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 a 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. 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 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. To achieve the IP address/service mobility requirement for host networking, MetroCluster supports both layer 2 (shared VLAN) and layer 3 (VIP/BGP) networking.

Advanced Drive Partitioning (ADP)

Note

For MetroCluster, ADP is only available on ETERNUS AX series systems in MetroCluster IP configurations.

ADP is a feature that enhances the storage efficiency of both HDDs and SSDs on ETERNUS AX/HX series systems. By enabling capacity sharing of physical drives between aggregates and controllers within a HA pair, ADP allows for expanding the capacity of both nodes with fewer SSDs, thereby improving efficiency and price. With ADP, more usable capacity is available for provisioning data aggregates, as less capacity is consumed by root aggregates. In addition, the sharing of parity and spare drives between both controllers further increases the capacity available within the HA pair. ADP is a more efficient method of provisioning storage capacity than using whole, unpartitioned drives.

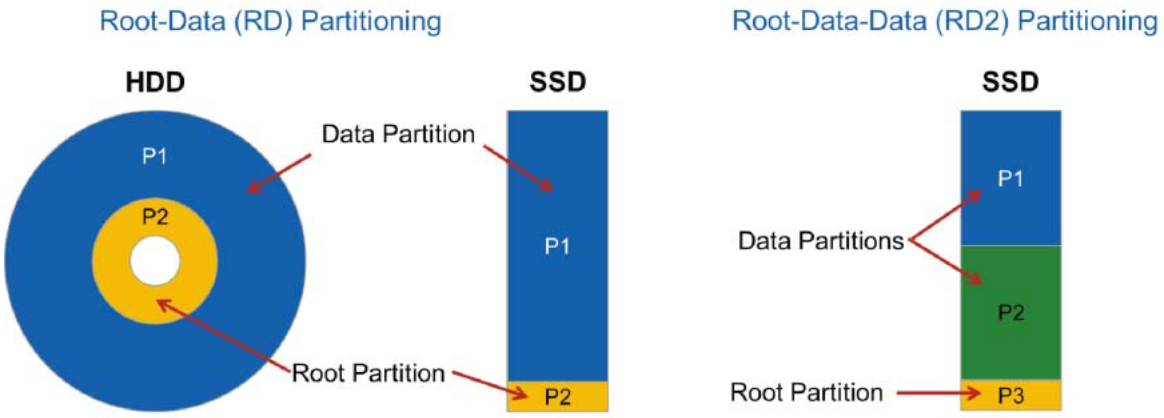
Benefits of Advanced ADP:

- Increases usable and effective capacity on ETERNUS AX/HX series systems.
- Improves storage efficiency (10-40% efficiency gain versus whole drive partitioning).
- Allows for expanding capacity to both nodes with fewer SSDs.
- Improves price and effective storage competitiveness.

Methods of partitioning:

- Root-data (RD) partitioning: Divides a drive into one root partition and one data partition.
- Root-data-data (RD2) partitioning: Divides a drive into one root partition and two data partitions.

Figure 9 Logical View of ADP methods



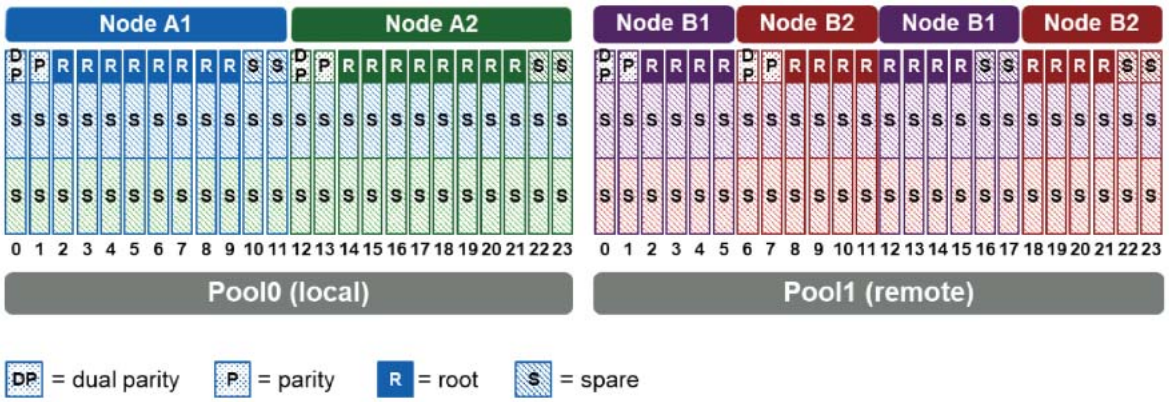
Note

Only ADP RD2 partitioning is supported on MetroCluster IP configurations on ETERNUS AX series systems with ONTAP 9.7 or later. ADP is applied by default at the time of MetroCluster initialization.

Root-Data-Data (RD2) Partitioning

Root-data-data (RD2) partitioning is a storage efficiency feature available in ONTAP 9.0 and later releases. RD2 efficiently provisions root aggregates by using partitions from multiple SSDs, resulting in more usable capacity for data aggregates. Each SSD is divided into 3 partitions: a smaller (thin) root partition and 2 larger (thick) data partitions. Having two data partitions per SSD enables the capacity and IOPS of a single drive to be used by both controllers in an ETRENUS AX series or all-SSD ETRENUS HX series system. The maximum number of SSDs that can be RD2 partitioned is 48, but more than 48 SSDs can use RD2 partitioning in an HA pair. The minimum number of SSDs required to use RD2 partitioning is 8, and 400GB SSDs or larger support RD2 partitioning. The root partition sizes in a system with RD2 partitioning vary according to controller model, ONTAP release, and the number of SSDs attached during system initialization. Spare root partitions can only be used to expand the root aggregate if additional space is required, while spare data partitions can be used as hot spares to replace failed data partitions. Figure 10 depicts an ADP example for a MetroCluster IP configuration with 1x AX2200 with 1x NS224 per site with 48x NVMe SSDs.

Figure 10 ADP example for 48 drive MetroCluster IP configuration

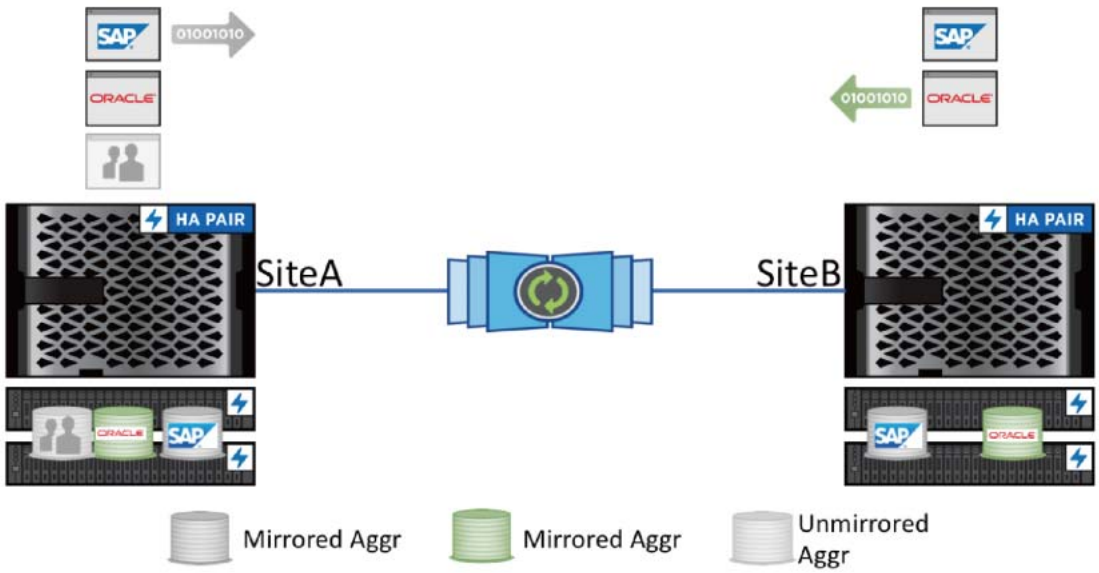


Unmirrored Aggregates

Beginning with ONTAP 9 for stretch and fabric-attached configurations and ONTAP 9.8 for IP configurations, MetroCluster supports unmirrored 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.

Figure 11 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 11 Unmirrored aggregates in MetroCluster



Note

When using ADP configured drives, it is critical to understand the specific rules regarding partition ownership by nodes and membership within pools. In addition, drives that are intended to be mirrored must be symmetric on both sides of the peering relationship. Using ADP configured partitions for unmirrored aggregates can lead to unintended and unpredictable failures. To avoid such issues, it is strongly recommended to deploy unmirrored aggregates on unpartitioned drives.

3. Deployment Options

MetroCluster is a fully redundant configuration with identical hardware required at each site. Additionally, MetroCluster offers flexibility of stretch, fabric-attached and IP 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
		Four-node or eight-node
Number of controllers	Four or eight	Four or eight
FC switch storage fabric	No	Yes
IP switch storage fabric	Yes	No
FC-to-SAS bridges	No	Yes
Direct- attached storage	Yes (local attached only)	No
Supports local HA	Yes	Yes
Supports automatic switchover	Yes (with mediator)	Yes
Supports unmirrored aggregates	Yes	Yes
Supports array LUNs	No	Yes

Stretch and Stretch-bridged Configurations

MetroCluster stretch configuration extends two storage nodes over a larger distance, typically up to 270m apart, and provides an elevated level of resiliency and disaster recovery capabilities. This is achieved by connecting the two clusters using high-speed links and synchronously mirroring the data between them. If one cluster fails or becomes unavailable, the other cluster takes over seamlessly, providing continuous data access to applications and users. MetroCluster stretch-bridged configurations extend the stretch configuration further, up to 500m, by adding FC-to-SAS bridges between the two primary clusters. The stretch-bridged configuration can be used to support more complex architectures or to span larger distances between data centers. Both of these configurations are ideal for data center deployments and have reduced infrastructure (e.g. cabling, FC switches, rack space) demands.

IP Configuration

MetroCluster IP uses IP networks for synchronous replication between two sites, up to 700km. Supported on ETERNUS AX/HX series systems, MetroCluster IP can be deployed in four- and eight-node architectures.

Note

For MetroCluster's latest supported hardware, software, sizing and limits please review Fusion.

4. Resiliency for Planned and Unplanned Events

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

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's storage resources, depending on the component that failed. Data recovery is described in the "ETERNUS AX/HX Series High-Availability Configuration Guide".

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 "ETERNUS AX/HX Series 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.

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.

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 3 describes failure types and the corresponding DR mechanism and recovery method.
AUSO is only supported on MetroCluster IP configurations when using ONTAP Mediator and ONTAP 9.7 or later.

Table 3 Failure types and recovery methods

Failure type	DR mechanism	Summary of recovery methods
Single-node failure	Local HA failure	Not required if automatic failover and giveback are enabled.
Site failure	MetroCluster switchover	After the node is restored, manual healing and switchback using the `metrocluster healing` and `metrocluster switchback` commands are required.
Sitewide controller failure	AUSO Only if the storage at the disaster site is accessible.	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. Depending on the component that failed, a forced switchover might not be required.	After the node is restored, manual healing and switchback using the `metrocluster healing` and `metrocluster switchback` commands are required.

Four-node and Eight-node Nondisruptive Operations

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

Because these 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.
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.

Should a local failover happen following a switchover, there is a risk of resource-related problems arising since only a single controller is responsible for handling data across all storage systems within the MetroCluster setup. The surviving controller is vulnerable to additional failures.

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 cluster. 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 drives that belong to the disaster site is changed to the DR partner. This situation is like the case of a local failover in an HA pair in which ownership of the drives 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.
  - 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 drives power up in a staggered manner.

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

MetroCluster Tiebreaker

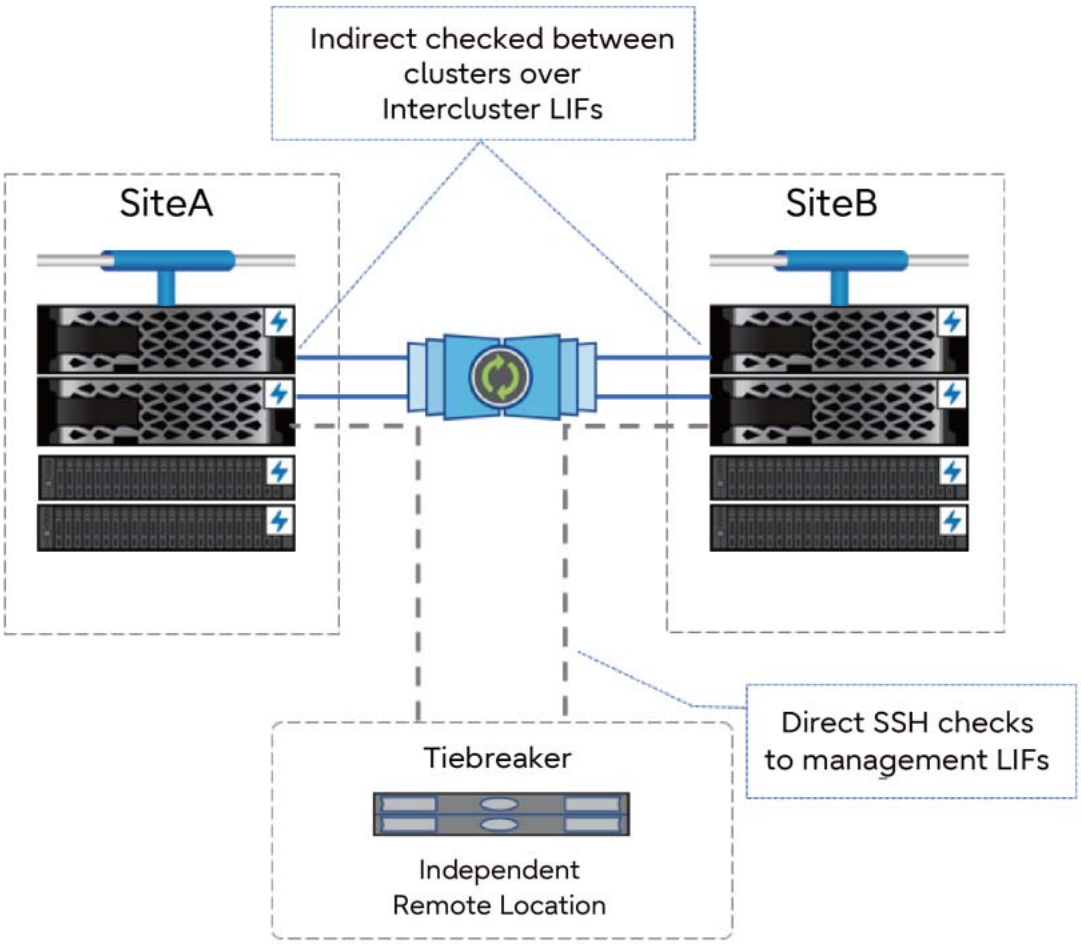
The MetroCluster Tiebreaker software alerts you if all connectivity between the sites is lost. The MetroCluster Tiebreaker software supports all the MetroCluster configurations that are supported in ONTAP 9.7 and later.

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.

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

The 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 12, so that the Tiebreaker does not initiate a switchover if the fabric is intact.

Figure 12 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:

- **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 the links to the partner cluster are severed before the Tiebreaker software triggers an alert.

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.

ONTAP Mediator

ONTAP 9.7 or later 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 the "ETERNUS AX/HX series 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. Technology Requirements

Hardware & Software Requirements

When configuring your MCC, it is important to carefully consider the hardware and software components that are supported. The specific hardware components used may vary depending on the customer's deployment. It is essential that the ONTAP systems, storage arrays, and FC switches used in MCC configurations meet the necessary requirements. The only required software component to implement MCC is ONTAP, which is a standard feature and does not require a separate license. Standard ONTAP licensing covers the client and host side protocols, as well as additional capabilities for SnapMirror to protect data using an asynchronous mirror to replicate data to a third cluster for backup data protection. For the latest information on hardware and software requirements, please consult the available technical resources.

For more information about the hardware and software requirements for MetroCluster IP, review the following document:

- ETERNUS AX/HX series MetroCluster IP Solution Architecture and Design

For the latest on hardware configuration and interoperability, please use the following tool.

- Fusion

6. Conclusion

The various deployment options for MetroCluster, including support for an IP fabric, provide the most flexibility, an elevated level of data protection, and seamless front-end integration for all protocols, applications, and virtualized environments.

Fujitsu Storage
ETERNUS AX series All-Flash Arrays,
ETERNUS HX series Hybrid Arrays
MetroCluster
Solution Architecture and Design

C140-0043-01ENZ3

Date of issuance: June 2023
Issuance responsibility: Fujitsu Limited


FUJITSU