Prepare for the MetroCluster installation

ONTAP MetroCluster

NetApp
February 13, 2026

Table of Contents

Prepare for the MetroCluster installation

ONTAP MetroCluster configurations support matrix

The various MetroCluster configurations have key differences in the required components.

In all configurations, each of the two MetroCluster sites are configured as an ONTAP cluster. In a two-node MetroCluster configuration, each node is configured as a single-node cluster.

Feature	**IP configurations**	**Fabric attached configurations**		**Stretch configurations**	
		Four- or eight-node	**Two-node**	**Two-node bridge-attached**	**Two-node direct-attached**
Number of controllers	Four or eight[1]	Four or eight	Two	Two	Two
Uses an FC switch storage fabric	No	Yes	Yes	No	No
Uses an IP switch storage fabric	Yes	No	No	No	No
Uses FC-to-SAS bridges	No	Yes	Yes	Yes	No
Uses direct-attached SAS storage	Yes (local attached only)	No	No	No	Yes
Supports ADP	Yes (beginning with ONTAP 9.4)	No	No	No	No
Supports local HA	Yes	Yes	No	No	No
Supports ONTAP automatic unplanned switchover (AUSO)	No	Yes	Yes	Yes	Yes

Supports unmirrored aggregates	Yes (beginning with ONTAP 9.8)	Yes	Yes	Yes	Yes
Supports ONTAP Mediator	Yes (beginning with ONTAP 9.7)	No	No	No	No
Supports MetroCluster Tiebreaker	Yes (not in combination with ONTAP Mediator)	Yes	Yes	Yes	Yes
Supports All SAN Arrays	Yes	Yes	Yes	Yes	Yes

Notes

1. Review the following considerations for eight-node MetroCluster IP configurations:
   - Eight-node configurations are supported beginning with ONTAP 9.9.1.
   - Only NetApp-validated MetroCluster switches (ordered from NetApp) are supported.
   - Configurations using IP-routed (layer 3) backend connections are not supported.

Support for All SAN Array systems in MetroCluster configurations

Some of the All SAN Arrays (ASAs) are supported in MetroCluster configurations. In the MetroCluster documentation, the information for AFF models applies to the corresponding ASA system. For example, all cabling and other information for the AFF A400 system also applies to the ASA AFF A400 system.

Supported platform configurations are listed in the NetApp Hardware Universe.

Differences between ONTAP Mediator and MetroCluster Tiebreaker

Beginning with ONTAP 9.7, you can use either the ONTAP Mediator-assisted automatic unplanned switchover (MAUSO) in the MetroCluster IP configuration or you can use the MetroCluster Tiebreaker software. It is not required to use the MAUSO or Tiebreaker software; however, if you choose to not use either of these services, you must perform a manual recovery if a disaster occurs.

The different MetroCluster configurations perform automatic switchover under different circumstances:

- **MetroCluster FC configurations using the AUSO capability (not present in MetroCluster IP configurations)**

  In these configurations, AUSO is initiated if controllers fail but the storage (and bridges, if present) remain operational.

- **MetroCluster IP configurations using ONTAP Mediator (ONTAP 9.7 and later)**

In these configurations, MAUSO is initiated in the same circumstances as AUSO, as described above, and also after a complete site failure (controllers, storage, and switches).

Learn about how the ONTAP Mediator supports automatic unplanned switchover.

- **MetroCluster IP or FC configurations using the Tiebreaker software in active mode**

  In these configurations, the Tiebreaker initiates unplanned switchover after a complete site failure.

  Before using the Tiebreaker software, review the MetroCluster Tiebreaker Software installation and configuration

Interoperability of ONTAP Mediator with other applications and appliances

You cannot use any third-party applications or appliances that can trigger a switchover in combination with ONTAP Mediator. In addition, monitoring a MetroCluster configuration with MetroCluster Tiebreaker software is not supported when using ONTAP Mediator.

Learn about remote storage and MetroCluster IP configurations

You should understand how the controllers access the remote storage and how the MetroCluster IP addresses work.

Access to remote storage in MetroCluster IP configurations

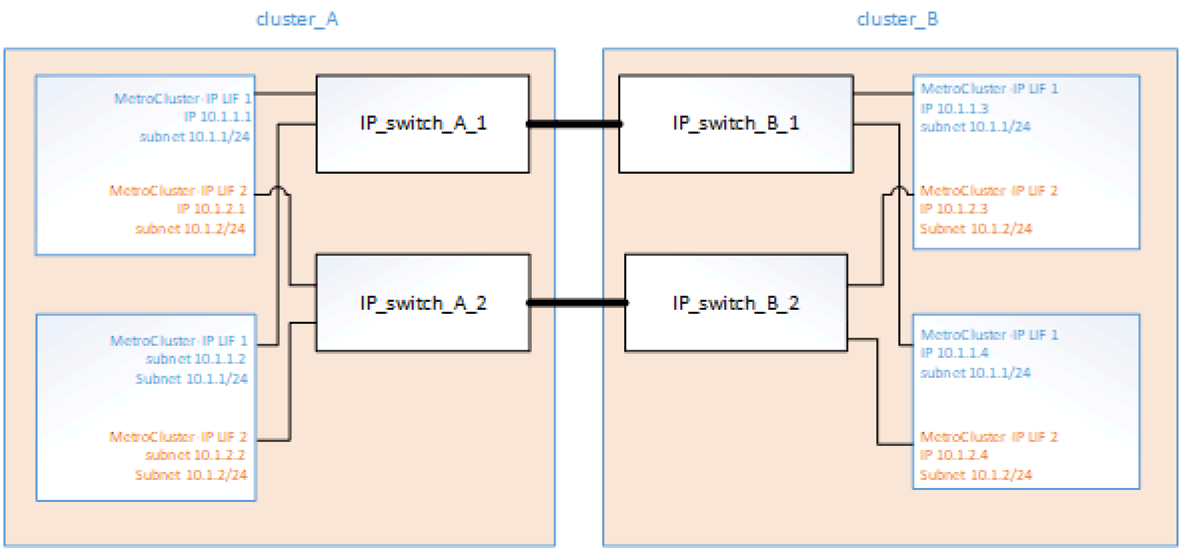
In MetroCluster IP configurations, the only way the local controllers can reach the remote storage pools is via the remote controllers. The IP switches are connected to the Ethernet ports on the controllers; they do not have direct connections to the disk shelves. If the remote controller is down, the local controllers cannot reach their remote storage pools.

This is different than MetroCluster FC configurations, in which the remote storage pools are connected to the local controllers via the FC fabric or the SAS connections. The local controllers still have access to the remote storage even if the remote controllers are down.

MetroCluster IP addresses

You should be aware of how the MetroCluster IP addresses and interfaces are implemented in a MetroCluster IP configuration, as well as the associated requirements.

In a MetroCluster IP configuration, replication of storage and nonvolatile cache between the HA pairs and the DR partners is performed over high-bandwidth dedicated links in the MetroCluster IP fabric. iSCSI connections are used for storage replication. The IP switches are also used for all intra-cluster traffic within the local clusters. The MetroCluster traffic is kept separate from the intra-cluster traffic by using separate IP subnets and VLANs. The MetroCluster IP fabric is distinct and different from the cluster peering network.



The MetroCluster IP configuration requires two IP addresses on each node that are reserved for the back-end MetroCluster IP fabric. The reserved IP addresses are assigned to MetroCluster IP logical interfaces (LIFs) during initial configuration, and have the following requirements:

> You must choose the MetroCluster IP addresses carefully because you cannot change them after initial configuration.

- They must fall in a unique IP range.

  They must not overlap with any IP space in the environment.

- They must reside in one of two IP subnets that separate them from all other traffic.

For example, the nodes might be configured with the following IP addresses:

Node	Interface	IP address	Subnet
node_A_1	MetroCluster IP interface 1	10.1.1.1	10.1.1/24
node_A_1	MetroCluster IP interface 2	10.1.2.1	10.1.2/24
node_A_2	MetroCluster IP interface 1	10.1.1.2	10.1.1/24
node_A_2	MetroCluster IP interface 2	10.1.2.2	10.1.2/24
node_B_1	MetroCluster IP interface 1	10.1.1.3	10.1.1/24
node_B_1	MetroCluster IP interface 2	10.1.2.3	10.1.2/24

node_B_2	MetroCluster IP interface 1	10.1.1.4	10.1.1/24
node_B_2	MetroCluster IP interface 2	10.1.2.4	10.1.2/24

Characteristics of MetroCluster IP interfaces

The MetroCluster IP interfaces are specific to MetroCluster IP configurations. They have different characteristics from other ONTAP interface types:

- They are created by the `metrocluster configuration-settings interface create` command as part the initial MetroCluster configuration.

  > ⓘ Beginning with ONTAP 9.9.1, if you are using a layer 3 configuration, you must also specify the `-gateway` parameter when creating MetroCluster IP interfaces. Refer to Considerations for layer 3 wide-area networks.

  They are not created or modified by the network interface commands.
- They do not appear in the output of the `network interface show` command.
- They do not fail over, but remain associated with the port on which they were created.
- MetroCluster IP configurations use specific Ethernet ports (depending on the platform) for the MetroCluster IP interfaces.

> ! Do not use 169.254.17.x or 169.254.18.x IP addresses when you create MetroCluster IP interfaces to avoid conflicts with system auto-generated interface IP addresses in the same range.

MetroCluster IP requirements for automatic drive assignment and ADP systems

Beginning with ONTAP 9.4, MetroCluster IP configurations support new installations using automatic disk assignment and ADP (Advanced Drive Partitioning).

You should be aware of the following when using ADP with MetroCluster IP configurations:

- ONTAP 9.4 and later is required to use ADP with MetroCluster IP configurations on AFF and ASA systems.
- ADPv2 is supported in MetroCluster IP configurations.
- The root aggregate must be located in Partition 3 for all nodes on both sites.
- Partitioning and disk assignment are performed automatically during the initial configuration of the MetroCluster sites.
- Pool 0 disk assignments are done at the factory.
- The unmirrored root is created at the factory.
- Data partition assignment is done at the customer site during the setup procedure.

- In most cases, drive assignment and partitioning is done automatically during the setup procedures.
- A disk and all of its partitions must be owned by nodes in the same high-availability (HA) pair. Partition or drive ownership within a single drive cannot be mixed between the local HA pair and the disaster recovery (DR) partner or DR auxiliary partner.

  Example of a supported configuration:

Drive/Partition	Owner
Drive:	`ClusterA-Node01`
Partition 1:	`ClusterA-Node01`
Partition 2:	`ClusterA-Node02`
Partition 3:	`ClusterA-Node01`

> When upgrading from ONTAP 9.4 to 9.5, the system recognizes the existing disk assignments.

Automatic partitioning

ADP is performed automatically during initial configuration of the system.

> Beginning with ONTAP 9.5, automatic assignment of disks must be enabled with the `storage disk option modify -autoassign on` command.

You must set the ha-config state to `mccip` before automatic provisioning to make sure that the correct partition sizes are selected to allow for appropriate root volume size. For more information, see Verifying the ha-config state of components.

A maximum of 96 drives can be automatically partitioned during installation. You can add extra drives after the initial installation.

> If you are using internal and external drives, you first initialize the MetroCluster with only the internal drives using ADP. You then manually connect the external shelf after you complete your installation or setup task.
>
> You must ensure that the internal shelves have the recommended minimum number of drives as outlined in ADP and disk assignment differences by system.
>
> For both the internal and external drives, you must populate the partially full shelves as described in How to populate partially-full shelves.

How shelf-by-shelf automatic assignment works

If there are four external shelves per site, each shelf is assigned to a different node and different pool, as shown in the following example:

- All of the disks on site_A-shelf_1 are automatically assigned to pool 0 of node_A_1
- All of the disks on site_A-shelf_3 are automatically assigned to pool 0 of node_A_2
- All of the disks on site_B-shelf_1 are automatically assigned to pool 0 of node_B_1

- All of the disks on site_B-shelf_3 are automatically assigned to pool 0 of node_B_2
- All of the disks on site_B-shelf_2 are automatically assigned to pool 1 of node_A_1
- All of the disks on site_B-shelf_4 are automatically assigned to pool 1 of node_A_2
- All of the disks on site_A-shelf_2 are automatically assigned to pool 1 of node_B_1
- All of the disks on site_A-shelf_4 are automatically assigned to pool 1 of node_B_2

How to populate partially-full shelves

If your configuration is using shelves that are not fully populated (have empty drive bays) you must distribute the drives evenly throughout the shelf, depending on the disk assignment policy. The disk assignment policy depends on how many shelves are at each MetroCluster site.

If you are using a single shelf at each site (or just the internal shelf on an AFF A800 system), disks are assigned using a quarter-shelf policy. If the shelf is not fully populated, install the drives equally on all quarters.

The following table shows an example of how to place 24 disks in a 48 drive internal shelf. The ownership for the drives is also shown.

The 48 drive bays are divided into four quarters:	Install six drives in the first six bays in each quarter…
Quarter 1: Bays 0-11	Bays 0-5
Quarter 2: Bays 12-23	Bays 12-17
Quarter 3: Bays 24-35	Bays 24-29
Quarter 4: Bays 36-47	Bays 36-41

The following table shows an example of how to place 16 disks in a 24 drive internal shelf.

The 24 drive bays are divided into four quarters:	Install four drives in the first four bays in each quarter…
Quarter 1: Bays 0-5	Bays 0-3
Quarter 2: Bays 6-11	Bays 6-9
Quarter 3: Bays 12-17	Bays 12-15
Quarter 4: Bays 18-23	Bays 18-21

If you are using two external shelves at each site, disks are assigned using a half-shelf policy. If the shelves are not fully populated, install the drives equally from either end of the shelf.

For example, if you are installing 12 drives in a 24-drive shelf, install drives in bays 0-5 and 18-23.

Manual drive assignment (ONTAP 9.5)

In ONTAP 9.5, manual drive assignment is required on systems with the following shelf configurations:

- Three external shelves per site.

  Two shelves are assigned automatically using a half-shelf assignment policy, but the third shelf must be assigned manually.

- More than four shelves per site and the total number of external shelves is not a multiple of four.

  Extra shelves above the nearest multiple of four are left unassigned and the drives must be assigned manually. For example, if there are five external shelves at the site, shelf five must be assigned manually.

You only need to manually assign a single drive on each unassigned shelf. The rest of the drives on the shelf are then automatically assigned.

Manual drive assignment (ONTAP 9.4)

In ONTAP 9.4, manual drive assignment is required on systems with the following shelf configurations:

- Fewer than four external shelves per site.

  The drives must be assigned manually to ensure symmetrical assignment of the drives, with each pool having an equal number of drives.

- More than four external shelves per site and the total number of external shelves is not a multiple of four.

  Extra shelves above the nearest multiple of four are left unassigned and the drives must be assigned manually.

When manually assigning drives, you should assign disks symmetrically, with an equal number of drives assigned to each pool. For example, if the configuration has two storage shelves at each site, you would one shelf to the local HA pair and one shelf to the remote HA pair:

- Assign half of the disks on site_A-shelf_1 to pool 0 of node_A_1.
- Assign half of the disks on site_A-shelf_1 to pool 0 of node_A_2.
- Assign half of the disks on site_A-shelf_2 to pool 1 of node_B_1.
- Assign half of the disks on site_A-shelf_2 to pool 1 of node_B_2.
- Assign half of the disks on site_B-shelf_1 to pool 0 of node_B_1.
- Assign half of the disks on site_B-shelf_1 to pool 0 of node_B_2.
- Assign half of the disks on site_B-shelf_2 to pool 1 of node_A_1.
- Assign half of the disks on site_B-shelf_2 to pool 1 of node_A_2.

Adding shelves to an existing configuration

Automatic drive assignment supports the symmetrical addition of shelves to an existing configuration.

When new shelves are added, the system applies the same assignment policy to newly added shelves. For example, with a single shelf per site, if an additional shelf is added, the systems applies the quarter-shelf assignment rules to the new shelf.

Related information

Required MetroCluster IP components and naming conventions

Disk and aggregate management

ADP and disk assignment differences by system in MetroCluster IP configurations

The operation of Advanced Drive Partitioning (ADP) and automatic disk assignment in MetroCluster IP configurations varies depending on the system model.

ⓘ In systems using ADP, aggregates are created using partitions in which each drive is partitioned in to P1, P2 and P3 partitions. The root aggregate is created using P3 partitions.

Review the following requirements before using the tables:

- You must meet the MetroCluster limits for the maximum number of supported drives and other guidelines. Refer to the NetApp Hardware Universe.
- If you are reusing an external disk shelf, verify that disk ownership on the external disk shelf has been removed before you attach it to the controller. Refer to Remove ONTAP ownership from a disk.

ADP and disk assignment on AFF A320 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition
Minimum recommended drives (per site)	48 drives	The drives on each external shelf are divided into two equal groups (halves). Each half-shelf is automatically assigned to a separate pool.	One shelf is used by the local HA pair. The second shelf is used by the remote HA pair. Partitions on each shelf are used to create the root aggregate. Each of the two plexes in the root aggregate includes the following partitions: • Eight partitions for data • Two parity partitions • Two spare partitions

Minimum supported drives (per site)	24 drives	The drives are divided into four equal groups. Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes the following partitions: • Three partitions for data • Two parity partitions • One spare partition

ADP and disk assignment on AFF A150, ASA A150, and AFF A220 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition
Minimum recommended drives (per site)	Internal drives only	The internal drives are divided into four equal groups. Each group is automatically assigned to a separate pool and each pool is assigned to a separate controller in the configuration. **Note:** Half of the internal drives remain unassigned before MetroCluster is configured.	Two quarters are used by the local HA pair. The other two quarters are used by the remote HA pair. The root aggregate includes the following partitions in each plex: • Three partitions for data • Two parity partitions • One spare partition

Minimum supported drives (per site)	16 internal drives	The drives are divided into four equal groups. Each quarter-shelf is automatically assigned to a separate pool. Two quarters on a shelf can have the same pool. The pool is chosen based on the node that owns the quarter: • If owned by the local node, pool0 is used. • If owned by the remote node, pool1 is used. For example: a shelf with quarters Q1 through Q4 can have following assignments: • Q1: node_A_1 pool0 • Q2: node_A_2 pool0 • Q3: node_B_1 pool1 • Q4:node_B_2 pool1 **Note:** Half of the internal drives remain unassigned before MetroCluster is configured.	Each of the two plexes in the root aggregate includes the following partitions: • Two partitions for data • Two parity partitions • No spares

ADP and disk assignment on AFF A250, AFF C250, ASA A250, ASA C250, FAS500f, AFF A20, AFF A30, and AFF C30 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition

Minimum recommended drives (per site)	48 drives (external drives only, no internal drives)	The drives on each external shelf are divided into two equal groups (halves). Each half-shelf is automatically assigned to a separate pool.	One shelf is used by the local HA pair. The second shelf is used by the remote HA pair. Partitions on each shelf are used to create the root aggregate. The root aggregate includes the following partitions in each plex: • Eight partitions for data • Two parity partitions • Two spare partitions
	48 drives (external and internal drives)	The internal partitions are divided into four equal groups (quarters). Each quarter is automatically assigned to a separate pool. The drives on the external shelves are divided into four equal groups (quarters). Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes: • Eight partitions for data • Two parity partitions • Two spare partitions
Minimum supported drives (per site)	16 internal drives	The drives are divided into four equal groups. Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes the following partitions: • Two partitions for data • Two parity partitions • No spare partitions

ADP and disk assignment on AFF A50 and AFF C60 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition

Minimum recommended drives (per site)	48 drives (external drives only, no internal drives)	The drives on each external shelf are divided into two equal groups (halves). Each half-shelf is automatically assigned to a separate pool.	The local HA pair uses one shelf. The remote HA pair uses the second shelf. Partitions on each shelf are used to create the root aggregate. The root aggregate includes the following partitions in each plex: • Eight partitions for data • Two parity partitions • Two spare partitions
	48 drives (external and internal drives)	The internal partitions are divided into four equal groups (quarters). Each quarter is automatically assigned to a separate pool. The drives on the external shelves are divided into four equal groups (quarters). Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes: • Eight partitions for data • Two parity partitions • Two spare partitions
Minimum supported drives (per site)	24 internal drives	The drives are divided into four equal groups. Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes the following partitions: • Two partitions for data • Two parity partitions • No spare partitions

ADP and disk assignment on AFF A300 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition

Minimum recommended drives (per site)	48 drives	The drives on each external shelf are divided into two equal groups (halves). Each half-shelf is automatically assigned to a separate pool.	One shelf is used by the local HA pair. The second shelf is used by the remote HA pair. Partitions on each shelf are used to create the root aggregate. The root aggregate includes the following partitions in each plex: • Eight partitions for data • Two parity partitions • Two spare partitions
Minimum supported drives (per site)	24 drives	The drives are divided into four equal groups. Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes the following partitions: • Three partitions for data • Two parity partitions • One spare partition

ADP and disk assignment on AFF C400, AFF A400, ASA C400, and ASA A400 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition
Minimum recommended drives (per site)	96 drives	Drives are automatically assigned on a shelf-by-shelf basis.	Each of the two plexes in the root aggregate includes: • 20 partitions for data • Two parity partitions • Two spare partitions
Minimum supported drives (per site)	24 drives	The drives are divided into four equal groups (quarters). Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes: • Three partitions for data • Two parity partitions • One spare partition

ADP and disk assignment on AFF A700 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition
Minimum recommended drives (per site)	96 drives	Drives are automatically assigned on a shelf-by-shelf basis.	Each of the two plexes in the root aggregate includes: • 20 partitions for data • Two parity partitions • Two spare partitions
Minimum supported drives (per site)	24 drives	The drives are divided into four equal groups (quarters). Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes: • Three partitions for data • Two parity partitions • One spare partition

ADP and disk assignment on AFF C800, ASA C800, ASA A800, and AFF A800 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root aggregate
Minimum recommended drives (per site)	Internal drives and 96 external drives	The internal partitions are divided into four equal groups (quarters). Each quarter is automatically assigned to a separate pool. The drives on the external shelves are automatically assigned on a shelf-by-shelf basis, with all of the drives on each shelf assigned to one of the four nodes in the MetroCluster configuration.	Each of the two plexes in the root aggregate includes: • Eight partitions for data • Two parity partitions • Two spare partitions **Note:** The root aggregate is created with 12 root partitions on the internal shelf.

Minimum supported drives (per site)	24 internal drives	The internal partitions are divided into four equal groups (quarters). Each quarter is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes: • Three partitions for data • Two parity partitions • One spare partitions **Note:** The root aggregate is created with 12 root partitions on the internal shelf.

ADP and disk assignment on AFF A70, AFF A90, and AFF C80 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root aggregate
Minimum recommended drives (per site)	Internal drives and 96 external drives	The internal partitions are divided into four equal groups (quarters). Each quarter is automatically assigned to a separate pool. The drives on the external shelves are automatically assigned on a shelf-by-shelf basis, with all of the drives on each shelf assigned to one of the four nodes in the MetroCluster configuration.	Each of the two plexes in the root aggregate includes: • Eight partitions for data • Two parity partitions • Two spare partitions
Minimum supported drives (per site)	24 internal drives	The internal partitions are divided into four equal groups (quarters). Each quarter is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes: • Three partitions for data • Two parity partitions • One spare partitions

ADP and disk assignment on AFF A900, ASA A900, and AFF A1K systems

Guideline	Shelves per site	Drive assignment rules	ADP layout for root partition

Minimum recommended drives (per site)	96 drives	Drives are automatically assigned on a shelf-by-shelf basis.	Each of the two plexes in the root aggregate includes: • 20 partitions for data • Two parity partitions • Two spare partitions
Minimum supported drives (per site)	24 drives	The drives are divided into four equal groups (quarters). Each quarter-shelf is automatically assigned to a separate pool.	Each of the two plexes in the root aggregate includes: • Three partitions for data • Two parity partitions • One spare partition

Disk assignment on FAS2750 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition
Minimum recommended drives (per site)	24 internal drives and 24 external drives	The internal and external shelves are divided into two equal halves. Each half is automatically assigned to different pool	Not applicable
Minimum supported drives (per site) (active/passive HA configuration)	Internal drives only	Manual assignment required	Not applicable

Disk assignment on FAS8200 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition
Minimum recommended drives (per site)	48 drives	The drives on the external shelves are divided into two equal groups (halves). Each half-shelf is automatically assigned to a separate pool.	Not applicable

Minimum supported drives (per site) (active/passive HA configuration)	24 drives	Manual assignment required.	Not applicable

Disk assignment on FAS500f systems

The same disk assignment guidelines and rules for AFF C250 and AFF A250 systems apply to FAS500f systems. For disk assignment on FAS500f systems, refer to the ADP and disk assignment on AFF A250, AFF C250, ASA A250, ASA C250, FAS500f, AFF A20, AFF A30, and AFF C30 systems table.

Disk assignment on FAS9000, FAS9500, FAS70, and FAS90 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition
Minimum recommended drives (per site)	96 drives	Drives are automatically assigned on a shelf-by-shelf basis.	Not applicable
Minimum supported drives (per site)	24 drives	The drives are divided into four equal groups (quarters). Each quarter-shelf is automatically assigned to a separate pool.	Not applicable

Disk assignment on FAS50 systems

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition

Minimum recommended drives (per site)	48 drives (external drives only, no internal drives)	The drives on each external shelf are divided into two equal groups (halves). Each half-shelf is automatically assigned to a separate pool.	Not applicable
	48 drives (external and internal drives)	The internal partitions are divided into four equal groups (quarters). Each quarter is automatically assigned to a separate pool. The drives on the external shelves are divided into four equal groups (quarters). Each quarter-shelf is automatically assigned to a separate pool.	Not applicable
Minimum supported drives (per site)	24 drives	The drives are divided into four equal groups. Each quarter-shelf is automatically assigned to a separate pool.	Not applicable

Disk assignment on FAS50 systems with Flash Cache

Beginning with ONTAP 9.18.1, Flash Cache is supported on FAS50 systems for MetroCluster IP configurations.

- You cannot have both data aggregates and the root aggregate with Flash Cache drives on the internal shelf.
- Slots 0 and 23 are used for Flash Cache drives.
- If you are reusing an external disk shelf, verify that disk ownership on the external disk shelf has been removed before you attach it to the controller. Refer to Remove ONTAP ownership from a disk.

Guideline	Drives per site	Drive assignment rules	ADP layout for root partition

Minimum recommended drives (per site)	48 drives (external drives only, no internal drives)	The drives on each external shelf are divided into two equal groups (halves). Each half-shelf is automatically assigned to a separate pool.	Not applicable
	36 drives (12 internal drives and 24 external drives – with the data aggregates on the external shelf and the root aggregate on the internal shelf)	The internal drives are divided into four equal groups (quarters). Each quarter is automatically assigned to a separate pool. The drives on the external shelves are divided into four equal groups (quarters). Each quarter-shelf is automatically assigned to a separate pool. **Notes:** • If you are using internal and external drives, you first need to install ONTAP and configure the local cluster with only the internal drives. You then manually connect the external shelf after you complete your installation or setup task. • A minimum of 12 internal drives is required for the root aggregate. You should place the root internal drives from slot 1. For example, for 12 internal drives, use slots 1-3, 6-8, 12-14, and 18-20.	Not applicable
Minimum supported drives (per site)	24 external drives	The drives are divided into four equal groups. Each quarter-shelf is automatically assigned to a separate pool.	Not applicable

Requirements for cluster peering in MetroCluster IP configurations

Each MetroCluster site is configured as a peer to its partner site. You must be familiar with the prerequisites and guidelines for configuring the peering relationships. This is important when deciding on whether to use shared or dedicated ports for those relationships.

Related information

Cluster and SVM peering express configuration

Prerequisites for cluster peering

Before you set up cluster peering, you should confirm that connectivity between port, IP address, subnet, firewall, and cluster-naming requirements are met.

Connectivity requirements

Every intercluster LIF on the local cluster must be able to communicate with every intercluster LIF on the remote cluster.

Although it is not required, it is typically simpler to configure the IP addresses used for intercluster LIFs in the same subnet. The IP addresses can reside in the same subnet as data LIFs, or in a different subnet. The subnet used in each cluster must meet the following requirements:

- The subnet must have enough IP addresses available to allocate to one intercluster LIF per node.

  For example, in a four-node cluster, the subnet used for intercluster communication must have four available IP addresses.

Each node must have an intercluster LIF with an IP address on the intercluster network.

Intercluster LIFs can have an IPv4 address or an IPv6 address.

> ONTAP 9 enables you to migrate your peering networks from IPv4 to IPv6 by optionally allowing both protocols to be present simultaneously on the intercluster LIFs. In earlier releases, all intercluster relationships for an entire cluster were either IPv4 or IPv6. This meant that changing protocols was a potentially disruptive event.

Port requirements

You can use dedicated ports for intercluster communication, or share ports used by the data network. Ports must meet the following requirements:

- All ports used to communicate with a given remote cluster must be in the same IPspace.

  You can use multiple IPspaces to peer with multiple clusters. Pair-wise full-mesh connectivity is required only within an IPspace.

- The broadcast domain used for intercluster communication must include at least two ports per node so that intercluster communication can fail over from one port to another port.

Ports added to a broadcast domain can be physical network ports, VLANs, or interface groups (ifgrps).

- All ports must be cabled.
- All ports must be in a healthy state.
- The MTU settings of the ports must be consistent.

Firewall requirements

Firewalls and the intercluster firewall policy must allow the following protocols:

- ICMP service
- TCP to the IP addresses of all the intercluster LIFs over the ports 10000, 11104, and 11105
- Bidirectional HTTPS between the intercluster LIFs

The default intercluster firewall policy allows access through the HTTPS protocol and from all IP addresses (0.0.0.0/0). You can modify or replace the policy if necessary.

Considerations when using dedicated ports

When determining whether using a dedicated port for intercluster replication is the correct intercluster network solution, you should consider configurations and requirements such as LAN type, available WAN bandwidth, replication interval, change rate, and number of ports.

Consider the following aspects of your network to determine whether using a dedicated port is the best intercluster network solution:

- If the amount of available WAN bandwidth is similar to that of the LAN ports, and the replication interval is such that replication occurs while regular client activity exists, then you should dedicate Ethernet ports for intercluster replication to avoid contention between replication and the data protocols.
- If the network utilization generated by the data protocols (CIFS, NFS, and iSCSI) is such that the network utilization is above 50 percent, then dedicate ports for replication to allow for nondegraded performance if a node failover occurs.
- When physical 10 GbE or faster ports are used for data and replication, you can create VLAN ports for replication and dedicate the logical ports for intercluster replication.

  The bandwidth of the port is shared between all VLANs and the base port.

- Consider the data change rate and replication interval and whether the amount of data, that must be replicated on each interval, requires enough bandwidth. This might cause contention with data protocols if sharing data ports.

Considerations when sharing data ports

When determining whether sharing a data port for intercluster replication is the correct intercluster network solution, you should consider configurations and requirements such as LAN type, available WAN bandwidth, replication interval, change rate, and number of ports.

Consider the following aspects of your network to determine whether sharing data ports is the best intercluster connectivity solution:

- For a high-speed network, such as a 40-Gigabit Ethernet (40-GbE) network, a sufficient amount of local LAN bandwidth might be available to perform replication on the same 40-GbE ports that are used for data access.

In many cases, the available WAN bandwidth is far less than the 10 GbE LAN bandwidth.

- All nodes in the cluster might have to replicate data and share the available WAN bandwidth, making data port sharing more acceptable.
- Sharing ports for data and replication eliminates the extra port counts required to dedicate ports for replication.
- The maximum transmission unit (MTU) size of the replication network will be the same size as that used on the data network.
- Consider the data change rate and replication interval and whether the amount of data, that must be replicated on each interval, requires enough bandwidth. This might cause contention with data protocols if sharing data ports.
- When data ports for intercluster replication are shared, the intercluster LIFs can be migrated to any other intercluster-capable port on the same node to control the specific data port that is used for replication.

ISL requirements

Inter-Switch Link requirements for MetroCluster IP configurations

You should verify that your MetroCluster IP configuration and network meets all Inter-Switch Link (ISL) requirements. Although certain requirements might not apply to your configuration, you should still be aware of all of the ISL requirements to gain a better understanding of the overall configuration.

The following table provides an overview of the topics covered in this section.

Title	Description
NetApp-validated and MetroCluster-compliant switches	Describes the switch requirements. Applies to all switches used in MetroCluster configurations, including backend switches.
Considerations for ISLs	Describes the ISL requirements. Applies to all MetroCluster configurations, regardless of network topology and whether you use NetApp-validated switches or MetroCluster-compliant switches.
Considerations when deploying MetroCluster in a shared layer 2 or layer 3 networks	Describes the requirements for shared layer 2 or layer 3 networks. Applies to all configurations except for MetroCluster configurations using NetApp-validated switches and using direct connected ISLs.
Considerations when using MetroCluster Compliant switches	Describes the requirements for MetroCluster-compliant switches. Applies to all MetroCluster configurations that are not using NetApp-validated switches.
Examples of MetroCluster network topologies	Provides examples of different MetroCluster network topologies. Applies to all MetroCluster configurations.

NetApp-validated and MetroCluster-compliant switches in a MetroCluster IP configuration

All of the switches used in your configuration, including backend switches, must either be NetApp-validated or MetroCluster-compliant.

NetApp-validated switches

A switch is NetApp-validated if it meets the following requirements:

- The switch is provided by NetApp as part of the MetroCluster IP configuration
- The switch is listed in the NetApp Hardware Universe as a supported switch under *MetroCluster-over-IP-connections*
- The switch is only used to connect MetroCluster IP controllers and, in some configurations, NS224 drive shelves
- The switch is configured using the Reference Configuration File (RCF) provided by NetApp

Any switch that does not meet these requirements is **not** a NetApp-validated switch.

MetroCluster-compliant switches

A MetroCluster-compliant switch is not NetApp-validated but can be used in a MetroCluster IP configuration if it meets certain requirements and configuration guidelines.

> ⓘ NetApp does not provide troubleshooting or configuration support services for any non-validated MetroCluster-compliant switch.

Requirements for Inter-Switch Links (ISLs) on MetroCluster IP configurations

Inter-Switch Links (ISLs) carrying MetroCluster traffic on all MetroCluster IP configurations and network topologies have certain requirements. These requirements apply to all ISLs carrying MetroCluster traffic, regardless of whether the ISLs are direct or shared between customer switches.

MetroCluster ISL requirements

The following applies to ISLs on all MetroCluster IP configurations:

- Both fabrics must have the same number of ISLs.
- ISLs on one fabric must all be the same speed and length.
- ISLs in both fabrics must be the same speed and length.
- The maximum supported difference in distance between fabric 1 and fabric 2 is 20km or 0.2ms.
- The ISLs must have the same topology. For example, they should all be direct links, or if the configuration uses WDM, then they must all use WDM.
- The minimum required ISL speed depends on the platform model:
  - Beginning with ONTAP 9.18.1, platforms with a MetroCluster IP backend port speed of 100G require a minimum ISL link speed of 100Gbps. Using a different ISL speed requires a Feature Variance Request (FPVR). To file an FPVR, please contact your NetApp sales team.

    - On all other platforms, the minimum supported ISL link speed is 10Gbps.
- There must be least one 10Gbps ISL port per fabric.

Latency and packet loss limits in the ISLs

The following applies to round-trip traffic between the MetroCluster IP switches at site_A and site_B, with the MetroCluster configuration in steady state operation:

- As the distance between two MetroCluster sites increases, latency increases, usually in the range of 1 ms round-trip delay time per 100 km (62 miles). Latency also depends on the network service level agreement (SLA) in terms of the bandwidth of the ISL links, packet drop rate, and jitter on the network. Low bandwidth, high jitter, and random packet drops lead to different recovery mechanisms by the switches, or the TCP engine on the controller modules, for successful packet delivery. These recovery mechanisms can increase overall latency. For specific information on round trip latency and maximum distance requirements for your configuration, refer to the Hardware Universe.
- Any device that contributes to latency must be accounted for.
- The Hardware Universe. provides the distance in km. You must allocate 1ms for every 100km. The maximum distance is defined by what is reached first, either the maximum round-trip time (RTT) in ms, or the distance in km. For example – if *The Hardware Universe* lists a distance of 300km, translating to 3ms, your ISL can be no further than 300km and the max RTT cannot exceed 3ms – whichever is reached first.
- Packet loss must be less than, or equal to, 0.01%. The maximum packet loss is the sum of all loss on all links on the path between the MetroCluster nodes, and the loss on the local MetroCluster IP interfaces.
- The supported jitter value is 3ms for round trip (or 1.5ms for one-way).
- The network should allocate and maintain the SLA amount of bandwidth required for MetroCluster traffic, regardless of microbursts and spikes in the traffic.
- If you are using ONTAP 9.7 or later, the intermediate network between the two sites must provide a minimum bandwidth of 4.5Gbps for the MetroCluster IP configuration.

Transceiver and cable considerations

Any SFPs or QSFPs supported by the equipment vendor are supported for the MetroCluster ISLs. SFPs and QSFPs provided by NetApp or the equipment vendor must be supported by the switch and switch firmware.

When connecting the controllers to the switches and the local cluster ISLs, you must use the transceivers and cables provided by NetApp with the MetroCluster.

When you use a QSFP-SFP adapter, whether you configure the port in breakout or native speed mode depends on the switch model and firmware. For example, using a QSFP-SFP adapter with Cisco 9336C switches running NX-OS firmware 9.x or 10.x requires that you configure the port in native speed mode.

> ⓘ If you configure an RCF, verify that you select the correct speed mode or use a port with an appropriate speed mode.

Using xWDM, TDM, and external encryption devices

When you use xWDM/TDM devices or devices providing encryption in a MetroCluster IP configuration your environment must meet the following requirements:

- When connecting the MetroCluster IP switches to the xWDM/TDM, the external encryption devices or xWDM/TDM equipment must be certified by the vendor for the switch and firmware. The certification must cover the operating mode (such as trunking and encryption).

- The overall end-to-end latency and jitter, including the encryption, cannot be more than the maximum amount stated in the IMT and in this documentation.

Supported number of ISLs and breakout cables

The following table shows the supported maximum number of ISLs that can be configured on a MetroCluster IP switch using the Reference Configuration File (RCF) configuration.

MetroCluster IP switch model	Port type	Maximum number of ISLs
Broadcom-supported BES-53248 switches	Native ports	4 ISLs using 10Gbps or 25Gbps
Broadcom-supported BES-53248 switches	Native ports (Note 1)	2 ISLs using 40Gbps or 100Gbps
Cisco 3132Q-V	Native ports	6 ISLs using 40Gbps
Cisco 3132Q-V	Breakout cables	16 ISLs using 10Gbps
Cisco 3232C	Native ports	6 ISLs using 40Gbps or 100Gbps
Cisco 3232C	Breakout cables	16 ISLs using 10Gbps or 25Gbps
Cisco 9336C-FX2 (not connecting NS224 shelves)	Native ports	6 ISLs using 40Gbps or 100Gbps
Cisco 9336C-FX2 (not connecting NS224 shelves)	Breakout cables	16 ISLs using 10Gbps or 25Gbps
Cisco 9336C-FX2 (connecting NS224 shelves)	Native ports (Note 2)	4 ISLs using 40Gbps or 100Gbps
Cisco 9336C-FX2 (connecting NS224 shelves)	Breakout cables (Note 2)	16 ISLs using 10Gbps or 25Gbps
NVIDIA SN2100	Native ports (Note 2)	2 ISLs using 40Gbps or 100Gbps
NVIDIA SN2100	Breakout cables (Note 2)	8 ISLs using 10Gbps or 25Gbps

Note 1: Using 40Gbps or 100Gbps ISLs on a BES-53248 switch requires an additional license.

Note 2: The same ports are used for native speed and breakout mode. You must choose to use ports in native speed mode or breakout mode when creating the RCF file.

- All ISLs on one MetroCluster IP switch must be the same speed. Using a mix of ISL ports with different speeds concurrently is not supported.
- For optimum performance, you should use at least one 40Gbps ISL per network. You should not use a single 10Gbps ISL per network for FAS9000, AFF A700, or other high capacity platforms.

NetApp recommends that you configure a small number of high bandwidth ISLs, rather than a high number of low bandwidth ISLs. For example, configuring one 40Gbps ISL instead of four 10Gbps ISLs is preferred. When using multiple ISLs, statistical load-balancing can impact the maximum throughput. Uneven balancing can reduce throughput to that of a single ISL.

Requirements to deploy MetroCluster IP configurations in shared layer 2 or layer 3 networks

Depending on your requirements, you can use shared layer 2 or layer 3 networks to deploy MetroCluster.

Beginning with ONTAP 9.6, MetroCluster IP configurations with supported switches can share existing networks for Inter-Switch Links (ISLs) instead of using dedicated MetroCluster ISLs. This topology is known as *shared layer 2 networks*.

Beginning with ONTAP 9.9.1, MetroCluster IP configurations can be implemented with IP-routed (layer 3) backend connections. This topology is known as *shared layer 3 networks*.

- Not all features are supported in all network topologies.
- You must verify that you have adequate network capacity and that the ISL size is appropriate for your configuration. Low latency is critical for replication of data between the MetroCluster sites. Latency issues on these connections can impact client I/O.
- All references to MetroCluster backend switches refer to switches that are NetApp-validated switches or MetroCluster-compliant. See NetApp-validated and MetroCluster-compliant switches for more details.

ISL requirements for layer 2 and layer 3 networks

The following applies to layer 2 and layer 3 networks:

- The speed and number of ISLs between the MetroCluster switches and the intermediate network switches does not need to match. Similarly, the speed between the intermediate network switches does not need to match.

  For example, MetroCluster switches can connect using one 40Gbps ISL to the intermediate switches, and the intermediate switches can connect to each other using two 100Gbps ISLs.

- Network monitoring should be configured on the intermediate network to monitor the ISLs for utilization, errors (drops, link flaps, corruption, and so on), and failures.
- The MTU size must be set to 9216 on all ports carrying MetroCluster end-to-end traffic.
- No other traffic can be configured with a higher priority than class of service (COS) 5.
- Explicit congestion notification (ECN) must be configured on all paths carrying end-to-end MetroCluster traffic.
- ISLs carrying MetroCluster traffic must be native links between the switches.

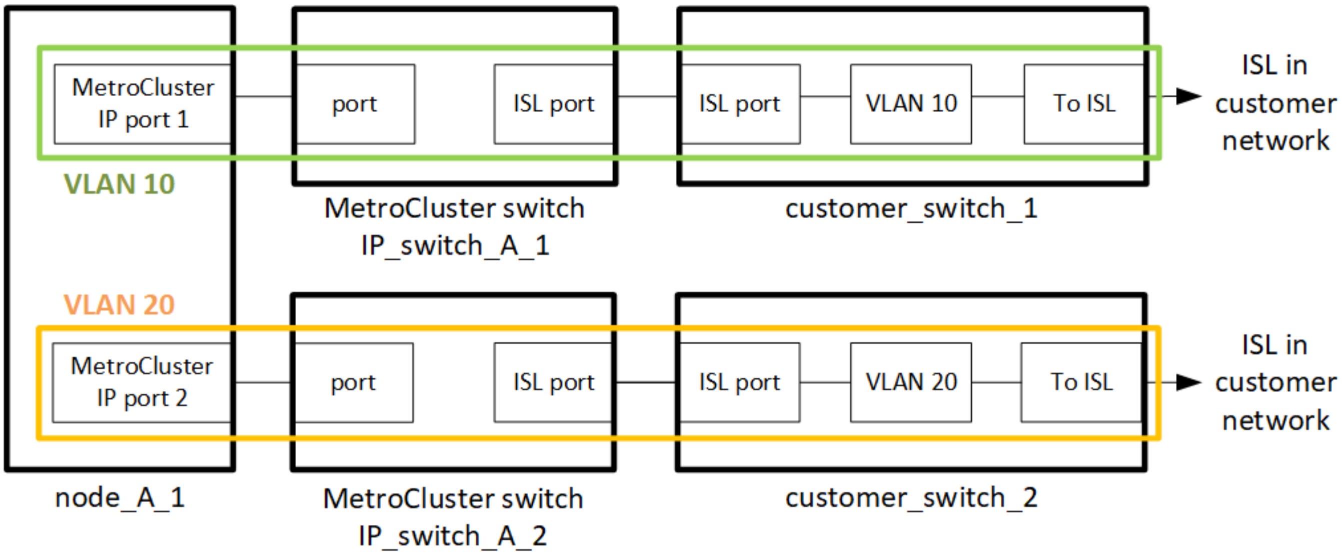
  Link sharing services such as Multiprotocol Label Switching (MPLS) links are not supported.

- The layer 2 VLANs must natively span the sites. VLAN overlay such as Virtual Extensible LAN (VXLAN) is not supported.

- The number of intermediate switches is not limited. However, NetApp recommends that you keep the number of switches to the minimum required.
- ISLs on MetroCluster switches are configured with the following:
    - Switch port mode 'trunk' as part of an LACP port-channel
    - The MTU size is 9216
    - No native VLAN is configured
    - Only VLANs carrying cross site MetroCluster traffic are allowed
    - The switch default VLAN is not allowed

Considerations for layer 2 networks

The MetroCluster backend switches are connected to the customer network.



The intermediate customer-provided switches must meet the following requirements:

- The intermediate network must provide the same VLANs between the sites. This must match the MetroCluster VLANs set in the RCF file.
- The RcfFileGenerator does not allow the creation of an RCF file using VLANs that are not supported by the platform.
- The RcfFileGenerator might restrict the use of certain VLAN IDs, for example, if they are intended for future use. Generally, reserved VLANs are up to and including 100.
- Layer 2 VLANs with IDs that match the MetroCluster VLAN IDs must span the shared network.

VLAN configuration in ONTAP

You can only specify the VLAN during interface creation. You can configure the default VLANs 10 and 20, or VLANs within the range 101 to 4096 (or the number supported by the switch vendor, whichever is the lower number). After the MetroCluster interfaces are created, you cannot change the VLAN ID.

ⓘ Some switch vendors might reserve the use of certain VLANs.

The following systems do not require VLAN configuration within ONTAP. The VLAN is specified by the switch port configuration:

- FAS8200 and AFF A300
- AFF A320
- FAS9000 and AFF A700
- AFF A800, ASA A800, AFF C800, and ASA C800

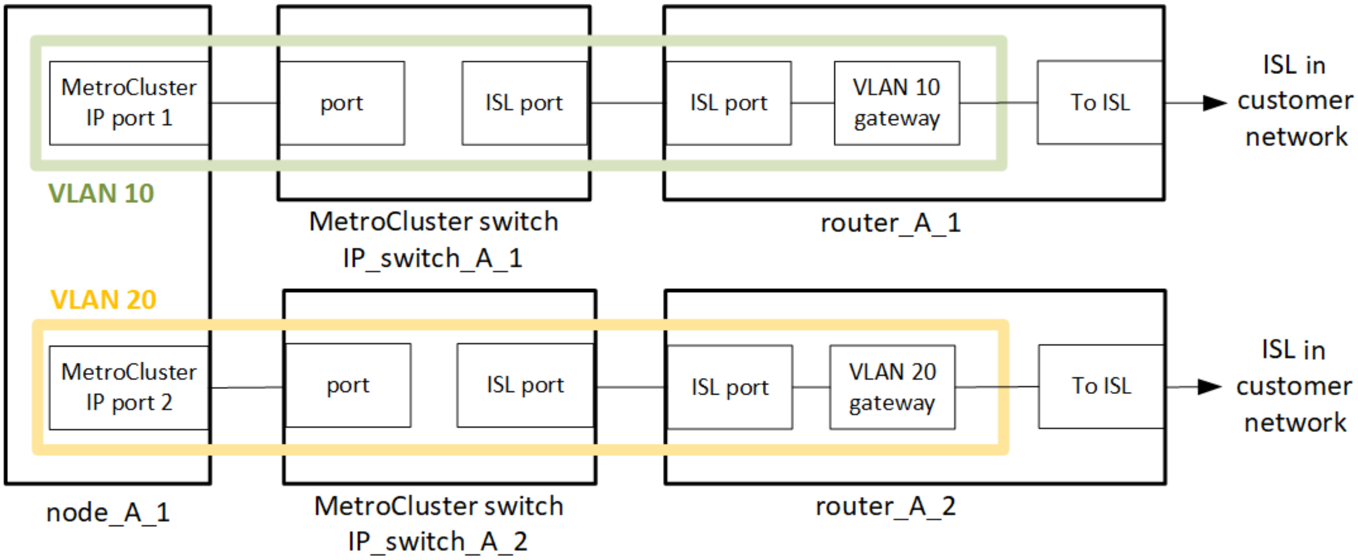
> The systems listed above might be configured using VLANs 100 and below. However, some VLANs in this range might be reserved for other or future use.

For all other systems, you must configure the VLAN when you create the MetroCluster interfaces in ONTAP. The following restrictions apply:

- The default VLAN is 10 and 20
- If you are running ONTAP 9.7 or earlier, you can only use the default VLAN 10 and 20.
- If you are running ONTAP 9.8 or later, you can use the default VLAN 10 and 20, and a VLAN over 100 (101 and higher) can also be used.

Considerations for layer 3 networks

The MetroCluster backend switches are connected to the routed IP network, either directly to routers (as shown in the following simplified example) or through other intervening switches.



The MetroCluster environment is configured and cabled as a standard MetroCluster IP configuration as described in Configure the MetroCluster hardware components. When you perform the installation and cabling procedure, you must perform the steps specific to a layer 3 configuration. The following applies to layer 3 configurations:

- You can connect MetroCluster switches directly to the router or to one or more intervening switches.
- You can connect MetroCluster IP interfaces directly to the router or to one of the intervening switches.
- The VLAN must be extended to the gateway device.
- You use the `-gateway parameter` to configure the MetroCluster IP interface address with an IP gateway address.
- The VLAN IDs for the MetroCluster VLANs must be the same at each site. However, the subnets can be

different.

- Dynamic routing is not supported for the MetroCluster traffic.
- The following features are not supported:
  - Eight-node MetroCluster configurations
  - Refreshing a four-node MetroCluster configuration
  - Transition from MetroCluster FC to MetroCluster IP
- Two subnets are required on each MetroCluster site—one in each network.
- Auto-IP assignment is not supported.

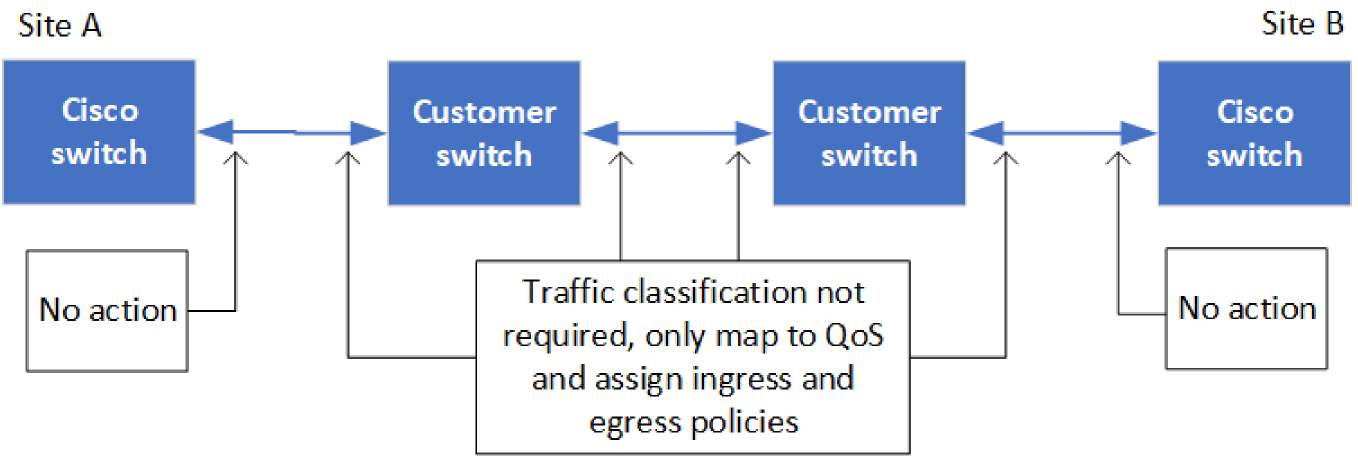
When you configure routers and gateway IP addresses, you must meet the following requirements:

- Two interfaces on one node cannot have the same gateway IP address.
- The corresponding interfaces on the HA pairs on each site must have the same gateway IP address.
- The corresponding interfaces on a node and its DR and AUX partners cannot have the same gateway IP address.
- The corresponding interfaces on a node and its DR and AUX partners must have the same VLAN ID.

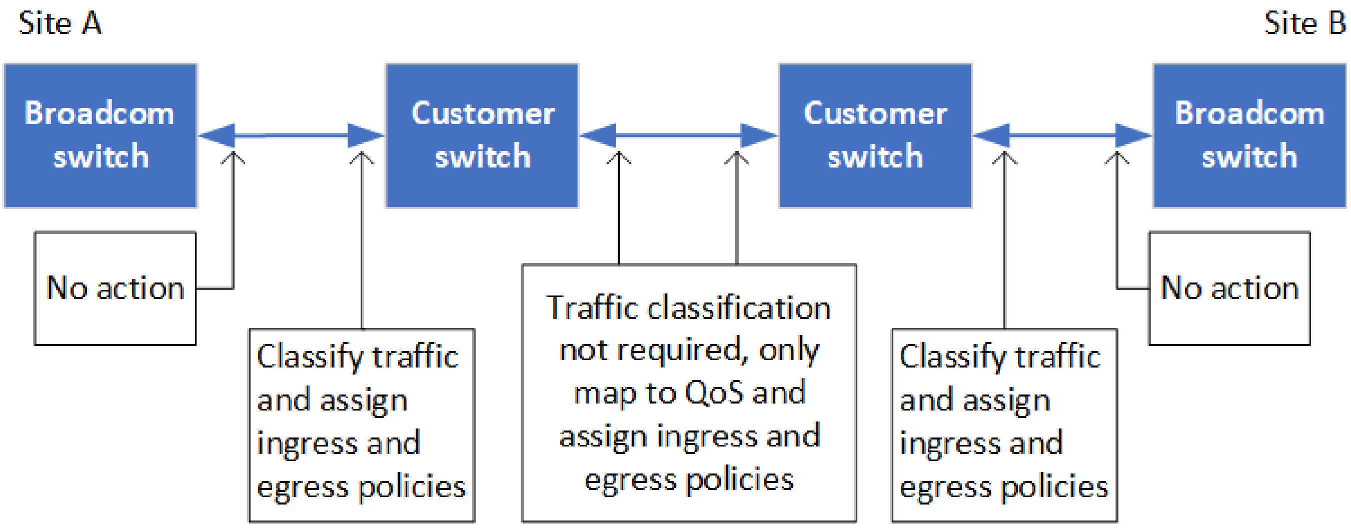
Required settings for intermediate switches

When MetroCluster traffic traverses an ISL in an intermediate network, you should verify that the configuration of the intermediate switches ensures that the MetroCluster traffic (RDMA and storage) meets the required service levels across the entire path between the MetroCluster sites.

The following diagram gives an overview of the required settings when using NetApp validated Cisco switches:


Site A
Site B
Cisco switch
Customer switch
Customer switch
Cisco switch
No action
Traffic classification not required, only map to QoS and assign ingress and egress policies
No action

The following diagram gives an overview of the required settings for a shared network when the external switches are Broadcom IP switches.


Site A
Site B
Broadcom switch
Customer switch
Customer switch
Broadcom switch
No action
Classify traffic and assign ingress and egress policies
Traffic classification not required, only map to QoS and assign ingress and egress policies
Classify traffic and assign ingress and egress policies
No action

In this example, the following policies and maps are created for MetroCluster traffic:

- The `MetroClusterIP_ISL_Ingress` policy is applied to ports on the intermediate switch that connects to the MetroCluster IP switches.

  The `MetroClusterIP_ISL_Ingress` policy maps the incoming tagged traffic to the appropriate queue on the intermediate switch.
- A `MetroClusterIP_ISL_Egress` policy is applied to ports on the intermediate switch that connect to ISLs between intermediate switches.
- You must configure the intermediate switches with matching QoS access-maps, class-maps, and policy-maps along the path between the MetroCluster IP switches. The intermediate switches map RDMA traffic to COS5 and storage traffic to COS4.

The following examples are for Cisco Nexus 3232C and 9336C-FX2 switches. Depending on your switch vendor and model, you must verify that your intermediate switches have an appropriate configuration.

Configure the class map for the intermediate switch ISL port

The following example shows the class map definitions depending on whether you need to classify or match traffic on ingress.

Classify traffic on ingress:

```
ip access-list rdma
  10 permit tcp any eq 10006 any
  20 permit tcp any any eq 10006
ip access-list storage
  10 permit tcp any eq 65200 any
  20 permit tcp any any eq 65200

class-map type qos match-all rdma
  match access-group name rdma
class-map type qos match-all storage
  match access-group name storage
```

Match traffic on ingress:

```
class-map type qos match-any c5
  match cos 5
  match dscp 40
class-map type qos match-any c4
  match cos 4
  match dscp 32
```

Create an ingress policy map on the ISL port of the intermediate switch:

The following examples show how to create an ingress policy map depending on whether you need to classify or match traffic on ingress.

Classify the traffic on ingress:

```
policy-map type qos MetroClusterIP_ISL_Ingress_Classify
  class rdma
    set dscp 40
    set cos 5
    set qos-group 5
  class storage
    set dscp 32
    set cos 4
    set qos-group 4
  class class-default
    set qos-group 0
```

Match the traffic on ingress:

```
policy-map type qos MetroClusterIP_ISL_Ingress_Match
  class c5
    set dscp 40
    set cos 5
    set qos-group 5
  class c4
    set dscp 32
    set cos 4
    set qos-group 4
  class class-default
    set qos-group 0
```

Configure the egress queuing policy for the ISL ports

The following example shows how to configure the egress queuing policy:

```
policy-map type queuing MetroClusterIP_ISL_Egress
   class type queuing c-out-8q-q7
      priority level 1
   class type queuing c-out-8q-q6
      priority level 2
   class type queuing c-out-8q-q5
      priority level 3
      random-detect threshold burst-optimized ecn
   class type queuing c-out-8q-q4
      priority level 4
      random-detect threshold burst-optimized ecn
   class type queuing c-out-8q-q3
      priority level 5
   class type queuing c-out-8q-q2
      priority level 6
   class type queuing c-out-8q-q1
      priority level 7
   class type queuing c-out-8q-q-default
      bandwidth remaining percent 100
      random-detect threshold burst-optimized ecn
```

These settings must be applied on all switches and ISLs carrying MetroCluster traffic.

In this example, Q4 and Q5 are configured with `random-detect threshold burst-optimized ecn`. Depending on your configuration, you might need to set the minimum and maximum thresholds, as shown in the following example:

```
class type queuing c-out-8q-q5
  priority level 3
  random-detect minimum-threshold 3000 kbytes maximum-threshold 4000
kbytes drop-probability 0 weight 0 ecn
class type queuing c-out-8q-q4
  priority level 4
  random-detect minimum-threshold 2000 kbytes maximum-threshold 3000
kbytes drop-probability 0 weight 0 ecn
```

Minimum and maximum values vary depending on the switch and your requirements.

Example 1: Cisco

If your configuration has Cisco switches, you do not need to classify on the first ingress port of the intermediate switch. You then configure the following maps and policies:

- `class-map type qos match-any c5`
- `class-map type qos match-any c4`

- `MetroClusterIP_ISL_Ingress_Match`

You assign the `MetroClusterIP_ISL_Ingress_Match` policy map to the ISL ports carrying MetroCluster traffic.

Example 2: Broadcom

If your configuration has Broadcom switches, you must classify on the first ingress port of the intermediate switch. You then configure the following maps and policies:

- `ip access-list rdma`
- `ip access-list storage`
- `class-map type qos match-all rdma`
- `class-map type qos match-all storage`
- `MetroClusterIP_ISL_Ingress_Classify`
- `MetroClusterIP_ISL_Ingress_Match`

You assign `the MetroClusterIP_ISL_Ingress_Classify` policy map to the ISL ports on the intermediate switch connecting the Broadcom switch.

You assign the `MetroClusterIP_ISL_Ingress_Match` policy map to the ISL ports on the intermediate switch that is carrying MetroCluster traffic but does not connect the Broadcom switch.

MetroCluster IP configuration network topology examples

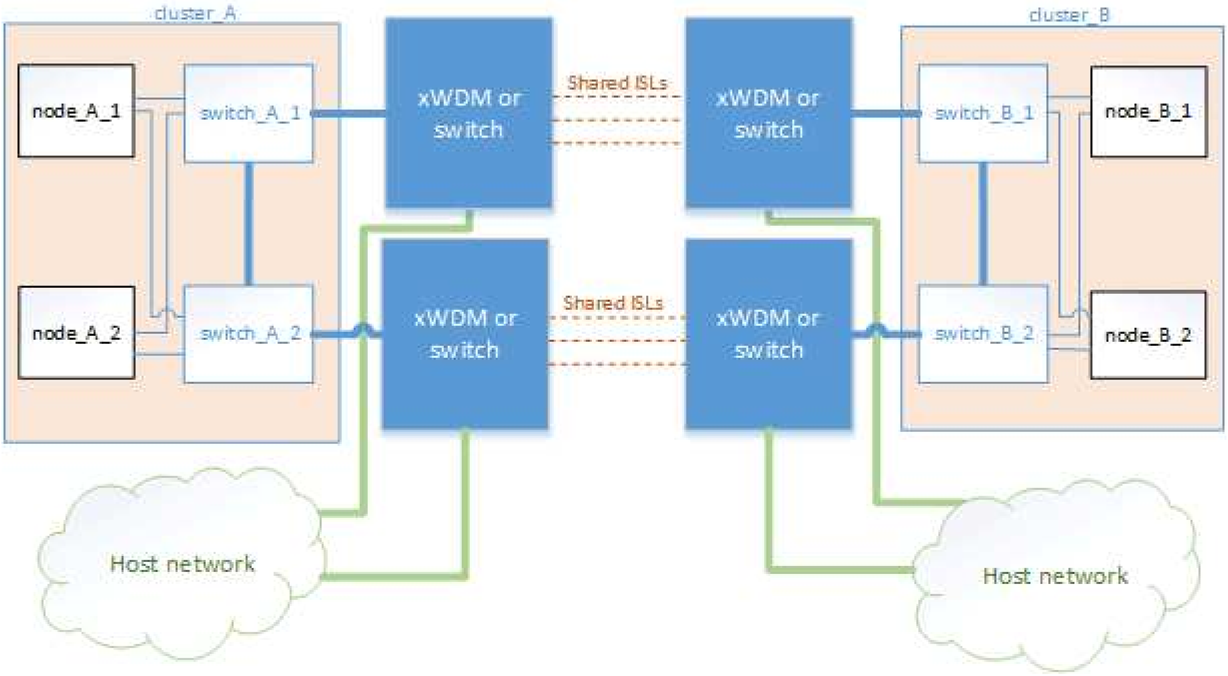
Beginning with ONTAP 9.6, some additional network configurations are supported for MetroCluster IP configurations. This section provides some examples of the supported network configurations. Not all of the supported topologies are listed.

In these topologies, it is assumed that the ISL and intermediate network is configured according to the requirements outlined in Considerations for ISLs.

> ⓘ If you are sharing an ISL with non-MetroCluster traffic, you must verify that the MetroCluster has at least the minimum required bandwidth available at all times.

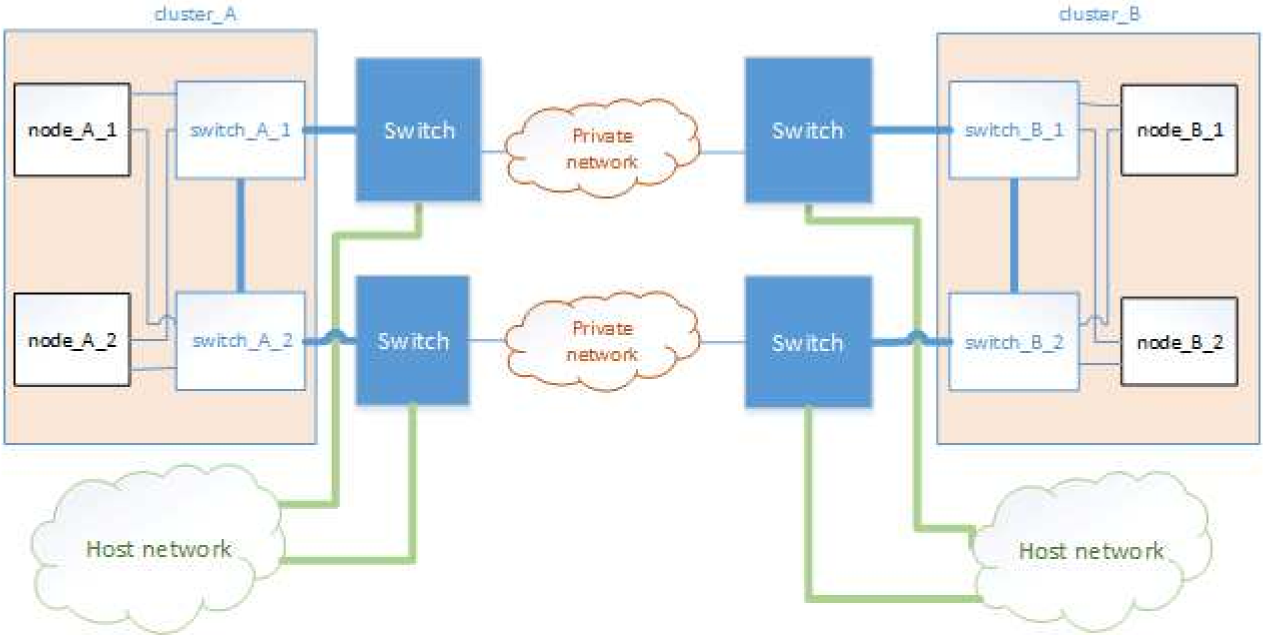
Shared network configuration with direct links

In this topology, two distinct sites are connected by direct links. These links can be between xWDM and TDM devices or switches. The capacity of the ISLs is not dedicated to the MetroCluster traffic but is shared with other non-MetroCluster traffic.



Shared infrastructure with intermediate networks

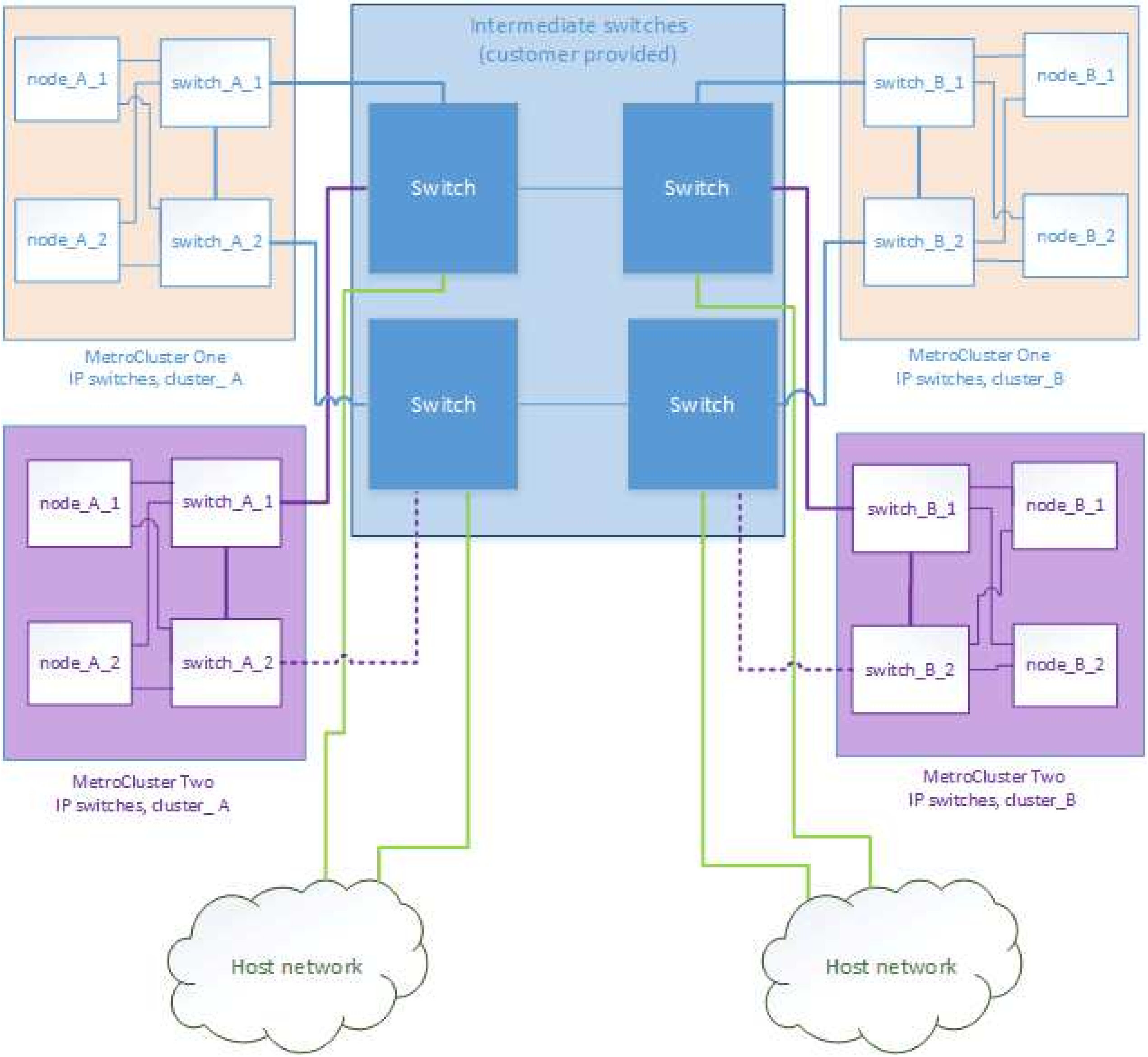
In this topology, the MetroCluster sites are not directly connected but MetroCluster and the host traffic travel through a network. The network can consist of a series of xWDM and TDM and switches, but unlike the shared configuration with direct ISLs, the links are not direct between the sites. Depending on the infrastructure between the sites, any combination of network configurations is possible.



Multiple MetroCluster configurations sharing an intermediate network

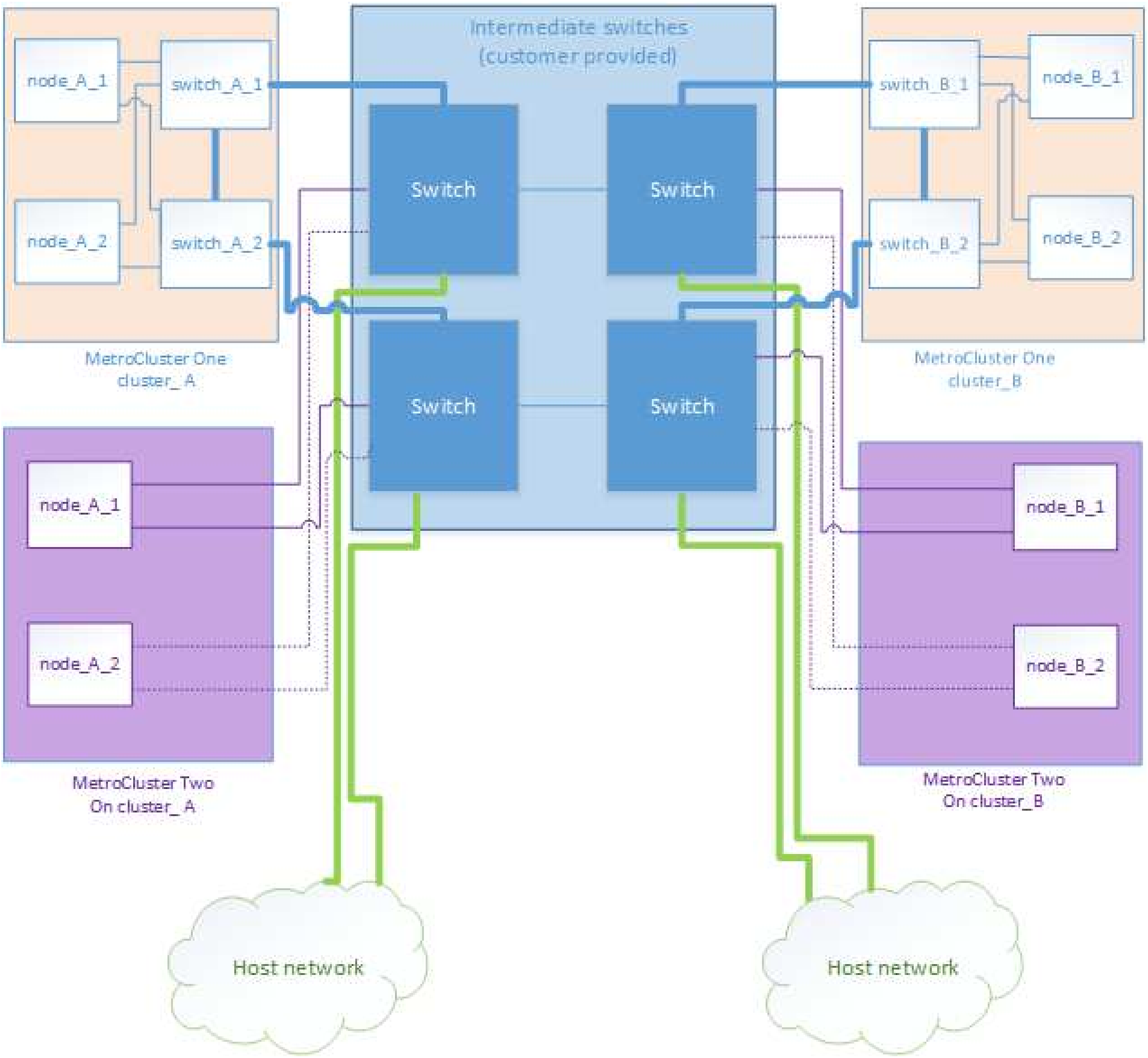
In this topology, two separate MetroCluster configurations are sharing the same intermediate network. In the example, MetroCluster one switch_A_1 and MetroCluster two switch_A_1, both connect to the same intermediate switch.

ⓘ Both “MetroCluster one” or “MetroCluster two” can be one eight-node MetroCluster configuration or two four-node MetroCluster configurations.



Combination of a MetroCluster configuration using NetApp validated switches and a configuration using MetroCluster-compliant switches

Two separate MetroCluster configurations share the same intermediate switch, where one MetroCluster is configured using NetApp validated switches in a shared layer 2 configuration (MetroCluster one), and the other MetroCluster is configured using MetroCluster-compliant switches connecting directly to the intermediate switches (MetroCluster two).



Considerations for using MetroCluster-compliant switches

Requirements and limitations for MetroCluster-compliant switches

Beginning with ONTAP 9.7, MetroCluster IP configurations can use MetroCluster-compliant switches. These are switches that are not NetApp-validated but are compliant with NetApp specifications. However, NetApp does not provide troubleshooting or configuration support services for any non-validated switch. You should be aware of the general requirements and limitations when using MetroCluster-compliant switches.

MetroCluster-compliant versus NetApp-validated switches

A switch is NetApp-validated if it meets the following requirements:

- The switch is provided by NetApp as part of the MetroCluster IP configuration
- The switch is listed in the NetApp Hardware Universe as a supported switch under *MetroCluster-over-IP-connections*

- The switch is only used to connect MetroCluster IP controllers and, in some configurations, NS224 drive shelves
- The switch is configured using the Reference Configuration File (RCF) provided by NetApp

Any switch that does not meet these requirements is **not** a NetApp-validated switch.

A MetroCluster-compliant switch is not NetApp-validated but can be used in a MetroCluster IP configuration if it meets certain requirements and configuration guidelines.

> NetApp does not provide troubleshooting or configuration support services for any non-validated MetroCluster-compliant switch.

General requirements for MetroCluster-compliant switches

The switch connecting the MetroCluster IP interfaces must meet the following general requirements:

- The switches must support quality of service (QoS) and traffic classification.
- The switches must support explicit congestion notification (ECN).
- The switches must support a load-balancing policy to preserve order along the path.
- The switches must support L2 Flow Control (L2FC).
- The switch port must provide a dedicated rate and must not be overallocated.
- The cables and transceivers connecting the nodes to the switches must be provided by NetApp. These cables must be supported by the switch vendor. If you are using optical cabling, the transceiver in the switch might not be provided by NetApp. You must verify that it is compatible with the transceiver in the controller.
- The switches connecting the MetroCluster nodes can carry non-MetroCluster traffic.
- Only platforms that provide dedicated ports for switchless cluster interconnects can be used with a MetroCluster-compliant switch. Platforms such as the FAS2750 and AFF A220 cannot be used because MetroCluster traffic and MetroCluster interconnect traffic share the same network ports.
- The MetroCluster-compliant switch must not be used for local cluster connections.
- The MetroCluster IP interface can be connected to any switch port that can be configured to meet the requirements.
- Four IP switches are required, two for each switch fabric. If you use directors, then you can use a single director at each side, but the MetroCluster IP interfaces must connect to two different blades in two different failure domains on that director.
- The MetroCluster interfaces from one node must connect to two network switches or blades. The MetroCluster interfaces from one node cannot be connected to the same network or switch or blade.
- The network must meet the requirements outlined in the following sections:
  - Considerations for ISLs
  - Considerations when deploying MetroCluster in shared layer 2 or layer 3 networks
- The maximum transmission unit (MTU) of 9216 must be configured on all switches that carry MetroCluster IP traffic.
- Reverting to ONTAP 9.6 or earlier is not supported.

Any intermediate switches that you use between the switches connecting the MetroCluster IP interfaces at both sites must meet the requirements and must be configured as outlined in Considerations when deploying

MetroCluster in shared layer 2 or layer 3 networks.

Limitations when using MetroCluster-compliant switches

You cannot use any configuration or feature that requires that local cluster connections are connected to a switch. For example, you cannot use the following configurations and procedures with a MetroCluster-compliant switch:

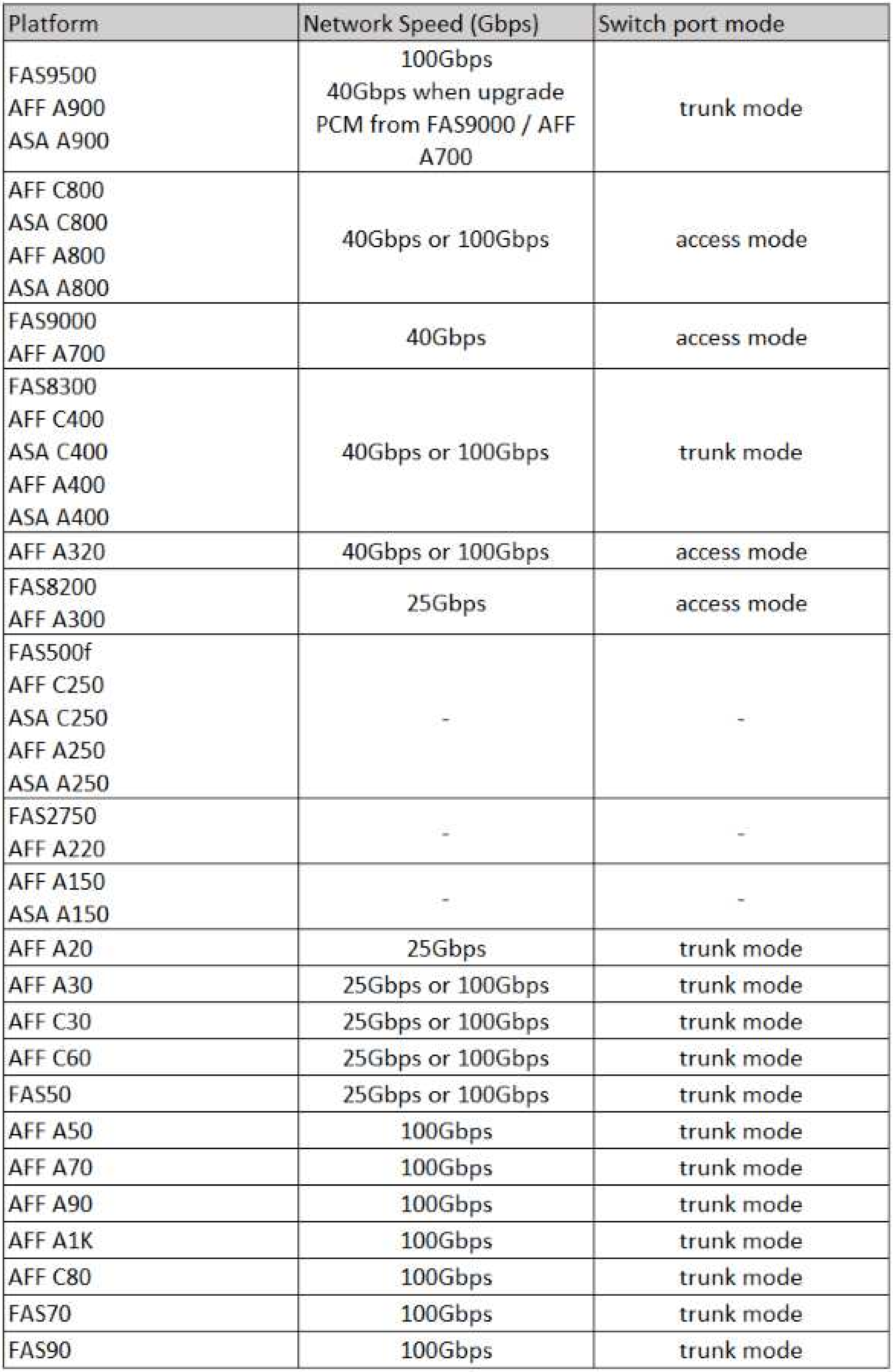
- Eight-node MetroCluster configurations
- Transitioning from MetroCluster FC to MetroCluster IP configurations
- Refreshing a four-node MetroCluster IP configuration
- Platforms sharing a physical interface for local cluster and MetroCluster traffic. Refer to Platform-specific network speeds and switch port modes for MetroCluster-compliant switches for supported speeds.

ONTAP platform-specific network speeds and switch port modes for MetroCluster-compliant switches

If you are using MetroCluster compliant switches, you should be aware of the platform-specific network speeds and switch port mode requirements.

The following table provides platform-specific network speeds and switch port modes for MetroCluster-compliant switches. You should configure the switch port mode according to the table.

- Missing values indicate that the platform cannot be used with a MetroCluster-compliant switch.
- AFF A30, AFF C30, AFF C60, and FAS50 systems require a QSFP-to-SFP+ adapter in the card on the controller to support a 25Gbps network speed.



Platform	Network Speed (Gbps)	Switch port mode
FAS9500 AFF A900 ASA A900	100Gbps 40Gbps when upgrade PCM from FAS9000 / AFF A700	trunk mode
AFF C800 ASA C800 AFF A800 ASA A800	40Gbps or 100Gbps	access mode
FAS9000 AFF A700	40Gbps	access mode
FAS8300 AFF C400 ASA C400 AFF A400 ASA A400	40Gbps or 100Gbps	trunk mode
AFF A320	40Gbps or 100Gbps	access mode
FAS8200 AFF A300	25Gbps	access mode
FAS500f AFF C250 ASA C250 AFF A250 ASA A250	-	-
FAS2750 AFF A220	-	-
AFF A150 ASA A150	-	-
AFF A20	25Gbps	trunk mode
AFF A30	25Gbps or 100Gbps	trunk mode
AFF C30	25Gbps or 100Gbps	trunk mode
AFF C60	25Gbps or 100Gbps	trunk mode
FAS50	25Gbps or 100Gbps	trunk mode
AFF A50	100Gbps	trunk mode
AFF A70	100Gbps	trunk mode
AFF A90	100Gbps	trunk mode
AFF A1K	100Gbps	trunk mode
AFF C80	100Gbps	trunk mode
FAS70	100Gbps	trunk mode
FAS90	100Gbps	trunk mode

MetroCluster IP switch configuration examples

Learn about the various switch port configurations.

> The following examples use decimal values and follow the table that applies to Cisco switches. Depending on the switch vendor, you might require different values for DSCP. Refer to the corresponding table for your switch vendor to confirm the correct value.

DSCP value	Decimal	Hex	Meaning
101 000	16	0x10	CS2
011 000	24	0x18	CS3
100 000	32	0x20	CS4
101 000	40	0x28	CS5

Switch port connecting a MetroCluster interface

- Classification for remote direct memory access (RDMA) traffic:
  - Match : TCP port 10006, source, destination, or both
  - Optional match: COS 5
  - Optional match: DSCP 40
  - Set DSCP 40
  - Set COS 5
  - Optional : rate shaping to 20Gbps
- Classification for iSCSI traffic:
  - Match : TCP port 62500, source, destination, or both
  - Optional match: COS 4
  - Optional match: DSCP 32
  - Set DSCP 32
  - Set COS 4
- L2FlowControl (pause), RX and TX

ISL ports

- Classification:
  - Match COS 5 or DSCP 40
    - Set DSCP 40
    - Set COS 5
  - Match COS 4 or DSCP 32
    - Set DSCP 32
    - Set COS 4

- Egress queuing
  - COS group 4 has a minimum configuration threshold of 2000 and a maximum threshold of 3000
  - COS group 5 has a minimum configuration threshold of 3500 and a maximum threshold of 6500.

    Configuration thresholds can vary depending on the environment. You must evaluate the configuration thresholds based on your individual environment.

  - ECN enabled for Q4 and Q5
  - RED enabled for Q4 and Q5

Bandwidth allocation (switch ports connecting MetroCluster interfaces and ISL ports)

- RDMA, COS 5 / DSCP 40: 60%
- iSCSI, COS 4 / DSCP 32: 40%
- Minimum capacity requirement per MetroCluster configuration and network: 10Gbps

If you use rate limits, the traffic should be **shaped** without introducing loss.

Examples for configuring switch ports connecting the MetroCluster controller

The example commands provided are valid for Cisco NX3232 or Cisco NX9336 switches. Commands vary according to the switch type.

If a feature or its equivalent shown in the examples is not available on the switch, the switch does not meet the minimum requirements and cannot be used to deploy a MetroCluster configuration. This is true for any switch connecting to a MetroCluster configuration and for all intermediate switches.

The following examples might only show the configuration for one network.

Basic configuration

A virtual LAN (VLAN) in each network must be configured. The following example shows how to configure a VLAN in network 10.

Example:

```
# vlan 10
The load balancing policy should be set so that order is preserved.
```

Example:

```
# port-channel load-balance src-dst ip-l4port-vlan
```

Examples for configuring classification

You must configure access and class maps to map RDMA and iSCSI traffic to the appropriate classes.

In the following example, all TCP traffic to and from the port 65200 is mapped to the storage (iSCSI) class. All TCP traffic to and from the port 10006 is mapped to the RDMA class. These policy-maps are used on switch ports connecting the MetroCluster interfaces.

Example:

```
ip access-list storage
  10 permit tcp any eq 65200 any
  20 permit tcp any any eq 65200
ip access-list rdma
  10 permit tcp any eq 10006 any
  20 permit tcp any any eq 10006

class-map type qos match-all storage
  match access-group name storage
class-map type qos match-all rdma
match access-group name rdma
```

You must configure an ingress policy. An ingress policy maps the traffic as classified to different COS groups. In this example, the RDMA traffic is mapped to COS group 5 and iSCSI traffic is mapped to COS group 4. The ingress policy is used on switch ports connecting the MetroCluster interfaces and on the ISL ports carrying MetroCluster traffic.

Example:

```
policy-map type qos MetroClusterIP_Node_Ingress
class rdma
  set dscp 40
  set cos 5
  set qos-group 5
class storage
  set dscp 32
  set cos 4
  set qos-group 4
```

NetApp recommends that you shape traffic on switch ports connecting a MetroCluster interface, as shown in the following example:

Example:

```
policy-map type queuing MetroClusterIP_Node_Egress
class type queuing c-out-8q-q7
  priority level 1
class type queuing c-out-8q-q6
  priority level 2
class type queuing c-out-8q-q5
  priority level 3
  shape min 0 gbps max 20 gbps
class type queuing c-out-8q-q4
  priority level 4
class type queuing c-out-8q-q3
  priority level 5
class type queuing c-out-8q-q2
  priority level 6
class type queuing c-out-8q-q1
  priority level 7
class type queuing c-out-8q-q-default
  bandwidth remaining percent 100
  random-detect threshold burst-optimized ecn
```

Examples for configuring the node ports

You might need to configure a node port in breakout mode. In the following example, ports 25 and 26 are configured in 4 x 25Gbps breakout mode.

Example:

```
interface breakout module 1 port 25-26 map 25g-4x
```

You might need to configure the MetroCluster interface port speed. The following example shows how to configure the speed to **auto** or into 40Gbps mode:

Example:

```
    speed auto

    speed 40000
```

The following example shows a switch port configured to connect a MetroCluster interface. It is an access mode port in VLAN 10, with an MTU of 9216 and is operating in native speed. It has symmetric (send and receive) flow control (pause) enabled and the MetroCluster ingress and egress policies assigned.

Example:

```
interface eth1/9
description MetroCluster-IP Node Port
speed auto
switchport access vlan 10
spanning-tree port type edge
spanning-tree bpduguard enable
mtu 9216
flowcontrol receive on
flowcontrol send on
service-policy type qos input MetroClusterIP_Node_Ingress
service-policy type queuing output MetroClusterIP_Node_Egress
no shutdown
```

On 25Gbps ports, you might need to set the Forward Error Correction (FEC) setting to "off", as shown in the following example.

Example:

```
fec off
```

Examples of configuration of ISL ports throughout the network

A MetroCluster-compliant switch is regarded as an intermediate switch, even it directly connects the MetroCluster interfaces. The ISL ports carrying MetroCluster traffic on the MetroCluster-compliant switch must be configured the same way as the ISL ports on an intermediate switch. Refer to Required settings on intermediate switches for guidance and examples.

> Some policy maps are the same for switch ports connecting MetroCluster interfaces and ISLs carrying MetroCluster traffic. You can use the same policy map for both of these port usages.

Learn about unmirrored aggregates in MetroCluster IP configurations

If your configuration includes unmirrored aggregates, you must be aware of potential access issues after switchover operations.

Unmirrored aggregates and hierarchical namespaces

If you are using hierarchical namespaces, you should configure the junction path so that all of the volumes in that path are either on mirrored aggregates only or on unmirrored aggregates only. Configuring a mix of unmirrored and mirrored aggregates in the junction path might prevent access to the unmirrored aggregates after the switchover operation.

Unmirrored aggregates and maintenance that requires power shutdown

If you perform a negotiated switchover for maintenance that requires a site-wide power shutdown, you should first manually offline any unmirrored aggregates owned by the disaster site.

If you don’t offline the unmirrored aggregates owned by the disaster site, nodes at the surviving site might go down due to multi-disk panics. This might occur if switched-over unmirrored aggregates go offline or are missing because of the loss of connectivity to storage at the disaster site if there’s a power shutdown or loss of ISLs.

Unmirrored aggregates, CRS metadata volumes, and data SVM root volumes

The configuration replication service (CRS) metadata volume and data SVM root volumes must be on a mirrored aggregate. You cannot move these volumes to an unmirrored aggregate. If they are on an unmirrored aggregate, negotiated switchover and switchback operations are vetoed and the `metrocluster check` command returns a warning.

Unmirrored aggregates and SVMs

You should configure SVMs on mirrored aggregates only or on unmirrored aggregates only. Configuring SVMs on a mix of both unmirrored and mirrored aggregates can result in a switchover operation that exceeds 120 seconds. This can lead to a data outage if the unmirrored aggregates don’t come online.

Unmirrored aggregates and SAN

Before ONTAP 9.9.1, a LUN should not be located on an unmirrored aggregate. Configuring a LUN on an unmirrored aggregate can result in a switchover operation that exceeds 120 seconds and a data outage.

Add storage shelves for unmirrored aggregates

If you add shelves and want to use them for unmirrored aggregates in a MetroCluster IP configuration, you must do the following:

1. Before starting the procedure to add the shelves, issue the following command:

   ```
   metrocluster modify -enable-unmirrored-aggr-deployment true
   ```

2. Verify that automatic disk assignment is off:

   ```
   disk option show
   ```

3. Follow the steps of the procedure to add the shelf.
4. Manually assign all disks from new shelf to the node that will own the unmirrored aggregate or aggregates.
5. Create the aggregates:

   ```
   storage aggregate create
   ```

6. After completing the procedure, issue the following command:

   ```
   metrocluster modify -enable-unmirrored-aggr-deployment false
   ```

7. Verify that automatic disk assignment is enabled:

   ```
   disk option show
   ```

Firewall port requirements for MetroCluster IP configurations

If you are using a firewall at a MetroCluster site, you must ensure access for certain required ports.

Considerations for firewall usage at MetroCluster sites

If you are using a firewall at a MetroCluster site, you must ensure access for required ports.

The following table shows TCP/UDP port usage in an external firewall positioned between two MetroCluster sites.

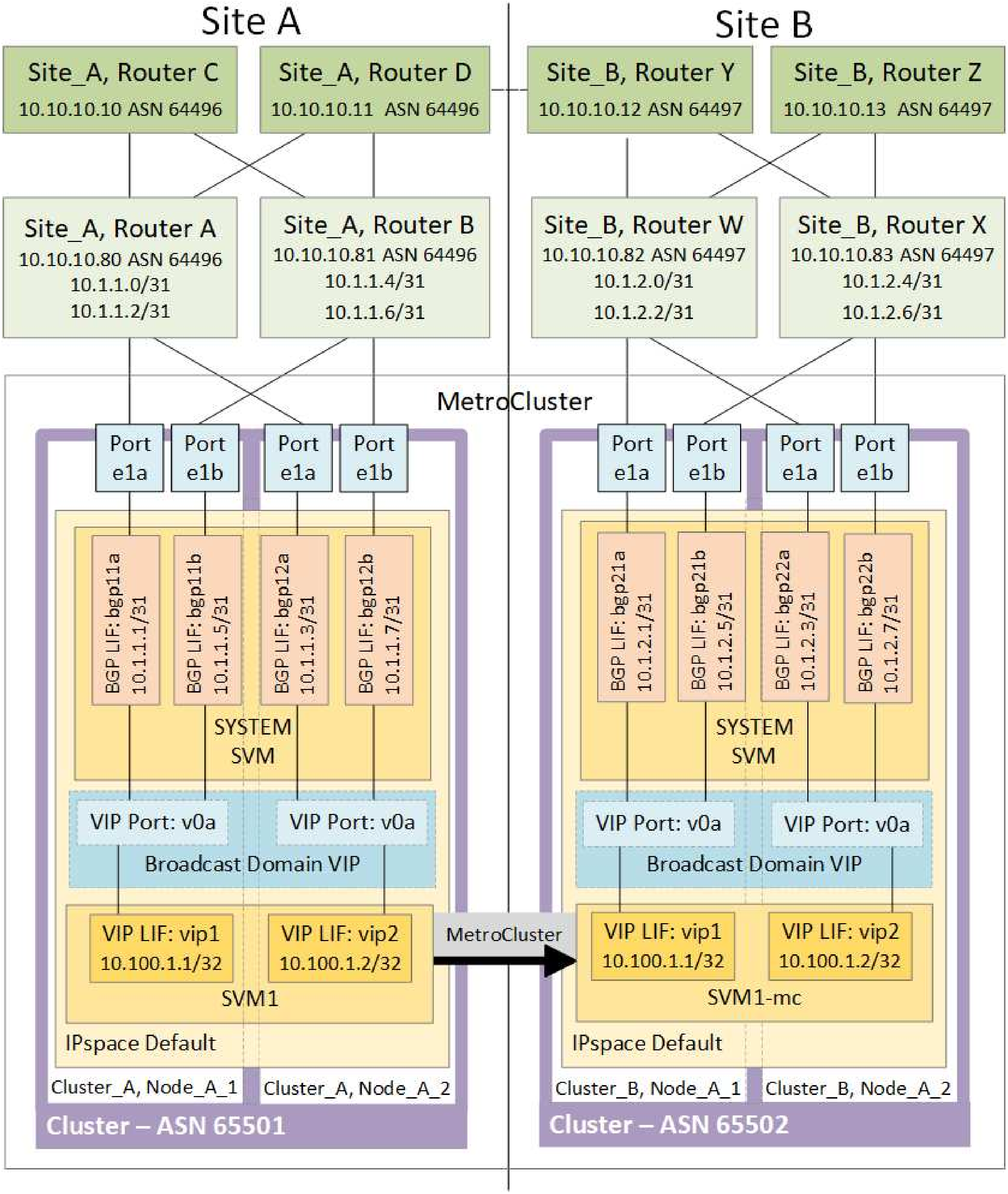
Traffic type	Port/services
Cluster peering	11104 / TCP 11105 / TCP
ONTAP System Manager	443 / TCP
MetroCluster IP intercluster LIFs	65200 / TCP 10006 / TCP and UDP
Hardware assist	4444 / TCP

Learn about using virtual IP and Border Gateway Protocol with a MetroCluster IP configuration

Beginning with ONTAP 9.5, ONTAP supports layer 3 connectivity using virtual IP (VIP) and Border Gateway Protocol (BGP). The combination VIP and BGP for redundancy in the front-end networking with the back-end MetroCluster redundancy provides a layer 3 disaster recovery solution.

Review the following guidelines and illustration when planning your layer 3 solution. For details on implementing VIP and BGP in ONTAP, refer to the following section:

Configuring virtual IP (VIP) LIFs



ONTAP limitations

ONTAP does not automatically verify that all nodes on both sites of the MetroCluster configuration are configured with BGP peering.

ONTAP does not perform route aggregation but announces all individual virtual LIF IPs as unique host routes at all times.

ONTAP does not support true AnyCast — only a single node in the cluster presents a specific virtual LIF IP (but is accepted by all physical interfaces, regardless of whether they are BGP LIFs, provided the physical port is part of the correct IPspace). Different LIFs can migrate independently of each other to different hosting nodes.

Guidelines for using this Layer 3 solution with a MetroCluster configuration

You must configure your BGP and VIP correctly to provide the required redundancy.

Simpler deployment scenarios are preferred over more complex architectures (for example, a BGP peering router is reachable across an intermediate, non-BGP router). However, ONTAP does not enforce network design or topology restrictions.

VIP LIFs only cover the frontend/data network.

Depending on your version of ONTAP, you must configure BGP peering LIFs in the node SVM, not the system or data SVM. In 9.8, the BGP LIFs are visible in the cluster (system) SVM and the node SVMs are no longer present.

Each data SVM requires the configuration of all potential first hop gateway addresses (typically, the BGP router peering IP address), so that the return data path is available if a LIF migration or MetroCluster failover occurs.

BGP LIFs are node specific, similar to intercluster LIFs — each node has a unique configuration, which does not need to be replicated to DR site nodes.

The existence of the v0a (v0b and so on) continuously validates the connectivity, guaranteeing that a LIF migrate or failover succeeds (unlike L2, where a broken configuration is only visible after the outage).

A major architectural difference is that clients should no longer share the same IP subnet as the VIP of data SVMs. An L3 router with appropriate enterprise grade resiliency and redundancy features enabled (for example, VRRP/HSRP) should be on the path between storage and clients for the VIP to operate correctly.

The reliable update process of BGP allows for smoother LIF migrations because they are marginally faster and have a lower chance of interruption to some clients

You can configure BGP to detect some classes of network or switch misbehaviors faster than LACP, if configured accordingly.

External BGP (EBGP) uses different AS numbers between ONTAP node(s) and peering routers and is the preferred deployment to ease route aggregation and redistribution on the routers. Internal BGP (IBGP) and the use of route reflectors is not impossible but outside the scope of a straightforward VIP setup.

After deployment, you must check that the data SVM is accessible when the associated virtual LIF is migrated between all nodes on each site (including MetroCluster switchover) to verify the correct configuration of the static routes to the same data SVM.

VIP works for most IP-based protocols (NFS, SMB, iSCSI).

Copyright information

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

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

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

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

Trademark information