nexenta™

Global Leader in Software Defined Storage.

MetroHA Administration Guide

Date: May, 2016

Subject: MetroHA Administration Guide

Software: NexentaStor

Software Version: 4.0.4

Part Number: 3000-nxsmetro-000001-A

www.nexenta.com

Nexenta reserves the right to make changes to this document at any time without notice and assumes no responsibility for its use. Nexenta products and services only can be ordered under the terms and conditions of Nexenta Systems’ applicable agreements. All of the features described in this document may not be available currently. Refer to the latest product announcement or contact your local Nexenta Systems sales office for information on feature and product availability. This document includes the latest information available at the time of publication.

This document applies to the following product versions:

Product	Versions supported
NexentaStor	4.0.4-FP5 or later
ATTO FibreBridge 6500	1.18 or later

Contents

Preface

This documentation presents information specific to Nexenta products. The information is for reference purposes and is subject to change.

About this Administration Guide

This guide is intended for Nexenta partners or end-users who may need to administer an existing MetroHA configuration. Administrative tasks for MetroHA include pool failover/failback, device replacement, and troubleshooting.

Typically, the initial installation and configuration of MetroHA is done by Nexenta personnel. Consequently, installation and configuration procedures are not covered in this document. For information on MetroHA installation and deployment, contact Nexenta Systems.

This documentation assumes that you have a working knowledge of UNIX. It also assumes that you have experience with data storage concepts, such as object storage, ZFS, iSCSI, NFS, CIFS, and so on.

Documentation History

The following table lists the released revisions of this documentation.

Table 1: Documentation Revision History

Revision	Date	Description
3000-nxsmetro-000001-A	May, 2016	GA

Comments

Your comments and suggestions to improve this documentation are greatly appreciated. Send any feedback to doc.comments@nexenta.com and include the documentation title, number, and revision. Refer to specific pages, sections, and paragraphs whenever possible.

MetroHA Overview

This chapter includes the following topics:

What is MetroHA?

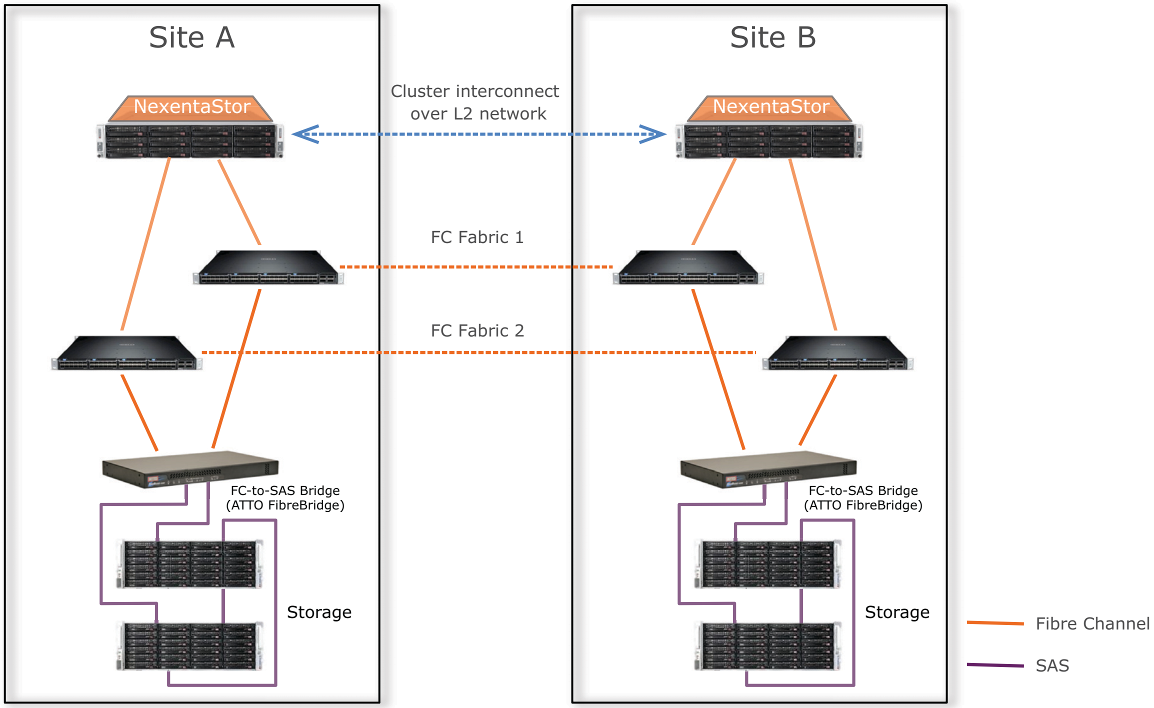
Nexenta Systems' Metro High Availability (MetroHA) is a solution that builds upon the NexentaStor HA Cluster feature to provide synchronous mirroring of data between two geographically remote sites.

This solution provides for single-instance mirroring of a data pool, with the mirrors separated by metro-wide distances (up to 50 miles / 80 kilometers, depending on inter-site latency) for storage redundancy, high availability, and disaster recovery, with minimal RTO/RPO. Since the data is mirrored across remote sites, it remains accessible to mission-critical applications even in the event of a data center failure or regional disruption.

MetroHA delivers this functionality through a combination of Nexenta products, including NexentaStor Enterprise Edition and the RSF-1 cluster plug-in. The MetroHA solution relies on a stretched HA cluster of NexentaStor head nodes (one per site) connected via Fibre Channel (FC) to an ATTO FibreBridge that bridges the FC to Serial Attached SCSI (SAS). The ATTO FibreBridge is connected via SAS to storage enclosures at each site.

Figure 1-1 shows a high-level illustration of the components in a MetroHA configuration. At a basic level, a MetroHA configuration consists of a set of NexentaStor head nodes, cascaded storage units, and FC-to-SAS bridges, deployed at sites connected via a stretched FC link.

Figure 1-1: Nexenta MetroHA Deployment


Site A
Site B
NexentaStor
NexentaStor
Cluster interconnect over L2 network
FC Fabric 1
FC Fabric 2
FC-to-SAS Bridge (ATTO FibreBridge)
FC-to-SAS Bridge (ATTO FibreBridge)
Storage
Storage
Fibre Channel
SAS

Key features of MetroHA include the following:

- Disk-level synchronous mirroring – Data at each site in a MetroHA configuration is constantly and seamlessly mirrored across the Fibre link. MetroHA makes use of four-way mirrors to provide data protection. Two devices from each four-way mirror are located in each data center, allowing for both local data center and complete data center disk component failure. All I/Os are checksummed, and the mirrors are self-healing.
- Automatic failover for high availability (HA) – MetroHA ensures continuity in the presence of service-level exceptional events, such as power outages or link failures. When MetroHA detects that the NexentaStor node is not functioning on one of the sites, it can automatically fail over storage services to the alternate node. This is made possible by the four-way synchronous disk mirroring layout for all storage pools.
- Disaster recovery (DR) – In the event of a major failure at one of the sites, such as the site going offline entirely, MetroHA makes it possible to make the storage services available from the other site. Unlike HA, DR is a manually triggered process; however, the actual data recovery process is fully automated.

Components of a MetroHA Configuration

Figure 1-1 above shows a high-level illustration of the components in a MetroHA configuration. At a basic level, a MetroHA configuration consists of a set of NexentaStor head nodes, SAS-connected storage enclosures, and FC-to-SAS bridges, deployed at sites connected via a stretched Fibre Channel link.

Contact Nexenta Systems for specific hardware that has been certified to work in a MetroHA configuration.

NexentaStor Head Nodes

The head nodes are servers running NexentaStor Enterprise Edition. The head nodes present the storage services to clients. There is a head node at each site in a MetroHA configuration. The two head nodes exchange heartbeat information over both storage and IP. An interruption in the heartbeat from one of the head nodes may indicate a failure condition and signal the other head node to take over from the failed node.

The NexentaStor nodes in a MetroHA configuration are running NexentaStor Enterprise edition version 4.0.4. The RSF-1 plug-in is required on both nodes. Fibre Channel HBAs are installed on the nodes to connect to the FC fabric(s).

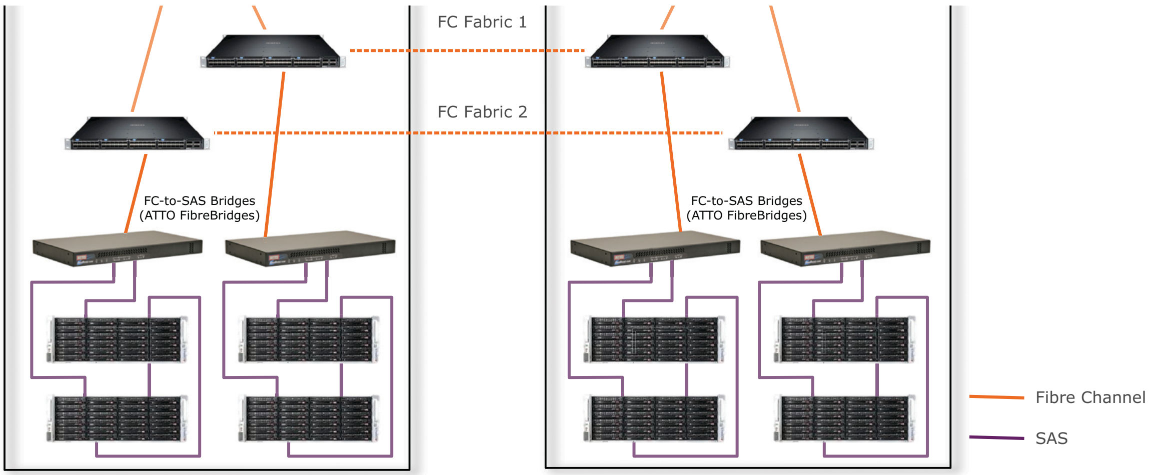
FC-to-SAS Bridges

The FC-to-SAS bridges allow SAS-connected storage enclosures to be connected to an FC fabric, providing the connection between the head nodes and the storage enclosures.

ATTO FibreBridges are used as the FC-to-SAS bridges in a MetroHA configuration. MetroHA has been tested with ATTO FibreBridge model 6500. A single ATTO FibreBridge 6500 can provide up to 75,000 IOPS. The ATTO FibreBridge 6500 functions as a bridge between SAS and FC topologies. It does so by taking SAS targets and presenting them as logical units (LUNs) behind a SCSI over Fibre Channel (FCP) target port. NexentaStor is still aware of the devices and presents them in terms of their World Wide Name (WWN), but it accesses them in terms of LUN numbers and Fibre Channel ports rather than a SAS port WWN.

At least two ATTO FibreBridges are required in a MetroHA deployment, one at each site. Figure 1-1 shows a two-FibreBridge deployment. For redundancy, MetroHA can also be deployed with four ATTO FibreBridges, two at each site (see Figure 1-2 below).

Figure 1-2: MetroHA Deployment with Four FC-to SAS Bridges



SAS-Connected Storage Enclosures

The storage enclosures, connected via SAS to the FC-to-SAS bridges, provide the actual storage in the MetroHA configuration. Multiple storage enclosures may be multipath cascaded to provide expansion capability. Nexenta recommends a minimum of two storage enclosures at each site to provide full data redundancy at a disk level.

See the *Hardware Compatibility List* for additional information about enclosures and devices supported for use with ATTO FibreBridge 6500s. See SAS Cable Connections for MetroHA for diagrams of how the SAS cables should be connected between the devices.

Four-Way Mirrors

The volume service in a MetroHA cluster is configured as a four-way mirror. The volume is constructed from four disk devices, with two devices each located in two storage enclosures at the local site and two storage enclosures at the remote site. In the event of a single storage enclosure failure, data remains available from the other local storage enclosure. In the event of a site failure, where both local storage enclosures are not available, clients can access data stored on the volume from the remaining mirrors on the remote site.

Stretched Fibre Link

To ensure synchronous mirroring of data, there must be a stretched fibre link between the remote sites for MetroHA traffic. The stretched fibre link ensures a reliable, predictable, low-latency connection between sites.

The primary requirement is that the link be sufficient to limit the round-trip delay (RTD) time between the sites to no more than 1 millisecond. In practice, this means that the maximum length of the stretched fibre link is approximately 80 kilometers. Lower latency between sites generally provides better overall performance.

The stretched fibre link can be described in terms of an Ethernet stretch and an FC stretch, as described below.

Ethernet Stretch

The Ethernet stretch applies to the IP-based data services (SMB, NFS, and iSCSI) and must be a contiguous segment with a VLAN spanning, not two subnets on different VLANs or segments. Heartbeat traffic can be routed, but Nexenta recommends using an isolated stretch VLAN dedicated to this purpose. Support for SSH binding between the nodes is required. This prevents data plane traffic from interfering with cluster communication at either level.

FC Stretch

The FC stretch between the sites in a MetroHA configuration (that is, the inter-site link or ISL) may be a series of fibre strands or DWDM channels.

Separate FC target ports are required for COMSTAR FC support. FC zoning should not allow visibility between target and initiator ports.

See the *Hardware Compatibility List* for specific requirements (zoning, buffer credits) for FC switches and fabrics.

Fault Tolerance for MetroHA Solution Components

MetroHA is designed to be a robust solution for both HA and DR. However, MetroHA works best when fault tolerance is taken into consideration for the components that make up the solution. Ideally there should be enough redundancy built in to the solution that a condition that requires failover is unlikely to occur.

To improve fault tolerance for individual MetroHA components, Nexenta recommends the following:

- Redundant links between NexentaStor nodes to ensure the flow of heartbeat traffic

  LACP or IPMP can be used for link redundancy if Ethernet/IP is used to present data services to clients.

- Dual FC fabrics
- Cascaded storage enclosures

  When two storage enclosures are connected to a FibreBridge, each storage enclosure must cascade through the other to provide a second path. See Figure 2-2 for an illustration.

- Full-path redundancy for inter-site links

  Inter-site links be redundant and independent, such that they are truly separate failure domains. For example, having redundant fibres in the same cable conduits still means that these cables can be lost at the same time. Concurrent or cumulative loss of inter-site links creates a site partition, which disrupts MetroHA's automatic failure capabilities and cross-site mirroring.

MetroHA Failover Scenarios

Table 1-1 describes how MetroHA handles various scenarios to facilitate high availability, failover, and disaster recovery.

The table describes how MetroHA deals with fundamental failure and recovery scenarios. Note that all of these scenarios may combine with previous or subsequent faults or failures to create more complex problems, resulting in outcomes other than those described in the table. For example, if a standby node does not have Ethernet redundancy, a cable or switch failure causes failover to abort, leaving the service in a broken/unsafe state.

Consequently, Nexenta advises following fault tolerance recommendations in the previous section and monitoring the MetroHA deployment for faults that may individually or cumulatively compromise failover and mirroring capabilities.

Table 1-1: MetroHA Failure/Recovery Scenarios

Failure/Fault Event	MetroHA Behavior
Single node fails	The other node automatically detects the failure and fails over the storage services that were provided by the failed node. Subsequent recovery of the affected node requires manual intervention to fail back storage services to it.
Single storage component fails (SSD, HDD, storage enclosure, ATTO FibreBridge)	The pool configurations for MetroHA are resilient to single component failures. In the worst case, all vdevs lose half of their mirrors. Though some or all vdevs end up in degraded mode, storage services are not affected and continue to be served through their respective nodes.
Single node hangs, then recovers	If a node becomes completely unresponsive, the other node detects it as a failed node and fails over services as necessary. When the hung node becomes responsive again, it detects that it has lost reservations on the service's storage and panics itself on the resulting reservation conflict. On reboot after panic, the node rejoins the cluster and resumes normal operation. Services must be failed back manually if required. If no services are running on the node when it hangs, there is no impact on the cluster.
Inter-site link fails	If the inter-site links are redundant and independent, it creates a risk of site partition; if the inter-site links are not redundant and independent, it creates a site partition.
Inter-site cluster heartbeat fails, but not the inter-site FC fabric link	The cluster continues normal operation and can continue to detect failover, but subsequent loss of remaining inter-site communication results in site partition.
Inter-site FC fabric link fails, but not the inter-site cluster heartbeat	
All inter-site links fail, causing a site partition	If a site partition occurs, MetroHA detects that it has lost both its partner node and its storage. On detecting these conditions, failover is aborted, requiring administrative intervention until the site partition is healed. Each pool loses half of its mirrors, but each site continues to operate and provide its set of services independently. After the inter-site links are restored, manual intervention is required to resync and recover the four-way mirrors for the affected storage pools.

Table 1-1: MetroHA Failure/Recovery Scenarios (Continued)

Failure/Fault Event	MetroHA Behavior
Site fails	MetroHA does not distinguish a site failure from a site partition, so it does not automate failover in case a site fails. An operator must determine that the situation is a site loss rather than a partition, and mark services that need to be recovered as safe for failover, since failover aborts upon detecting a site partition. Following a site failure, all vdevs in all pools lose half of their mirrors, disrupting cross-site replication. They are available in degraded mode and can continue to support all storage services.
Loss of link for the service interface	Failover takes place automatically when loss of link for the service interface is detected. If COMSTAR FC is running, FC port monitoring should be enabled, which also triggers failover when link loss is detected.
Loss of access to pool devices	Loss of access to pool devices equivalent to loss of the parent vdev on an active node triggers panic upon failure of an operation that needs to touch the device; for example, writes and uncached reads.

If MetroHA loses contact with both the partner node and its half of the storage, it requires operator intervention to determine whether the event is a site partition or a site failure. In case an operator incorrectly determines that a site has failed when it has in fact been partitioned, a *split brain* condition results, in which the two sides of the cluster continue without coordination, which can allow diverging and irreconcilable updates to each side, possibly leading to data loss or corruption.

2

Administering MetroHA

This chapter includes the following topics:

- Basic MetroHA Cluster Maintenance
- Administering the MetroHA Cluster
- Service Modes and States for the HA Cluster Plug-In

Basic MetroHA Cluster Maintenance

Installation and initial configuration of MetroHA are typically done by Nexenta Professional Services personnel. Following installation and configuration, there are administrative tasks related to MetroHA that you may need to undertake independently. These tasks include the following:

- Confirming Cluster SSH Bindings
- Adding New HA Cluster Plug-In Licenses
- Modifying Heartbeat Properties to reflect changes to your network environment
- Considerations for Managing Block Target Services in a MetroHA environment, including additional licensing requirements for FC target mode, ALUA support, management of host and initiator groups, and defining storage views

Confirming Cluster SSH Bindings

MetroHA uses three distinct types of interconnects for clustering:

- SSH bindings, which allow the appliance layer to communicate across the cluster
- Network heartbeats
- Volume heartbeats

The HA Cluster plug-in's user interface and general operation depends on an SSH binding being established. Nexenta Professional Services will configure SSH bindings; end-users only need to confirm them.

❖ *To confirm SSH bindings, using NMC:*

1. Type:

```
nmc:/$ show network ssh-bindings
```

System response:

```
HOST                         PINGABLE  SSH-ACCESSIBLE  IS-APPLIANCE
root@metronx02               Yes       Yes             Yes
```

The command displays the following output:

`PINGABLE`	Indicates whether the host is accessible via ICMP.
`SSH-ACCESSIBLE`	Indicates that the bound node responds and accepts requests via the SSH bind.
`IS-APPLIANCE`	Indicates whether the other side responds as a NexentaStor system.

2. Check the following for each node:

- If a node is down, all three columns indicate `No`.
- If a node is up and fully functional, all three columns indicate `Yes`.
- If a node is up but not indicating `Yes` in all three columns, confirm that there are no issues with connectivity in your network environment. If you have eliminated your network or if you have made networking changes and are still having problems, contact Nexenta Support.

Adding New HA Cluster Plug-In Licenses

The HA Cluster plug-in is a separately licensed feature of MetroHA that comes pre-installed with NexentaStor. You must have a valid HA Cluster license, which may be either a temporary trial license for pre-production use or a permanent license for production use.

❖ *To register the HA Cluster, using NMV:*

1. Click **Settings > HA Cluster**.
2. Accept the HA Cluster license agreement.
3. If you see a warning that the HA Cluster is not running, click **Start RSF**.

   NexentaStor starts the HA Cluster daemon.
4. Request an HA Cluster license key or type an existing one.

   When prompted, select one of the following options:

   - If you have an Internet connection:

     1)Click **Confirm**.

Note:	If the other HA Cluster node is unlicensed, NMV notifies you about that. Click **Yes** to install the HA license to the other node.

     2)When prompted, accept the HA Cluster license agreement.

     3)Type your e-mail address.

     The temporary HA Cluster license automatically registers on your NexentaStor appliance. The license package is also sent to the e-mail address you provided.

     4)When prompted, click **OK**.

   - If you do not have an access to the Internet:

     1)Click **Manual**.

2)Select an HA node.

3)Type the HA Cluster license.

You must have a hard copy of the HA Cluster license to use this functionality.

4)Click **OK**.

5)Repeat Step 1 - Step 4 for the second node.

You can install a temporary 45-days trial license.

5. Alternatively, you can type a permanent HA Cluster license.

Modifying Heartbeat Properties

Heartbeat properties are defined at installation, but may need to be changed if a device providing a storage heartbeat is replaced or in case of networking changes.

Note: A NexentaStor node can be part of only one cluster at a time.

❖ *To change heartbeat properties, using NMV:*

1. Click **Settings > HA Cluster**.
2. In **Cluster Settings**, click **Heartbeats**.

Adding Volume Heartbeats

Volume heartbeats are storage devices that are all part of a single virtual device (vdev) for a volume. At installation, all devices from the first vdev are defined as volume heartbeats. If any of these devices are subsequently replaced, the replacement device must be defined as a volume heartbeat, since this does not happen automatically as part of other device replacement procedures.

❖ *To add a volume heartbeat, using NMV:*

1. Click the **Volume Heartbeats** tab.
2. Right click on a disk and select a vdev.
3. Click **Save Settings**.

Adding Network Heartbeats

Network heartbeats allow the cluster control process on each node to communicate with the partner node using IP networking. Nexenta recommends using a dedicated, non-routed VLAN, spanned across sites that can also support the SSH binding between the nodes.

Note: Normally, you should not need to change network heartbeats. If you are planning substantial changes to your network environment, contact Nexenta Professional Services.

❖ *To add a network heartbeat, using NMV:*

1. Click the **Appliance Heartbeats** tab.

2. Right click on **Network heartbeats**, and select **Add a network heartbeat**.
3. In the **Create network heartbeat** dialog, type the IP address or a hostname available on a remote NexentaStor appliance.
4. Optionally, click **Test**.
5. Click **OK**.
6. Click **Save Settings**.

❖ *To change heartbeat properties, using NMC:*

1. Type:

   ```
   nmc:/$ setup group rsf-cluster <cluster_name> hb_properties
   ```

   System response:

   ```
   • Enable inter-appliance heartbeat through primary interfaces?:
   ```

2. Follow the on-screen instructions.

Considerations for Managing Block Target Services

MetroHA has specific considerations for block target protocols. The following sections describe protocol issues specific to Fibre Channel and iSCSI targets. Note that Fibre Channel and iSCSI target services cannot be provided by the same cluster, because they have conflicting configuration requirements for Asymmetric LUN Access (ALUA) support.

- Configuring iSCSI Targets for Failover
- Configuring Fibre Channel Targets for Failover

Configuring iSCSI Targets for Failover

You can use HA Cluster to fail over iSCSI volumes from one cluster node to another. The target IQN moves as part of the failover.

Setting up iSCSI failover involves setting up a zvol in the shared volume.

Note:	Note that you perform the process of creating a zvol and sharing it through iSCSI separately from the HA Cluster configuration.

If you create iSCSI zvols before marking the zvol’s volume as a shared cluster volume, then when you share the cluster volume as an active iSCSI session, it may experience some delays. Depending on the network, application environment and active workload, you may also see command level failures or disconnects during this period.

When you add a shared volume to a cluster that uses zvols as the backing store for iSCSI target LUNs, it is vital that you configure all client iSCSI initiators, regardless of the operating system, to access those targets using the shared logical hostname that is specified when the volume service was created, rather than a real hostname associated with one of the appliances.

Note that the cluster manages all aspects of the shared logical hostname configuration. Therefore, do not configure the shared logical hostname manually. Furthermore, unless the shared volume service is running, the shared logical hostname is not present on the network.

❖ *To configure iSCSI targets on the active appliance, using NMV:*

1. Click **Data Management > SCSI Target.**
2. In the zvols panel, click **create**.
3. Make the virtual block device > 200MB.

   HA Cluster automatically migrates the newly created zvol to the other appliance on failover. Therefore, you do not have to duplicate it manually.
4. From the iSCSI pane, click **iSCSI > Target Portal Groups** and define a target portal group.

Note:	It is critical that the IPv4 portal address is the shared logical hostname specified when the volume service was created, instead of a real hostname associated with one of the appliances.

HA Cluster automatically replicates the newly created target portal group to the other appliance.

❖ *To create an iSCSI target and add it to the target portal group, using NMV:*

1. Click **iSCSI > Targets**.

   This limits zvol visibility from initiators to the target portal group. The newly created iSCSI target is automatically replicated to the other appliance.
2. Type a name and an alias.

   The newly created iSCSI target displays in the Targets page.

❖ *To create a LUN mapping to the zvol, using NMV:*

1. From the SCSI Target pane, click **Mappings.**

   This creates a LUN mapping to the zvol for use as the backing store for the iSCSI target. The newly created LUN mapping is automatically migrated to the other appliance on failover.
2. On the client, configure the iSCSI initiator to use both the IQN of the iSCSI target created and the shared logical hostname associated with both the volume service and the target portal group to access the zvol through iSCSI.

Configuring Fibre Channel Targets for Failover

To configure the Fibre Channel targets for HA Cluster failover, you complete the following tasks:

- Setting the HA Cluster to ALUA Mode
- Changing the HBA Port Mode
- Creating a Target Group
- Adding WWNs to an Initiator Group
- Creating a Zvol

- Mapping a Zvol

Note:	Using MetroHA as a Fibre Channel target requires the separately licensed FC target plug-in.

Setting the HA Cluster to ALUA Mode

For IP-based storage services, the cluster maintains a separate network identity or series of identities using a virtual IP address that either node can own. In contrast, Fibre Channel target mode is provided in terms of target port World Wide Names (WWNs), which are hardware addresses belonging to the HBAs on each node. Asymmetric LUN Access (ALUA) allows initiators to be aware of both nodes and to communicate with them to understand which WWN, if any, owns target services (`ACTIVE` state) and to see an inactive node as a standby (`STANDBY` state) that can nevertheless proxy I/Os to the `ACTIVE` node if there is one.

COMSTAR presents the exported LUNs on the second node in `STANDBY` state. This is similar to having the same LUN presented via two different IP addresses for iSCSI. Normal running mode is to have the first head doing the FC work with the FC port with the LUNs in `ACTIVE` mode on it, and the second NexentaStor node with the same LUNs in `STANDBY` mode. In case of a failover, the `ACTIVE` port becomes `STANDBY` and the `STANDBY` becomes `ACTIVE`.

ALUA support is disabled by default, since iSCSI does not support it, so it must be enabled for FC target support.

Warning:	Before you configure ALUA, verify that you do not have iSCSI targets configured on both HA Cluster nodes.

❖ *To set the HA Cluster to ALUA mode:*

1. Log in to an HA Cluster node.
2. Click **Settings > HA Cluster**.
3. Select Advanced > **GLobal Cluster Properties**.
4. Select the **Enable ALUA mode** checkbox.

Changing the HBA Port Mode

For MetroHA, one set of Fibre Channel HBA ports should be configured in initiator mode for access to the storage used to provide services. When FC target mode is used, an additional separate set of HBA ports must be configured in target mode.

Fibre Channel switch zoning must be configured so that target and initiator ports do not see one another. Setting up target mode ports should generally be done by Nexenta Professional Services in initial deployment, but HBA ports will need to be configured to target mode after HBA hardware replacement.

❖ *To change the HBA port mode, using NMV:*

1. Click **Data Management > SCSI Target Plus**
2. **Select Fibre Channel > Ports.**
3. Select **Target** from the Mode dropdown menu.
4. Once you change the HBA port modes of both appliances from Initiator mode to Target mode, reboot both appliances so the Target mode changes can take effect.

Creating a Target Group

To use ALUA, create an FC target using the desired FC ports from both nodes. Failure to create a target group with FC ports from both nodes may result in the inability of an initiator to maintain access to the storage following a failover event.

Target FC ports are identified as:

- **Local**

  An FC port that resides on the node that you are configuring.

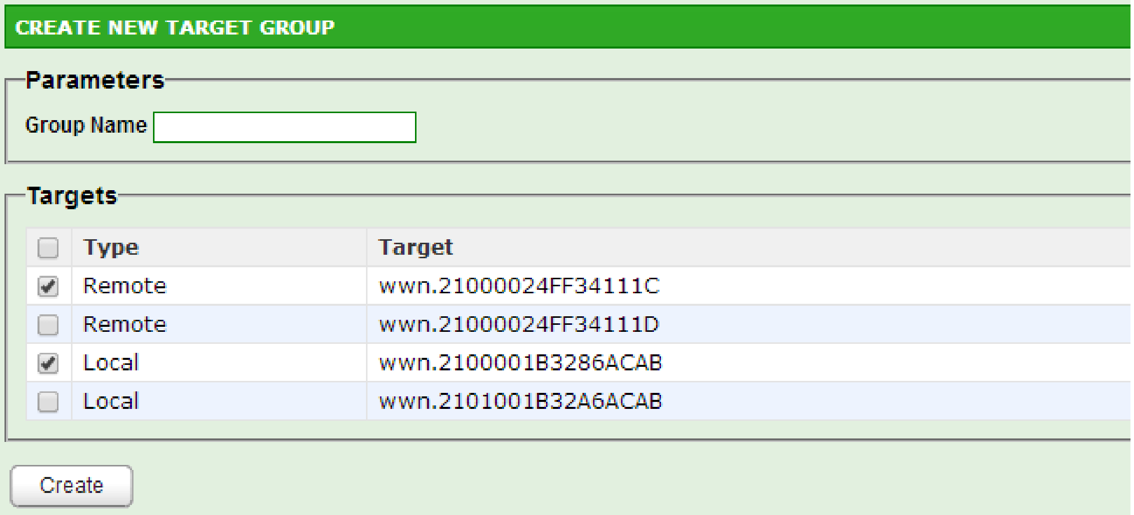
- **Remote**

  An FC port that resides on the second node in the same cluster.

To ensure proper failover, a Target Group must have at least one local and one remote port defined within it.

❖ *To create a target group, using NMV:*

1. Click **Data Management > SCSI Target Plus**.
2. In the SCSI Target panel, click **Target groups**.
3. Click **Create** or **here**.
4. In the **Group Name** field, type the name of the target group.
5. Select at least one local and one remote FC ports.



CREATE NEW TARGET GROUP

Parameters

Group Name

Targets

☐	Type	Target
☑	Remote	wwn.21000024FF34111C
☐	Remote	wwn.21000024FF34111D
☑	Local	wwn.2100001B3286ACAB
☐	Local	wwn.2101001B32A6ACAB

Adding WWNs to an Initiator Group

NexentaStor supports LUN masking so that only a specified initiator or group of initiators can see a LUN provided by a target group. Nexenta recommends that all initiators be grouped, even when only a single initiator is a member, such that all LUN masks are a relationship between a LUN, the target group that presents it, and the initiator group that accesses it.

You must configure an initiator group for each individual initiator or group of initiators. Depending on the configuration of your SAN, some SCSI initiators may not be visible on some nodes. Therefore, you may need to manually add the initiator information for initiators not visible to the node that you are working on.

Alternatively, you can create an initiator group on one node and assign initiators that are visible to that node. Then you need to modify the initiator group from the other node by adding initiators visible from that second node.

❖ *To create an initiator group, using NMV:*

1. Log in to an HA Cluster node.
1. Click **Data Management > SCSI Target**.
2. In the **SCSI Target** panel, click **Initiator Groups**.
3. In the **Manage Groups of Remote Initiators** window, click **here**.
4. In the **Create New Initiator Group** window, specify a custom group name.
5. In the Additional Initiators field, type the WWNs of additional initiators, not visible to this node, separated by commas.
6. Click **Create**.

Creating a Zvol

A zvol is a unit of storage allocated to back a LUN. The zvol itself has a block size and a number of ZFS attributes, such as compression settings.

Create a zvol using the required FC LUNs.

❖ *To create a zvol, using NMV:*

1. Click **Data Management > SCSI Target**.
2. In the **ZVOLS** panel, click **Create**.
3. In the **Create a New ZVOL** window, fill in the required fields and click **Create**.
4. Proceed to Mapping a Zvol.

Mapping a Zvol

Map the zvol that you created using the FC LUNs to appropriate initiator and target groups to ensure LUN visibility and failover capability.

❖ *To map a zvol:*

1. Click **Data Management > SCSI Target**.
2. In the **ZVOLS** panel, click **Mapping**.
3. In the **Manage Mappings** window, click **here**.
4. In the **Create New Mapping** dialog, fill the required fields.
5. Click **Create**.

Note:	If you do not have target or initiator groups, you can Share the zvol. If you share a zvol, it is visible to all network clients.

Administering the MetroHA Cluster

This section covers:

- Networking Management Tasks for the NexentaStor nodes
- Device Replacement for MetroHA

Networking Management Tasks

The following network management tasks apply to a MetroHA configuration:

- Adding a Virtual IP Address
- Modifying the Default Netmasks
- Adding a Virtual IP Address and Hostname

Adding a Virtual IP Address

There is a name associated with a shared volume service that is referred to as a **virtual** shared service hostname, or virtual IP address (VIP). The network clients use the virtual hostname to connect to the shared volume.

You must specify a VIP and a corresponding virtual shared service hostname in the NexentaStor IP table.

The appliances in the HA cluster group must be resolvable to each other. This means they must be able to detect each other on the network and communicate. When you create an SSH-binding, NexentaStor automatically adds records to the IP host tables on both appliances. You can verify that the records were added correctly when you add a VIP.

For each host a single line should be present with the following information:

```
IP_address    hostname    [aliases...]
```

❖ *To verify hostnames, using NMC:*

1. Log in to the NMC on one of the NexentaStor appliances.
2. Type the following to open the `/etc/hosts` file:

   ```
   nmc:/$ setup appliance hosts
   ```

3. Using the `vim` editor commands, type a virtual IP address for the shared volume.

   Example:

   ```
   Internet host table
   ::1 localhost
   127.0.0.1 localhost
   192.168.11.1  <nodeA nodeA.example.com> loghost 192.168.10.1  <nodeB
   nodeB.example.com>
   192.168.1.1  <shared_hostname>
   ```

Note: Use the failover hostname to add a shared volume.

Modifying the Default Netmasks

When you add a volume service to the HA Cluster, NexentaStor assigns a default netmask for the class of IP network it is on. However, you may need to change the default netmasks according to your configuration changes.

❖ *To modify the default netmask, using NMC:*

1. Type the following to open the `/etc/netmasks` file for nodeA:

   ```
   nmc:/$ setup appliance netmasks
   ```

2. Add the netmask for each network address:

   Example:

   ```
   192.168.1.0      255.255.255.0
   192.168.13.0     255.255.255.0
   192.168.0.0      255.255.0.0
   ```

   Repeat Step 1 — Step 2 for nodeB.

Adding a Virtual IP Address and Hostname

You can add a VIP, or shared hostname, when you create an HA Cluster. You can also add additional VIPs later. Additional VIPs provide the access to a shared volume using an additional IP address.

❖ *To add a virtual IP address, using NMV:*

1. In the **Cluster Settings** panel, click **Advanced**.
2. Click **Additional Virtual Hostnames**.
3. Select a shared volume from the drop-down list.
4. Click **Add a new virtual hostname**.
5. Select an interface for each node.
6. Click **Add**.
7. If prompted, type the IP address of the failover node.Click **Add**.

Note:	Type the IP address that is not in use and that is accessible from both nodes of the HA Cluster. You can add the hostname and IP address pair to the NexentaStor host tables. See Modifying the Default Netmasks.

8. Click **Save Settings.**
9. Click **OK** to confirm the modifications.

❖ *To add a virtual IP address, using NMC:*

1. Type:

   ```
   nmc:/$ setup group rsf-cluster <HA Cluster> vips add
   ```

2. Select the HA Cluster service.

3. Type a virtual hostname.
4. If you type an IP address or hostname that one or more HA Cluster nodes cannot resolve, NexentaStor prompts you to modify the local host tables.
   - If you want to modify the local host tables:
     1. Type **y**.
     2. Type the IP address and host name.
   - Alternatively, you can configure the DNS server settings.
     1. Type **n**.
     2. Log in to your DNS server and add the host name and IP address pair to the DNS settings.
     3. Repeat Step 1 — Step 3 and Step 5 — Step 10.
5. Select a network interface for this node.

   Nexenta recommends that you configure additional network interfaces rather than specifying the primary network interface.
6. Select network interface for the remote node.
7. Type the failover netmask.
8. Confirm the settings by typing **y**.

   System response:

   ```
   Stop adding VIPs?  (y/n)
   ```
9. Type **y** to finish adding VIPs.
10. Alternatively, type **n** to add more VIPs and repeat Step 1 — Step 9.

Device Replacement for MetroHA

This section covers the following:

- MetroHA Device Replacement Overview
- Replacing a Device
- Restoring a Path Following a Device Replacement or Disconnected SAS Cable

MetroHA Device Replacement Overview

Nexenta Professional Services is responsible for MetroHA pool device layout. Contact Nexenta Professional Services via Nexenta Support if you want to add redundancy or capacity to pools, since these changes are not currently supported as customer operations. You can replace faulted devices in the same slot of the same enclosure. For any other operational changes, including enclosure replacement, you should contact Nexenta Professional Services.

MetroHA does not support hot-standby devices because they do not guarantee correct device selection in terms of site location and fault tolerance within a site. Nexenta Professional Services lays out pools that are optimal for the initial install, for subsequent additions of SAS enclosures and/or ATTO FibreBridges.

MetroHA currently supports only four-way mirroring of regular virtual devices in the pool topology. Dedicated intent log (SLOG) virtual devices can be either mirrored either two or four ways, but two-way mirroring means that a configuration does not have fault tolerance after a site partition or failure. Loss of the remaining physical device in a SLOG virtual device can cause system panics (because writes cannot be transferred from the intent log to regular devices as part of regular transaction flushes) and prevent a pool from being imported successfully, as writes have to be de-staged from the intent log to regular devices on import and cannot be where the intent log is not available.

L2ARC devices are an optimization to allow in-memory caching to spill onto faster devices so that they can be retrieved more quickly, so there is no mirroring of L2ARC. The loss of an L2ARC device is recoverable, although there will likely be some performance impact, since recovery contents into the primary cache now requires a read from regular pool devices, which generally have lower performance. L2ARC can therefore be provisioned with only the capacity needed to support the pool in its primary location, although device topologies will need to be reconfigured in case of site loss or partition or if the service is failed over.

Replacing a Device

This section contains the procedure for replacing a drive for MetroHA. This procedure assumes that the replacement drive is installed in the same slot as the drive that was removed. In this example, disk `c0t50000393E8CABBF0d0` has failed in pool `DATA1`.

❖ *To replace a device for MetroHA, using NMC:*

1. Identify faulted logical drives:

   ```
   root@NodeA-43:# show volume DATA1 status
   ```

   System response:

   ```
   volume : DATA1
    state : ONLINE
     scan : resilvered 128K in 0h0m with 0 errors on Mon May  9 19:18:01 2016
   config : NAME                        STATE     READ WRITE CKSUM
            metro-repl                  ONLINE       0     0     0
              mirror-0                  ONLINE       0     0     0
                c0t5000C50069009207d0   ONLINE       0     0     0
                c0t5000C500893FBAD3d0   ONLINE       0     0     0
                c0t5000C50041ACCA97d0   ONLINE       0     0     0
                c0t5000C50041ACB4B3d0   ONLINE       0     0     0
              mirror-1                  ONLINE       0     0     0
                c0t5000C50056D5C683d0   ONLINE       0     0     0
                c0t5000C50057339C33d0   ONLINE       0     0     0
                c0t5000C50069010F17d0   ONLINE       0     0     0
                c0t5000C50088A91F6Fd0   ONLINE       0     0     0
              mirror-2                  ONLINE       0     0     0
                c0t5000C500212ADBEBd0   ONLINE       0     0     0
                c0t5000C50041ABCF9Bd0   ONLINE       0     0     0
                c0t5000C5006900A317d0   ONLINE       0     0     0
                c0t5000C500893FB93Fd0   ONLINE       0     0     0
              mirror-3                  ONLINE       0     0     0
   ```

```
             c0t5000C50057339CD7d0  ONLINE       0     0     0
             c0t5000C50057339F07d0  ONLINE       0     0     0
             c0t5000C50069010F1Fd0  ONLINE       0     0     0
             c0t5000C500893FBBB3d0  ONLINE       0     0     0
           mirror-4                 ONLINE       0     0     0
             c0t5000C50041AC5817d0  ONLINE       0     0     0
             c0t5000C5005733A90Fd0  ONLINE       0     0     0
             c0t5000C50069011CA7d0  ONLINE       0     0     0
             c0t5000C50088A9136Bd0  ONLINE       0     0     0
           mirror-5                 ONLINE       0     0     0
             c0t5000C50056F9DEFBd0  ONLINE       0     0     0
             c0t5000C50057339B77d0  ONLINE       0     0     0
             c0t5000C50069010DA3d0  ONLINE       0     0     0
             c0t5000C500893FC48Bd0  ONLINE       0     0     0
           mirror-6                 ONLINE       0     0     0
             c0t5000C50041ABE39Bd0  ONLINE       0     0     0
             c0t5000C50056F9E753d0  ONLINE       0     0     0
             c0t5000C50069010EFFd0  ONLINE       0     0     0
             c0t5000C50088568893d0  ONLINE       0     0     0
           mirror-7                 ONLINE       0     0     0
             c0t5000C50041AC6CDFd0  ONLINE       0     0     0
             c0t5000C5005733979Bd0  ONLINE       0     0     0
             c0t5000C50069010EE7d0  ONLINE       0     0     0
             c0t5000C500893FBCCBd0  ONLINE       0     0     0
           mirror-8                 ONLINE       0     0     0
             c0t5000C50056F9B41Bd0  ONLINE       0     0     0
             c0t5000C50057338E2Fd0  ONLINE       0     0     0
             c0t5000C50069010F33d0  ONLINE       0     0     0
             c0t5000C500893FB6ABd0  ONLINE       0     0     0
errors : No known data errors
```

Options for finding drives are the following:

- If you have vendor-provided slot mapping, identify the slot number:

```
root@NodeA-43:# show lun slotmap | grep c0t5000C500893FB93Fd0
c0t5000C500893FB93Fd0   jbod:3   5       id1,sd@n5000c500893fb93f
```

- Alternatively, if you do not have slot mapping pre-configured, you can identify the faulted drive using the following command:

```
root@NodeA-43:# show lun c0t5000C500893FB93Fd0 blink
The disk 'c0t5000C500893FB93Fd0' is currently part of the volume: 'metro-
repl'
Warning! Please make sure there is no IO activity on this volumes before
continuing ...
Proceed to blink?  Yes
Enabled blinking LED activity for disk 'c0t5000C500893FB93Fd0' (press
Ctrl-C to interrupt)...
```

The failed disk should blink.

- Use `dd` to identify the location of the drive within the JBOD. As root from bash, enter:

```
while true
do
```

```
dd if=/dev/rdsk/c0t5000C500893FB93Fd0s0 of=/dev/null bs=8192
count=100000
sleep 5
done
```

Note that this command will loop continuously. To stop the loop, two breaks must be sent together: one to interrupt the `dd` command, the other to interrupt the shell loop.

• If you are still unable to locate the disk, you will need to locate it by a process of elimination by blinking the remaining disks from NMV and seeing which one remains unlit.

2. Once you have identified the physical location of the faulted drive, enter:

```
root@NodeA-43:# setup volume metro-repl offline-lun
c0t5000C500893FB93Fd0
   volume : metro-repl
    state : DEGRADED
    scan : resilvered 8.7M in 0h0m with 0 errors on Mon May 3 5:42:53 2016
   config : NAME                       STATE     READ WRITE CKSUM
            metro-repl                 DEGRADED     0     0     0
              mirror-0                 ONLINE       0     0     0
                c0t5000C50069009207d0  ONLINE       0     0     0
                c0t5000C500893FBAD3d0  ONLINE       0     0     0
                c0t5000C50041ACCA97d0  ONLINE       0     0     0
                c0t5000C50041ACB4B3d0  ONLINE       0     0     0
              mirror-1                 ONLINE       0     0     0
                c0t5000C50056D5C683d0  ONLINE       0     0     0
                c0t5000C50057339C33d0  ONLINE       0     0     0
                c0t5000C50069010F17d0  ONLINE       0     0     0
                c0t5000C50088A91F6Fd0  ONLINE       0     0     0
              mirror-2                 DEGRADED     0     0     0
                c0t5000C500212ADBEBd0  ONLINE       0     0     0
                c0t5000C50041ABCF9Bd0  ONLINE       0     0     0
                c0t5000C5006900A317d0  ONLINE       0     0     0
                c0t5000C500893FB93Fd0  OFFLINE      0     0     0
              mirror-3                 ONLINE       0     0     0
                c0t5000C50057339CD7d0  ONLINE       0     0     0
                c0t5000C50057339F07d0  ONLINE       0     0     0
                c0t5000C50069010F1Fd0  ONLINE       0     0     0
                c0t5000C500893FBBB3d0  ONLINE       0     0     0
              mirror-4                 ONLINE       0     0     0
                c0t5000C50041AC5817d0  ONLINE       0     0     0
                c0t5000C5005733A90Fd0  ONLINE       0     0     0
                c0t5000C50069011CA7d0  ONLINE       0     0     0
                c0t5000C50088A9136Bd0  ONLINE       0     0     0
              mirror-5                 ONLINE       0     0     0
                c0t5000C50056F9DEFBd0  ONLINE       0     0     0
                c0t5000C50057339B77d0  ONLINE       0     0     0
                c0t5000C50069010DA3d0  ONLINE       0     0     0
                c0t5000C500893FC48Bd0  ONLINE       0     0     0
              mirror-6                 ONLINE       0     0     0
                c0t5000C50041ABE39Bd0  ONLINE       0     0     0
                c0t5000C50056F9E753d0  ONLINE       0     0     0
                c0t5000C50069010EFFd0  ONLINE       0     0     0
                c0t5000C50088568893d0  ONLINE       0     0     0
              mirror-7                 ONLINE       0     0     0
                c0t5000C50041AC6CDFd0  ONLINE       0     0     0
                c0t5000C5005733979Bd0  ONLINE       0     0     0
                c0t5000C50069010EE7d0  ONLINE       0     0     0
                c0t5000C500893FBCCBd0  ONLINE       0     0     0
              mirror-8                 ONLINE       0     0     0
                c0t5000C50056F9B41Bd0  ONLINE       0     0     0
                c0t5000C50057338E2Fd0  ONLINE       0     0     0
                c0t5000C50069010F33d0  ONLINE       0     0     0
                c0t5000C500893FB6ABd0  ONLINE       0     0     0
   errors : No known data errors
```

3. From bash, enter the following to identify the new disk being inserted:

   ```
   tail -f /var/adm/messages
   ```

   Output such as the following is shown when removing a disk:

   ```
   May 23 15:26:42 metronx02 scsi: [ID 243001 kern.warning] WARNING: /
   pci@0,0/pci8086,3c08@3/pci1077,15d@0/fp@0,0 (fcp0):
   May 23 15:26:42 metronx02      INQUIRY to D_ID=0x80600 lun=0x32 failed:
   sense key=Illegal_Request, ASC=5, ASCQ=0. Giving up
   May 23 15:26:42 metronx02 scsi: [ID 243001 kern.warning] WARNING: /
   pci@0,0/pci8086,3c08@3/pci1077,15d@0,1/fp@0,0 (fcp2):
   May 23 15:26:42 metronx02      INQUIRY to D_ID=0xa0600 lun=0x32 failed:
   sense key=Illegal_Request, ASC=5, ASCQ=0. Giving up
   May 23 15:26:42 metronx02 scsi: [ID 243001 kern.warning] WARNING: /
   pci@0,0/pci8086,3c08@3/pci1077,15d@0/fp@0,0 (fcp0):
   May 23 15:26:42 metronx02      INQUIRY to D_ID=0x80600 lun=0x4c failed:
   sense key=Illegal_Request, ASC=5, ASCQ=0. Giving up
   May 23 15:26:42 metronx02 scsi: [ID 243001 kern.warning] WARNING: /
   pci@0,0/pci8086,3c08@3/pci1077,15d@0,1/fp@0,0 (fcp2):
   May 23 15:26:42 metronx02      INQUIRY to D_ID=0xa0600 lun=0x4c failed:
   sense key=Illegal_Request, ASC=5, ASCQ=0. Giving up
   May 23 15:26:42 metronx02 scsi: [ID 243001 kern.info] /pci@0,0/
   pci8086,3c08@3/pci1077,15d@0/fp@0,0 (fcp0):
   May 23 15:26:42 metronx02      offlining lun=4c (trace=0), target=80600
   (trace=b00001)
   May 23 15:26:42 metronx02 scsi: [ID 243001 kern.info] /pci@0,0/
   pci8086,3c08@3/pci1077,15d@0,1/fp@0,0 (fcp2):
   May 23 15:26:42 metronx02      offlining lun=4c (trace=0), target=a0600
   (trace=b00001)
   May 23 15:26:42 metronx02 scsi: [ID 243001 kern.info] /pci@0,0/
   pci8086,3c08@3/pci1077,15d@0/fp@0,0 (fcp0):
   May 23 15:26:42 metronx02      offlining lun=32 (trace=0), target=80600
   (trace=b00001)
   May 23 15:26:42 metronx02 scsi: [ID 243001 kern.info] /pci@0,0/
   pci8086,3c08@3/pci1077,15d@0,1/fp@0,0 (fcp2):
   May 23 15:26:42 metronx02      offlining lun=32 (trace=0), target=a0600
   (trace=b00001)
   May 23 15:26:42 metronx02 genunix: [ID 483743 kern.info] /scsi_vhci/
   disk@g5000c50041acc9ab (sd175) multipath status: optimal: path 122 fp0/
   disk@w21000010866275lc,4c is offline
   ```

```
May 23 15:26:42 metronx02 genunix: [ID 483743 kern.info] /scsi_vhci/
disk@g5000c50041acc9ab (sd175) multipath status: optimal: path 146 fp0/
disk@w21000010866275lc,32 is offline
May 23 15:26:42 metronx02 genunix: [ID 483743 kern.info] /scsi_vhci/
disk@g5000c50041acc9ab (sd175) multipath status: degraded: path 215 fp2/
disk@w22000010866275lc,4c is offline
May 23 15:26:42 metronx02 genunix: [ID 408114 kern.info] /scsi_vhci/
disk@g5000c50041acc9ab (sd175) offline
May 23 15:26:42 metronx02 genunix: [ID 483743 kern.info] /scsi_vhci/
disk@g5000c50041acc9ab (sd175) multipath status: failed: path 239 fp2/
disk@w22000010866275lc,32 is offline
```

Output such as the following is shown when inserting a disk (additional output may appear):

```
May 23 15:40:48 metronx02 genunix: [ID 408114 kern.info] /scsi_vhci/
disk@g5000c500893fb93f (sd105) online
May 23 15:40:48 metronx02 genunix: [ID 483743 kern.info] /scsi_vhci/
disk@g5000c500893fb93f (sd105) multipath status: degraded: path 47 fp0/
disk@w2100001086618fc6,5e is online
May 23 15:40:48 metronx02 genunix: [ID 530209 kern.info] /scsi_vhci/
disk@g5000c500893fb93f (sd105) multipath status: optimal: path 72 fp0/
disk@w2100001086618fc6,31 is online: Load balancing: logical-block,
region-size: 18
May 23 15:40:58 metronx02 scsi: [ID 583861 kern.info] sd105 at
scsi_vhci0: unit-address g5000c500893fb93f: f_sym
May 23 15:40:58 metronx02 genunix: [ID 936769 kern.info] sd105 is /
scsi_vhci/disk@g5000c500893fb93f
```

4. Remove the failed drive from its slot and install the new disk in the same slot.

5. Enter the following command:

   ```
   root@NodeA-43:# lunsync -r
   ```

   If this is a cluster, enter `lunsync -r` on the other node.

6. Check that slot mapping is working and up to date. If it is not, enter the following command (for a cluster, enter it on both nodes).

   ```
   root@NodeA-43:# setup jbod rescan
   ```

7. Replace the drive in the pool:

   ```
   root@NodeA-43:# setup volume metro-repl replace-lun -o
   c0t5000C500893FB93Fd0 -n c0t5000C500893FC3ABd0 -r
   Replace 'c0t5000C500893FB93Fd0' with 'c0t5000C500893FC3ABd0' in the
   volume 'metro-repl'?  Yes
   The operation will take some time. Run 'show volume metro-repl status'
   to see the in-progress status.
   ```

Restoring a Path Following a Device Replacement or Disconnected SAS Cable

In a MetroHA configuration, the NexentaStor head nodes and storage enclosures are connected via FC-to-SAS bridges, specifically ATTO FibreBridge units. The NexentaStor nodes connect to the FibreBridge units over FC, and the FibreBridge units connect to the storage enclosures over SAS (see Figure 1-1).

If the SAS connection between a FibreBridge unit and a storage enclosure is disrupted, which can happen when a device is replaced or the SAS cable is detached, the devices affected by the disrupted connection may be shown as “unusable” on the NexentaStor nodes, and the contents may not be visible to the system.

In such a case, manual action is necessary to restore the path from the FibreBridge unit to the device: after resolving the SAS cable issue or installing a new drive, you enter commands at the NMC console to identify faulted devices and recover the path to them.

❖ *To recover the path:*

1. Correct the issue that disrupted the SAS connection between the FibreBridge unit and the storage enclosure (that is, install the new drive or reconnect the SAS cable).
2. On the NexentaStor node, enter bash by typing:

```
root@NodeA-43:# option expert_mode =1
root@NodeA-43:# !bash
You are about to enter the Unix ("raw") shell and execute low-level Unix
command(s). Warning: using low-level Unix commands is not recommended!
Execute? Yes
```

3. Use the `fmadm faulty` command to display the list of faults in the system. If there are failed paths, the output of the command will look like the following:

```
root@NodeA-43:# fmadm faulty
--------------- ------------------------------------  -------------- ---------
TIME            EVENT-ID                              MSG-ID         SEVERITY
--------------- ------------------------------------  -------------- ---------
Aug 03 17:49:07 e13c5edb-a25c-cb9a-b2ff-fcc1efca383c  ZFS-8000-JQ    Major

Host        : nodea-43
Platform    : PowerEdge-R720    Chassis_id  : FJS14Y1
Product_sn  :

Fault class : fault.fs.zfs.io_failure_continue
Affects     : zfs://pool=myvol02
              faulted but still in service
Problem in  : zfs://pool=myvol02
              faulted but still in service

Description : The ZFS pool has experienced currently unrecoverable I/O
              failures.  Refer to http://illumos.org/msg/ZFS-8000-JQ for
              more information.

Response    : No automated response will be taken.

Impact      : Read and write I/Os cannot be serviced.

Action      : Make sure the affected devices are connected, then run
              'zpool clear'.
```

4. Enter the `zpool clear <pool>` command, which clears all faults raised by ZFS.
5. After clearing all of the faults, identify disks in “unusable” condition:

```
root@NodeA-43:# cfgadm -al -o show_SCSI_LUN
```

```
Ap_Id                         Type        Receptacle   Occupant     Condition
...
c10::210000108662751c,3        ESI          connected    configured   unknown
c10::210000108662751c,4        processor    connected    configured   unknown
c8::2100001086618fc7,0         disk         connected    configured   unusable
```

6. For disks that have a condition of “unusable”, use the following command to recover the path to the disks:

   ```
   cfgadm -c configure <disk-id-prefix>
   ```

   For example:

   ```
   root@NodeA-43:# cfgadm -c configure c8::2100001086618fc7
   ```

Service Modes and States for the HA Cluster Plug-In

A service can be in automatic or manual mode for a cluster node. If a service is in automatic mode on a node, it is automatically started if it is not already running somewhere in the cluster. If a service is in manual mode, the cluster does not attempt to take any action on the service except by operator intervention.

A service can have a state of *stopped*, *starting*, *running*, *stopping*, or *broken*. The *stopped* and *running* states indicate that a service is offline or online on a node. The *starting* and *stopping* states indicate that a service is transitioning to an online or offline state, respectively. A *broken* state indicates that a state had an unexpected state change or failed to make a state change because failover automation encountered an exception it could not correct. Operator intervention is required to resolve the underlying problem resulting in a broken state, which may be in software, hardware, or underlying infrastructure, at which point the service should be marked as repaired and can be recovered.

Site partitions and failures are treated as special cases of *broken* state, requiring administrative intervention to confirm that a node should recover services. Incorrect determination of site partition as site loss results in a split-brain condition.

Operator tasks for setting service modes for Metro HA include the following:

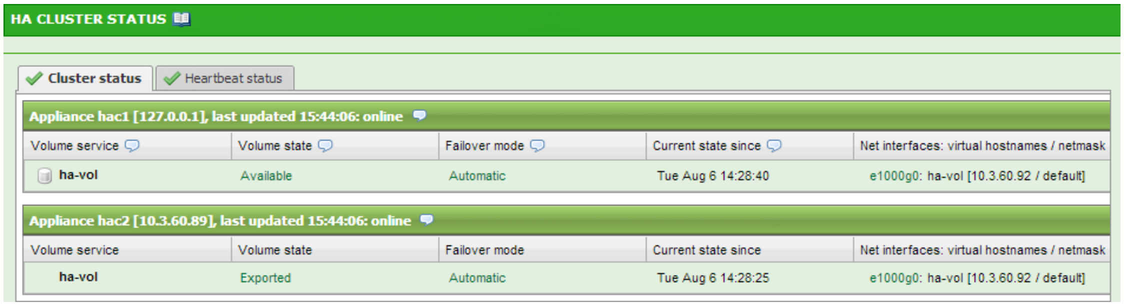
- Verifying HA Status
- Setting Failover Mode
- Managing Global Cluster Properties
- Manually Triggering a Failover
- Repairing a Broken Cluster Service

Verifying HA Status

Verify the status on the shared volume service using NMV or NMC.

❖ *To view the status of a shared volume, using NMV:*

1. In the **Cluster Settings** panel, click **Status**.



Appliance hac1 [127.0.0.1], last updated 15:44:06: online

Volume service	Volume state	Failover mode	Current state since	Net interfaces: virtual hostnames / netmask
ha-vol	Available	Automatic	Tue Aug 6 14:28:40	e1000g0: ha-vol [10.3.60.92 / default]

Appliance hac2 [10.3.60.89], last updated 15:44:06: online

Volume service	Volume state	Failover mode	Current state since	Net interfaces: virtual hostnames / netmask
ha-vol	Exported	Automatic	Tue Aug 6 14:28:25	e1000g0: ha-vol [10.3.60.92 / default]

❖ *To view the status of a shared volume, using NMC:*

- Type:

```
nmc:/$ show group rsf-cluster
```

System response:

```
PROPERTY                   VALUE
name                     : HA-Cluster
appliances               : [metronx02 metronx01]
machinesigs              : {"metronx02":"46BJAHEMB","metronx01":"2D79KCMLD"}
creator                  : metronx02.mydomain.local
hbipifs                  : metronx02:metronx01: metronx01:metronx02:
fcmon                    : 1
netmon                   : 1
info                     : Nexenta HA-Cluster
generation               : 3
refresh_timestamp        : 1448557558.09891
type                     : rsf-cluster
creation                 : Nov 26 09:05:58 2015

SHARED VOLUME: metro-repl
svc-metro-repl-shared-vol-name : metro-repl
svc-metro-repl-ipdevs          : metro-repl/255.255.255.0 metronx01:ixgbe2000
metronx02:ixgbe2000
svc-metro-repl-ipdevs-IPv6     :
svc-metro-repl-attached-vols   :
svc-metro-repl-main-node       : metronx02
svc-metro-repl-inittimeout     : 20
svc-metro-repl-runtimeout      : 8
svc-metro-repl-mhdc-disable    : n
svc-metro-repl-monitor         :
{"metronx02":{"monitor":"","ipdevs":{"ixgbe2000":""}},"metronx01"
svc-metro-repl-resdisks        :

HA CLUSTER STATUS: HA-Cluster
metronx01:
 metro-repl    stopped      auto    unblocked     metro-repl     ixgbe2000     20  8
metronx02:
 metro-repl    running      auto    unblocked     metro-repl     ixgbe2000     20  8
```

Setting Failover Mode

The failover mode defines whether or not an appliance attempts to start a service when it is not running. There are separate failover mode settings for each appliance that can run a service.

Note:	Set failover mode to manual every time you perform any maintenance to avoid unwanted failover events.

You can set the failover to the following modes:

- Setting Manual Failover Mode
- Setting Automatic Failover Mode

Setting Manual Failover Mode

In manual mode, the HA Cluster service does not initiate the failover when it detects a failure. However, it generates warnings when the parallel appliance is not available. If the appliance cannot obtain a definitive answer about the state of the service, or the service is not running anywhere else, the appropriate timeout must expire before you can take any action. The primary service failover modes are typically set to automatic to ensure that an appliance starts its primary service(s) on boot up.

Note:	Setting a service to manual mode when the service is already running does not stop that service, it only prevents the service from starting on that appliance.

❖ *To set the failover mode to manual, using NMV:*

1. Click **Advanced Setup > Cluster Operations > Set all Manual**.
2. Click **Yes** to confirm.

Note:	Before HAC performs an operation, it saves the state of the services in the cluster, which you can later re-apply to the cluster using the restore button. Once HA Cluster restores the service state, it clears the saved state.

Setting Automatic Failover Mode

In automatic mode, the appliance attempts to start the service when it detects that there is no available parallel appliance running in the cluster. Automatic failover mode is the default setting.

❖ *To set the failover mode to automatic, using NMV:*

1. Click **Advanced Setup > Cluster Operations > Set all Automatic**
2. Click **Yes** to confirm.

❖ *To stop all services in the HA Cluster, using NMV:*

1. Click **Stop All Services**.
2. Click **Yes** to confirm.

Managing Global Cluster Properties

You can manage the global cluster properties in using from the Advanced tab in NMV add an advanced level of control for fine-tuning the HA cluster.

❖ *To manage global cluster properties using NMV:*

1. Click **Settings > HA Cluster**.
2. Select the **Advanced** > **Global Cluster Properties** tab.
3. Modify the properties as required.

Table 2-1: Global Cluster Properties

Property	Description
Force iSCSI group creation	Force iSCSI view creation using the Group All function if an iSCSI view refers to a group that is not found on the node to which the volume is failing over. By default, if NexentaStor is unable to find elements in a LUN mask on failover, it removes previous restrictions and makes the LUN visible to all initiators. This is a preference for restoring availability ahead of configuration integrity. If this behavior does not meet the security needs of the site, this property should be disabled.
Failover State is sticky	If enabled, during the failover transfers the failover mode to the alternate node. If disabled the failover mode is not transferred. HA Cluster has the following failover modes: • Automatic • Manual

Manually Triggering a Failover

You can manually trigger a failover between systems when needed. Performing a failover from the current appliance to the specified appliance causes the volume sharing service to stop on the current appliance, and the opposite actions to take place on the passive appliance. Additionally, the volume exports to the other node.

Note: You must first set all cluster operations to manual mode.

Note: You should not manually fail over multiple services at the same time. Services should be failed over sequentially.

❖ *To manually trigger a failover, using NMC:*

1. Verify that shared volume is in healthy state by typing:

   ```
   nmc:/$ zpool status <shared-volume>
   ```

   Example:

   ```
   pool : <shared-volume>
   state : ONLINE
   scan : none requested
   config :
            NAME                       STATE     READ WRITE CKSUM
            <shared-volume>            ONLINE       0     0     0
              mirror-0                 ONLINE       0     0     0
                c0t5000C50069009207d0  ONLINE       0     0     0
                c0t5000C500893FBAD3d0  ONLINE       0     0     0
                c0t5000C50041ACCA97d0  ONLINE       0     0     0
                c0t5000C50041ACB4B3d0  ONLINE       0     0     0
              mirror-1                 ONLINE       0     0     0
                c0t5000C50056D5C683d0  ONLINE       0     0     0
                c0t5000C50057339C33d0  ONLINE       0     0     0
                c0t5000C50069010F17d0  ONLINE       0     0     0
                c0t5000C50088A91F6Fd0  ONLINE       0     0     0
              mirror-2                 ONLINE       0     0     0
                c0t5000C500212ADBEBd0  ONLINE       0     0     0
                c0t5000C50041ABCF9Bd0  ONLINE       0     0     0
                c0t5000C5006900A317d0  ONLINE       0     0     0
                c0t5000C500893FB93Fd0  ONLINE       0     0     0
              mirror-3                 ONLINE       0     0     0
                c0t5000C50057339CD7d0  ONLINE       0     0     0
                c0t5000C50057339F07d0  ONLINE       0     0     0
                c0t5000C50069010F1Fd0  ONLINE       0     0     0
                c0t5000C500893FBBB3d0  ONLINE       0     0     0
              mirror-4                 ONLINE       0     0     0
                c0t5000C50041AC5817d0  ONLINE       0     0     0
                c0t5000C5005733A90Fd0  ONLINE       0     0     0
                c0t5000C50069011CA7d0  ONLINE       0     0     0
                c0t5000C50088A9136Bd0  ONLINE       0     0     0
              mirror-5                 ONLINE       0     0     0
                c0t5000C50056F9DEFBd0  ONLINE       0     0     0
                c0t5000C50057339B77d0  ONLINE       0     0     0
                c0t5000C50069010DA3d0  ONLINE       0     0     0
                c0t5000C500893FC48Bd0  ONLINE       0     0     0
              mirror-6                 ONLINE       0     0     0
                c0t5000C50041ABE39Bd0  ONLINE       0     0     0
                c0t5000C50056F9E753d0  ONLINE       0     0     0
                c0t5000C50069010EFFd0  ONLINE       0     0     0
                c0t5000C50088568893d0  ONLINE       0     0     0
              mirror-7                 ONLINE       0     0     0
                c0t5000C50041AC6CDFd0  ONLINE       0     0     0
                c0t5000C5005733979Bd0  ONLINE       0     0     0
                c0t5000C50069010EE7d0  ONLINE       0     0     0
                c0t5000C500893FBCCBd0  ONLINE       0     0     0
              mirror-8                 ONLINE       0     0     0
   ```

```
             c0t5000C50056F9B41Bd0  ONLINE        0      0      0
             c0t5000C50057338E2Fd0  ONLINE        0      0      0
             c0t5000C50069010F33d0  ONLINE        0      0      0
             c0t5000C500893FB6ABd0  ONLINE        0      0      0
errors : No known data errors
```

Warning: If any disk drive from the shared volume is in `DEGRADED` state, you must replace the faulted drive(s) before executing failover. Otherwise, failover may take long time or your system may freeze.

2. Type:

   ```
   nmc:/$ setup group rsf-cluster <cluster_name> failover
   ```

Repairing a Broken Cluster Service

NexentaStor tracks various appliance components and their state. If and when failover occurs (or any service changes to a broken state), NexentaStor sends an email to the administrator describing the event.

You can execute the repair command for a cluster service. The repair command forces the import operation of a shared volume. Therefore, the shared volume must be exported on both nodes.

Note: During the NexentaStor installation, you set up SMTP configuration and test so that you can receive emails from the appliance.

There are two broken states:

- **Broken_Safe**

  A problem occurred while starting the service on the server, but it was stopped safely and you can run it elsewhere.

- **Broken_Unsafe**

  A fatal problem occurred while starting or stopping the service on the server. The service cannot run on any other server in the cluster until it is repaired.

Warning: Manually verify and troubleshoot the volume before marking the state as repaired. Failure to do so could result in cross-mounting of the volume or storage partition, leading to data corruption or loss.

❖ *To repair a shared volume that is in broken state, using NMC:*

1. Verify that the volume is exported on both HA Cluster nodes by typing:

   ```
   nmc:/$ zpool status
   ```

   The output should not include the information about the shared volume.

2. Repeat Step 1 on other node.

3. Execute the volume repair operation:

   ```
   nmc:/$ setup group rsf-cluster shared-volume repair <cluster_name>
   <volume_name>
   ```

This initiates and runs the repair process.

3

Troubleshooting MetroHA

This chapter includes the following topics:

Overview

This chapter contains procedures you can use to troubleshoot a MetroHA topology. When troubleshooting MetroHA, Nexenta recommends using an inside-out approach, which can minimize errors and reduce diagnostic time in localizing storage problems in terms of the device topology. For MetroHA, the inside-out approach means that you start with the device's immediate connection point in the SAS storage enclosure and move further out, one layer at a time, so that you are certain of the previous layer's state when diagnosing the next layer. For MetroHA, the layers are:

1. Storage devices
2. SAS enclosures
3. ATTO FibreBridges
4. FC switches
5. NexentaStor head nodes

Some steps may require assistance from Nexenta support or whoever operates your FC switch environment.

Troubleshooting a MetroHA Configuration

Use the procedures in this section to check the following:

- All disks in the configuration are visible
- All SAS enclosures are visible

Confirming Appliance View of JBOD Inventory and Device Census

Use the NMC command `show jbod` to verify that all expected disks are visible in the storage enclosure. If a disk is installed in one of the slots, but not visible to the system, the command may report a state of `not-installed` for the slot.

This check is effective if the problem is only at the slot or device level, but connectivity is otherwise fine. If that is not the case, the JBOD SES inventory can be checked in a more cursory fashion on the FibreBridge (see Checking SES Inventory).

For example:

```
root@NodeA-43:# show jbod jbod:1 all
...
ELEMENT   SENSOR         VALUE      STATE
jbod      state          -          ok
slot:1    state          -          ok
slot:2    state          -          ok
slot:3    state          -          ok
slot:4    state          -          ok
slot:5    state          -          ok
slot:6    state          -          ok
slot:7    state          -          not-installed
slot:8    state          -          ok
...
```

Checking SAS Enclosure Visibility

Use the FibreBridge CLI command `SASEnclosures` to verify enclosure visibility.

```
Ready.
SASEnclosures
6
;Idx Enclosure Addr   # Devs  Start LUN # Vendor   Product
-------------------------------------------------------------------
   0 500304800111283f  45        1          SMCI     SC216BE2CJBOD
   1 50030480091bda3f  45       91          SMCI     SC216BE2CJBOD
   2 50030480011128bf  45      136          SMCI     SC216BE2CJBOD
   3 50030480091bdabf  45       46          SMCI     SC216BE2CJBOD
```

The command shows redundant SAS expanders in an enclosure separately. In this example `500304800111283f` and `50030480011128bf` are the two expanders in the first storage enclosure, and `50030480091bda3f` and `50030480091bdabf` are the two in the second. Although these are twenty-four bay expanders, SES reports 45 bays; thus the FibreBridge allocates LUN numbers for 44 bays on each expander, plus a LUN number for the expander so that SES data can be bridged.

Troubleshooting the ATTO FibreBridges

Use the procedures in this section to check the following:

- The FibreBridge has FC connectivity
- The links to the ATTO FibreBridges are up
- The storage enclosures are visible from the FibreBridge
- All SAS targets are visible
- The SAS-to-FC mappings are correct

If necessary, you can gather diagnostic information from the FibreBridge and submit it to Nexenta Support for analysis.

Verifying Network Connectivity

Use the following FibreBridge CLI commands to check FibreBridge FC connectivity.

```
Ready.
FCPortList
4
; Port Status
;===============
1 Up
2 Up

Ready.
get FCConnMode all
4
; Port Conn Mode
;==================
1 ptp-loop
2 ptp-loop

Ready.
get FCDataRate all
4
; Port Data Rate
;==================
1 auto
2 auto
```

The FCPortList command shows the ports as up or down; the FCConnMode command shows what the ports are configured to do, which may include negotiations with outcomes that aren't displayed.

To confirm that the switch sees the bridges as connected, enter the command switchshow output on the bridge, which displays the WWPN and the port state. You should see online FC F-Ports for edge devices. Anything else, and it may be that the FCConnMode setting on the bridge is set to something other than ptp-loop, such that it prefers or exclusively accepts loop mode.

Confirming ATTO FibreBridge SAS Port Connectivity

Use the FibreBridge CLI command SASPortList to verify the SAS connections are up. There should be redundant SAS and FC connections throughout the deployment. Therefore, you should see two online SAS ports negotiated up to 6 Gb.

For example:

```
Ready.
SASPortList
10
;Connector      PHY     Link    Speed   SAS Address
;========================================================
Device  A       1       Up      6Gb     5001086000618fc6
Device  A       2       Up      6Gb     5001086000618fc6
Device  A       3       Up      6Gb     5001086000618fc6
Device  A       4       Up      6Gb     5001086000618fc6
Device  B       1       Up      6Gb     5001086000618fc6
Device  B       2       Up      6Gb     5001086000618fc6
Device  B       3       Up      6Gb     5001086000618fc6
Device  B       4       Up      6Gb     5001086000618fc6
...
```

Checking SES Inventory

The FibreBridge maps devices in terms of slot population using SES data, so you need to confirm that all slots are properly reported. If SES is not presented properly to the host, output from the `show jbod` command (see Confirming Appliance View of JBOD Inventory and Device Census above) is not available, but you can still check the enclosure inventory via the FibreBridge in case the problem is further out in the stack. You can see this data for individual expanders by entering the `SASEnclosures` command on the FibreBridge with an index specified.

For example:

```
Ready.
SASEnclosures 0
47
;Slot #  Device Description
-------------------------------------------------------------------
    1    5000c50069009205   0 SEAGATE  ST800FM0043      P3G111370000T0000000
    2    5000c50069010f15   0 SEAGATE  ST800FM0043      P3G111180000T0000000
    3    5000c50069010f1d   0 SEAGATE  ST800FM0043      P3G111700000T0000000
    4    5000c50069011ca5   0 SEAGATE  ST800FM0043      P3G113460000T0000000
    5    5000c50069010da1   0 SEAGATE  ST800FM0043      P3G1301C0000T0000000
    6    5000c50069010efd   0 SEAGATE  ST800FM0043      P3G111280000T0000000
    7    5000c50069010ee5   0 SEAGATE  ST800FM0043      P3G1115E0000T0000000
    8    5000c50069010f31   0 SEAGATE  ST800FM0043      P3G110F60000T0000000
    9    5000c50069011bcd   0 SEAGATE  ST800FM0043      P3G112640000T0000000
   10    5000c50069010da9   0 SEAGATE  ST800FM0043      P3G130050000T0000000
   11    5000c50069010db5   0 SEAGATE  ST800FM0043      P3G130090000T0000000
   12    5000c50069010d91   0 SEAGATE  ST800FM0043      P3G130300000T0000000
   13    5000c50069010d61   0 SEAGATE  ST800FM0043      P3G130190000T0000000
   14    5000c50069003339   0 SEAGATE  ST800FM0043      P3G102330000T0000000
   15    5000c50069010f49   0 SEAGATE  ST800FM0043      P3G111270000T0000000
   16    5000c5003012e239   0 SEAGATE  ST800FM0043      Z3G013BS0000Z3G013BS
   17    5000c5003012e24d   0 SEAGATE  ST800FM0043      Z3G013JC0000Z3G013JC
   18    5000c50069010dad   0 SEAGATE  ST800FM0043      P3G1300C0000T0000000
   19    5000c50069008a95   0 SEAGATE  ST800FM0043      P3G1110B0000T0000000
   20    5000c50069011b59   0 SEAGATE  ST800FM0043      P3G1126F0000T0000000
   21    5000c5003012e23d   0 SEAGATE  ST800FM0043      Z3G013G40000Z3G013G4
   22    500003945831060a   0 TOSHIBA  AL13SEB900       X2U0A01NFTR6
   23    50000394583105de   0 TOSHIBA  AL13SEB900       X2U0A01HFTR6
   24    *NOT PRESENT*
   25    *NOT PRESENT*
   26    *NOT PRESENT*
   27    *NOT PRESENT*
   28    *NOT PRESENT*
   29    *NOT PRESENT*
   30    *NOT PRESENT*
   31    *NOT PRESENT*
   32    *NOT PRESENT*
   33    *NOT PRESENT*
   34    *NOT PRESENT*
   35    *NOT PRESENT*
   36    *NOT PRESENT*
   37    *NOT PRESENT*
   38    *NOT PRESENT*
   39    *NOT PRESENT*
   40    *NOT PRESENT*
   41    *NOT PRESENT*
   42    *NOT PRESENT*
   43    *NOT PRESENT*
  UNK    5000c5006900a315   0 SEAGATE  ST800FM0043      P3G111FE0000T0000000
  UNK    500304800111283d   0 SMCI     SC216BE2CJBOD    SMC01020304050607080
```

In this example, the command shows output for a fully populated 24-bay SuperMicro SC216BE2CJBOD. For the enclosure, the expander is at the bottom, which does not have a slot index (slot UNK). The enclosure indicates it has 44 bays, but the FibreBridge believes its first slot is 1 rather than 0, so the device in slot 0 appears second from bottom in an unknown slot.

Checking SAS Target Visibility

There are two separate items to look at when checking device visibility via the FibreBridge:

- Is the device in a working slot and properly reported by SES? See "Checking SES Inventory" above.
- Is the device connected to the FibreBridge via SAS?

Use the FibreBridge CLI command SASTargets to check target visibility. Since the ATTO FibreBridge does not reduce multipathed entities to a single instance, redundantly connected SAS devices appear as the same serial number twice. Note that the command does not show through which SAS port the target is visible, so if a target is not visible, you should check whether both FibreBridge ports are online and connected (see Confirming ATTO FibreBridge SAS Port Connectivity.)

```
Ready.
SASTargets
101
; Tgt VendorID ProductID        Type          SerialNumber
    0 TOSHIBA  AL13SEB900       Disk          X2U0A01HFTR6
    1 TOSHIBA  AL13SEB900       Disk          X2U0A01HFTR6
    2 TOSHIBA  AL13SEB900       Disk          X2U0A01NFTR6
    3 TOSHIBA  AL13SEB900       Disk          X2U0A01NFTR6
    4 STEC     S842Z32M2        Disk          STM000196CBE
    5 STEC     S842Z32M2        Disk          STM000196CBE
    6 SEAGATE  ST800FM0043      Disk          Z3G013BS0000Z3G013BS
    7 SEAGATE  ST800FM0043      Disk          Z3G013BS0000Z3G013BS
    8 SEAGATE  ST800FM0043      Disk          Z3G013G40000Z3G013G4
    9 SEAGATE  ST800FM0043      Disk          Z3G013G40000Z3G013G4
   10 SEAGATE  ST800FM0043      Disk          Z3G013JC0000Z3G013JC
   11 SEAGATE  ST800FM0043      Disk          Z3G013JC0000Z3G013JC
   12 SEAGATE  ST800FM0043      Disk          P3G102330000T0000000
   13 SEAGATE  ST800FM0043      Disk          P3G102330000T0000000
   14 SEAGATE  ST800FM0043      Disk          P3G1110B0000T0000000
   15 SEAGATE  ST800FM0043      Disk          P3G1110B0000T0000000
   16 SEAGATE  ST800FM0043      Disk          P3G111370000T0000000
   17 SEAGATE  ST800FM0043      Disk          P3G111370000T0000000
...
```

Checking SAS-to-FC Mappings

Use the FibreBridge CLI command `RouteDisplay FC` to display the mappings of SAS targets to FC LUNs. In a MetroHA configuration, each SAS port is presented out of each FC port, so the host sees the devices four times but knows to present a device only once per path. On the FibreBridge side, everything is seen twice from the SAS side, which is then doubled when presented by redundant FC ports.

The following example shows a working set of mappings:

```
Ready.
RouteDisplay FC
103
; FL   Device Address   LUN VendorID ProductID        SerialNumber
;================================================================
   0   Bridge
   1   5000c50069009205   0 SEAGATE  ST800FM0043      P3G111370000T0000000
   2   5000c50069010f15   0 SEAGATE  ST800FM0043      P3G111180000T0000000
   3   5000c50069010f1d   0 SEAGATE  ST800FM0043      P3G111700000T0000000
   4   5000c50069011ca5   0 SEAGATE  ST800FM0043      P3G113460000T0000000
   5   5000c50069010da1   0 SEAGATE  ST800FM0043      P3G1301C0000T0000000
   6   5000c50069010efd   0 SEAGATE  ST800FM0043      P3G111280000T0000000
   7   5000c50069010ee5   0 SEAGATE  ST800FM0043      P3G1115E0000T0000000
   8   5000c50069010f31   0 SEAGATE  ST800FM0043      P3G110F60000T0000000
   9   5000c50069011bcd   0 SEAGATE  ST800FM0043      P3G112640000T0000000
  10   5000c50069010da9   0 SEAGATE  ST800FM0043      P3G130050000T0000000
  11   5000c50069010db5   0 SEAGATE  ST800FM0043      P3G130090000T0000000
  12   5000c50069010d91   0 SEAGATE  ST800FM0043      P3G130300000T0000000
  13   5000c50069010d61   0 SEAGATE  ST800FM0043      P3G130190000T0000000
  14   5000c50069003339   0 SEAGATE  ST800FM0043      P3G102330000T0000000
  15   5000c50069010f49   0 SEAGATE  ST800FM0043      P3G111270000T0000000
  16   5000c5003012e239   0 SEAGATE  ST800FM0043      Z3G013BS0000Z3G013BS
  17   5000c5003012e24d   0 SEAGATE  ST800FM0043      Z3G013JC0000Z3G013JC
  18   5000c50069010dad   0 SEAGATE  ST800FM0043      P3G1300C0000T0000000
  19   5000c50069008a95   0 SEAGATE  ST800FM0043      P3G1110B0000T0000000
  20   5000c50069011b59   0 SEAGATE  ST800FM0043      P3G1126F0000T0000000
  21   5000c5003012e23d   0 SEAGATE  ST800FM0043      Z3G013G40000Z3G013G4
  22   500003945831060a   0 TOSHIBA  AL13SEB900       X2U0A01NFTR6
  23   50000394583105de   0 TOSHIBA  AL13SEB900       X2U0A01HFTR6
  44   5000c5006900a315   0 SEAGATE  ST800FM0043      P3G111FE0000T0000000
  45   500304800111283d   0 SMCI     SC216BE2CJBOD    SMC0102030405060708O
  46   5000a72b3009cd21   0 STEC     S842Z32M2        STM000196CBE
  47   5000c500893fbad2   0 SEAGATE  ST1200MM0018     S4003J410000K5346FD1
  48   5000c50088a91f6e   0 SEAGATE  ST1200MM0018     S4003ES10000K5325DGW
...
```

In this example, the expander itself is mapped as LUN 45, which allows the appliance to see the JBOD device census (see Confirming Appliance View of JBOD Inventory and Device Census above).

The FC mappings from the `RouteDisplay FC` command show 15 SAS attachments each per FC port (that is, the same 30 target attachments, accounting for everything in the `SASTargets` output) plus two bridge mappings, which means the full SAS inventory is mapped.

Gathering Diagnostics From ATTO FibreBridges

Nexenta Support may request diagnostics from the ATTO FibreBridge 6500s. You should retain records of the IP addresses and root passwords for the FibreBridges so that you can log in to them as directed. You can access an ATTO FibreBridge 6500 for remote management using a Web browser or a Telnet client. NexentaStor does not ship with a Telnet client by default, so you should install one from your desktop if you do not wish to use a Web browser.

❖ *To gather diagnostic information for an ATTO FibreBridge:*

1. Using a Web browser, log into the FibreBridge by entering its IP address or host name and then entering root's login credentials.
2. Once logged in, navigate to the **Advanced** item in the list on the left-hand side, which presents the Advanced CLI configuration page.
3. Use the `DumpConfiguration` command.

   The `DumpConfiguration` command generally produces more output than can be buffered by a Telnet client or the browser interface, but it also generates a file that can be retrieved via FTP.
4. After running the `DumpConfiguration` command, log into the FibreBridge via FTP with the same address or name and credentials. Note that the ATTO FibreBridge 6500 only supports one concurrent login, and the previous login session may need to expire before another login can succeed.
5. Once connected by FTP, retrieve the `dumpcfg.txt` file, which contains the diagnostic information.

Host Troubleshooting for MetroHA

This section describes how to check whether the NexentaStor node has its FC ports online and how to identify and correct faults.

Checking Host FC Target Port Visibility

If the FibreBridge sees storage via SAS, has correct SES data, and reflects this in its mappings, confirm that the host has its FC ports online and properly connected and that it can see the expected remote ports. To determine this, enter bash and use the following shell loop with the `fcinfo` command:

```
root@NodeA-43:# option expert_mode =1
root@NodeA-43:# !bash
```

```
You are about to enter the Unix ("raw") shell and execute low-level Unix command(s).
Warning: using low-level Unix commands is not recommended! Execute? Yes

root@NodeA-43:# for wwpn in $(fcinfo hba-port -i | nawk '$1 == "HBA" {print $NF}');
do fcinfo hba-port $wwpn; fcinfo remote-port -p $wwpn; done

HBA Port WWN: 21000024ff453d7e
        Port Mode: Initiator
        Port ID: 80500
        OS Device Name: /dev/cfg/c3
        Manufacturer: QLogic Corp.
        Model: QLE2562
        Firmware Version: 05.03.01
        FCode/BIOS Version:  BIOS: 3.00; EFI: 2.16;
        Serial Number: BFD1219A40544
        Driver Name: qlc
        Driver Version: 20100408-3.01
        Type: N-port
        State: online
        Supported Speeds: 2Gb 4Gb 8Gb
        Current Speed: 8Gb
        Node WWN: 20000024ff453d7e
        Max NPIV Ports: 254
        NPIV port list:
Remote Port WWN: 210000108662751c
        Active FC4 Types: SCSI
        SCSI Target: yes
        Port Symbolic Name:
        Node WWN: 200000108662751c
Remote Port WWN: 2100001086618fc6
        Active FC4 Types: SCSI
        SCSI Target: yes
        Port Symbolic Name:
        Node WWN: 2000001086618fc6
HBA Port WWN: 21000024ff453d7f
        Port Mode: Initiator
        Port ID: a0500
        OS Device Name: /dev/cfg/c4
        Manufacturer: QLogic Corp.
        Model: QLE2562
        Firmware Version: 05.03.01
        FCode/BIOS Version:  BIOS: 3.00; EFI: 2.16;
        Serial Number: BFD1219A40544
        Driver Name: qlc
        Driver Version: 20100408-3.01
        Type: N-port
        State: online
        Supported Speeds: 2Gb 4Gb 8Gb
        Current Speed: 8Gb
        Node WWN: 20000024ff453d7f
        Max NPIV Ports: 254
        NPIV port list:
Remote Port WWN: 220000108662751c
        Active FC4 Types: SCSI
        SCSI Target: yes
        Port Symbolic Name:
        Node WWN: 200000108662751c
Remote Port WWN: 2200001086618fc6
        Active FC4 Types: SCSI
        SCSI Target: yes
        Port Symbolic Name:
        Node WWN: 2000001086618fc6
```

This example shows two HBAs (or a single dual-ported HBA) with ports `c3` and `c4`. It is connected to a fabric in the correct mode (N-port type), the connect is working (`State: online`), can run at 2, 4, or 8 Gb, and is now running at 8 Gb. Port `c3` sees `210000108662751c` and `2100001086618fc6`, and port `c4` sees `220000108662751c` and `2200001086618fc6`, all of which are advertising SCSI target services.

FMA Overview

In terms of storage faults, FMA receives events from two places: the SCSI initiator (sd) and ZFS. Events generated by the SCSI initiator use the prefix `DISK`, while events generated by ZFS use the prefix `ZFS`.

The following shows a pool that has just had one of two storage enclosures behind a FibreBridge powered off as it appears in the output of the `zpool status` command:

```
root@NodeA-43:# zpool status -v metro-repl
  pool: metro-repl
 state: DEGRADED
status: One or more devices are faulted in response to persistent errors.
        Sufficient replicas exist for the pool to continue functioning in a
        degraded state.
action: Replace the faulted device, or use 'zpool clear' to mark the device
        repaired.
  scan: resilvered 1.05M in 0h0m with 0 errors on Mon Apr 11 16:56:28 2016
config:
        NAME                       STATE     READ WRITE CKSUM
        metro-repl                 DEGRADED     0     0     0
          mirror-0                 DEGRADED     0     0     0
            c0t5000C50069009207d0  ONLINE       0     0     0
            c0t5000C500893FBAD3d0  FAULTED      0     0     0  too many errors
            c0t5000C50041ACCA97d0  ONLINE       0     0     0
            c0t5000C50041ACB4B3d0  ONLINE       0     0     0
          mirror-1                 DEGRADED     0     0     0
            c0t5000C50056D5C683d0  ONLINE       0     0     0
            c0t5000C50057339C33d0  ONLINE       0     0     0
            c0t5000C50069010F17d0  ONLINE       0     0     0
            c0t5000C50088A91F6Fd0  REMOVED      0     0     0
          mirror-2                 DEGRADED     0     0     0
            c0t5000C500212ADBEBd0  ONLINE       0     0     0
            c0t5000C50041ABCF9Bd0  ONLINE       0     0     0
            c0t5000C5006900A317d0  ONLINE       0     0     0
            c0t5000C500893FB93Fd0  FAULTED      0    64     0  too many errors
          mirror-3                 DEGRADED     0     0     0
            c0t5000C50057339CD7d0  ONLINE       0     0     0
            c0t5000C50057339F07d0  ONLINE       0     0     0
            c0t5000C50069010F1Fd0  ONLINE       0     0     0
            c0t5000C500893FBBB3d0  FAULTED      0     0     0  too many errors
          mirror-4                 DEGRADED     0     0     0
            c0t5000C50041AC5817d0  ONLINE       0     0     0
            c0t5000C5005733A90Fd0  ONLINE       0     0     0
            c0t5000C50069011CA7d0  ONLINE       0     0     0
            c0t5000C50088A9136Bd0  FAULTED      0     0     0  too many errors
          mirror-5                 DEGRADED     0     0     0
            c0t5000C50056F9DEFBd0  ONLINE       0     0     0
            c0t5000C50057339B77d0  ONLINE       0     0     0
            c0t5000C50069010DA3d0  ONLINE       0     0     0
            c0t5000C500893FC48Bd0  FAULTED      0     0     0  too many errors
          mirror-6                 DEGRADED     0     0     0
            c0t5000C50041ABE39Bd0  ONLINE       0     0     0
            c0t5000C50056F9E753d0  ONLINE       0     0     0
            c0t5000C50069010EFFd0  ONLINE       0     0     0
            c0t5000C50088568893d0  FAULTED      0     0     0  too many errors
          mirror-7                 DEGRADED     0     0     0
            c0t5000C50041AC6CDFd0  ONLINE       0     0     0
            c0t5000C5005733979Bd0  ONLINE       0     0     0
            c0t5000C50069010EE7d0  ONLINE       0     0     0
            c0t5000C500893FBCCBd0  FAULTED      0     0     0  too many errors
          mirror-8                 DEGRADED     0     0     0
            c0t5000C50056F9B41Bd0  ONLINE       0     0     0
            c0t5000C50057338E2Fd0  ONLINE       0     0     0
            c0t5000C50069010F33d0  ONLINE       0     0     0
            c0t5000C500893FB6ABd0  FAULTED      0     4     0  too many errors
errors: No known data errors
```

This is the corresponding `fmadm faulty -s` output:

```
root@NodeA-43:# fmadm faulty -s
--------------- ------------------------------------  -------------- ---------
TIME            EVENT-ID                              MSG-ID         SEVERITY
--------------- ------------------------------------  -------------- ---------
Apr 19 07:58:43 3fa07abb-5baa-62a5-d004-c3350dd7a750  ZFS-8000-FD    Major
Apr 19 07:58:28 1761ac01-56ef-6db9-af46-f18c3983c6f7  ZFS-8000-D3    Major
Apr 19 07:58:43 fafa0dfe-b438-6d35-996c-b1e2876fe99f  ZFS-8000-FD    Major
Apr 19 07:58:43 dbe44680-37a7-6c33-afc4-ce2a071ec2fa  ZFS-8000-FD    Major
Apr 19 07:58:43 d524ba96-12cf-efab-e85c-e5b1d1309d9a  ZFS-8000-FD    Major
Apr 19 07:58:43 cf84cc85-e651-edef-fa7d-d9aa4a0748f1  ZFS-8000-FD    Major
Apr 19 07:58:43 b1c32ff5-70c7-c841-e61e-ba40705fed70  ZFS-8000-FD    Major
Apr 19 07:58:43 ab906344-9865-cf5f-d703-8a99e48d095d  ZFS-8000-FD    Major
Apr 19 07:58:43 a866aedb-ba54-e261-fb82-f636ae6eaf4d  ZFS-8000-FD    Major
```

In cases of SCSI initiator faults, the underlying device is retired and cannot be used by ZFS until it has been recovered and marked as repaired. The steps to recover a device depend on the diagnosed problem, but the repair must always be marked for each individual fault using `fmadm acquit <UUID>`. SCSI initiator faults generally include the following problems for active devices:

- A device is not responsive to commands
- A device responds with cumulative SCSI errors that suggest that the device is no longer reliable
- A device returns SCSI errors that are not known to and handled by the SCSI initiator driver
- A device is no longer visible via any host initiator ports

- A device disappears from the ATTO FibreBridge LUN inventory

It is important to keep in mind the significance of device activity and the implications of a bridged topology such as MetroHA. The SCSI protocol generally requires activity to signal exceptions. In a bridged topology, events that the topology itself may present to an initiator port (for example, SAS disconnection) must be mediated in terms of a different topology, and this generally happens as SCSI events rather than topology events. For example: When a SAS device is disconnected, in a directly connected SAS topology, it is a SAS event immediately visible to an initiator port, whereas a bridged SAS device makes it into a LUN inventory change that will only be presented via SCSI when attempting to address any LUN associated with the same target port.

In case of ZFS faults against regular devices, the `zpool clear` command recovers all devices in the pool in a single action. In case of faults against L2ARC and SLOG devices, `fmadm acquit` must be executed for the individual faults. ZFS faults generally include the following:

- A device returned data that did not match its checksum
- A device read or write operation failed
- A device could not be found or did not respond when ZFS attempted a reopen as part of its transaction cycle or was unavailable when the pool was imported (for example, at failover or start)

cfgadm Overview

The `show_SCSI_LUN` cfgadm subcommand shows accessible FC target ports, their LUN inventory, and the state of those LUNs.

This is the corresponding `cfgadm -al -o show_SCSI_LUN` output for the example above:

```
root@NodeA-43:# option expert_mode =1
root@NodeA-43:# !bash
You are about to enter the Unix ("raw") shell and execute low-level Unix command(s).
Warning: using low-level Unix commands is not recommended! Execute? Yes

root@NodeA-43:# cfgadm -al -o show_SCSI_LUN
Ap_Id                          Type         Receptacle   Occupant     Condition
c3                             fc-fabric    connected    configured   unknown
c3::2100001086618fc6,0         processor    connected    configured   unusable
c3::2100001086618fc6,1         disk         connected    configured   unknown
c3::2100001086618fc6,2         disk         connected    configured   unknown
c3::2100001086618fc6,45        ESI          connected    configured   unknown
c3::2100001086618fc6,46        unavailable  connected    configured   unusable
c3::2100001086618fc6,47        unavailable  connected    configured   unusable
c3::2100001086618fc6,48        unavailable  connected    configured   unusable
c3::2100001086618fc6,49        unavailable  connected    configured   unusable
c3::2100001086618fc6,50        unavailable  connected    configured   unusable
c3::2100001086618fc6,51        unavailable  connected    configured   unusable
c3::210000108662751c,98        ESI          connected    configured   unknown
c4                             fc-fabric    connected    configured   unknown
c4::2200001086618fc6,0         processor    connected    configured   unusable
c4::2200001086618fc6,1         disk         connected    configured   unknown
c4::2200001086618fc6,2         disk         connected    configured   unknown
c4::2200001086618fc6,3         disk         connected    configured   unknown
c4::2200001086618fc6,4         disk         connected    configured   unknown
```

```
c4::2200001086618fc6,5          disk         connected    configured   unknown
c4::2200001086618fc6,6          disk         connected    configured   unknown
c4::2200001086618fc6,7          disk         connected    configured   unknown
c4::2200001086618fc6,8          disk         connected    configured   unknown
c4::2200001086618fc6,9          disk         connected    configured   unknown
c4::2200001086618fc6,10         disk         connected    configured   unknown
c4::2200001086618fc6,11         disk         connected    configured   unknown
c4::2200001086618fc6,12         disk         connected    configured   unknown
c4::2200001086618fc6,13         disk         connected    configured   unknown
c4::2200001086618fc6,14         disk         connected    configured   unknown
c4::22000010862751c,98          ESI          connected    configured   unknown
```

This output shows identical subsets of LUNs behind FC target WWPNs `2100001086618fc6`, seen through controller `c3`, and `2200001086618fc6`, seen through controller `c4`, are unavailable. All of this is consistent with loss of a storage enclosure.

When the storage enclosure is powered back on, rescan the LUN inventory using this command:

```
root@NodeA-43:# cfgadm -c configure c3::2100001086618fc6 c4::2200001086618fc6
cfgadm: Library error: failed to create device node: 2100001086618fc6: I/O error
```

Note that the `cfgadm` command incorrectly raises errors with device node creation, which can be ignored. When you check the LUN inventory again with the `cfgadm -al -o show_SCSI_LUN` command, no devices are listed with `unavailable` in the `Type` column and `unusable` in the `Condition` column.

Having confirmed visibility of the devices, you can now clear their state in the pool with this command:

```
root@NodeA-43:# zpool clear metro-repl
```

The `zpool clear` command clears all regular device faults from FMA, and the pool is restored to normal function, as shown by the following commands:

```
root@NodeA-43:# fmadm faulty
root@NodeA-43:# zpool status -v metro-repl
  pool: metro-repl
 state: ONLINE
  scan: resilvered 1.05M in 0h0m with 0 errors on Mon Apr 11 16:56:28 2016
config:
        NAME                        STATE     READ WRITE CKSUM
        metro-repl                  ONLINE       0     0     0
          mirror-0                  ONLINE       0     0     0
            c0t5000C50069009207d0   ONLINE       0     0     0
            c0t5000C500893FBAD3d0   ONLINE       0     0     0
            c0t5000C50041ACCA97d0   ONLINE       0     0     0
            c0t5000C50041ACB4B3d0   ONLINE       0     0     0
          mirror-1                  ONLINE       0     0     0
            c0t5000C50056D5C683d0   ONLINE       0     0     0
            c0t5000C50057339C33d0   ONLINE       0     0     0
            c0t5000C50069010F17d0   ONLINE       0     0     0
            c0t5000C50088A91F6Fd0   ONLINE       0     0     0
          mirror-2                  ONLINE       0     0     0
            c0t5000C500212ADBEBd0   ONLINE       0     0     0
            c0t5000C50041ABCF9Bd0   ONLINE       0     0     0
            c0t5000C5006900A317d0   ONLINE       0     0     0
            c0t5000C500893FB93Fd0   ONLINE       0     0     0
          mirror-3                  ONLINE       0     0     0
```

```
            c0t5000C50057339CD7d0  ONLINE       0     0     0
            c0t5000C50057339F07d0  ONLINE       0     0     0
            c0t5000C50069010F1Fd0  ONLINE       0     0     0
            c0t5000C500893FBBB3d0  ONLINE       0     0     0
          mirror-4                 ONLINE       0     0     0
            c0t5000C50041AC5817d0  ONLINE       0     0     0
            c0t5000C5005733A90Fd0  ONLINE       0     0     0
            c0t5000C50069011CA7d0  ONLINE       0     0     0
            c0t5000C50088A9136Bd0  ONLINE       0     0     0
          mirror-5                 ONLINE       0     0     0
            c0t5000C50056F9DEFBd0  ONLINE       0     0     0
            c0t5000C50057339B77d0  ONLINE       0     0     0
            c0t5000C50069010DA3d0  ONLINE       0     0     0
            c0t5000C500893FC48Bd0  ONLINE       0     0     0
          mirror-6                 ONLINE       0     0     0
            c0t5000C50041ABE39Bd0  ONLINE       0     0     0
            c0t5000C50056F9E753d0  ONLINE       0     0     0
            c0t5000C50069010EFFd0  ONLINE       0     0     0
            c0t5000C50088568893d0  ONLINE       0     0     0
          mirror-7                 ONLINE       0     0     0
            c0t5000C50041AC6CDFd0  ONLINE       0     0     0
            c0t5000C5005733979Bd0  ONLINE       0     0     0
            c0t5000C50069010EE7d0  ONLINE       0     0     0
            c0t5000C500893FBCCBd0  ONLINE       0     0     0
          mirror-8                 ONLINE       0     0     0
            c0t5000C50056F9B41Bd0  ONLINE       0     0     0
            c0t5000C50057338E2Fd0  ONLINE       0     0     0
            c0t5000C50069010F33d0  ONLINE       0     0     0
            c0t5000C500893FB6ABd0  ONLINE       0     0     0
errors: No known data errors
```

In most cases, state and LUN inventory changes are signaled by SCSI responses, but these devices require activity that may not be happening because of previous faults or for lack of ZFS activity; for example, if the node is not running services. In some cases, it is therefore necessary to rescan for target ports or to query individual target ports for their current LUN inventory using the configure subcommand.

In the above output `c3` and `c4` are host FC initiator ports, `c3::2100001086618fc6` and `c4::2200001086618fc6` are FC target ports, and `c3::2100001086618fc6,0` and `c4::220000108662751c,98` are LUNs. Rescanning for target ports uses the `configure` subcommand against a host initiator port, as in `cfgadm -c configure c<n>`. Rescanning a target port for its LUN inventory requires using a the FC target port, as in `cfgadm -c configure c<n>:<WWPN>`.

4

SAS Cable Connections for MetroHA

This chapter includes the following topics:

- Overview
- One FibreBridge and One Storage Enclosure
- One FibreBridge and Two Storage Enclosures
- Two FibreBridges With Two Storage Enclosures Each

Overview

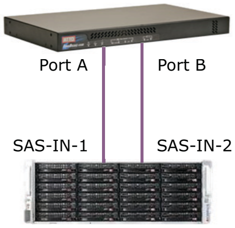
In a MetroHA configuration, ATTO FibreBridges provide the connectivity between the FC fabric(s) and the SAS storage enclosures. MetroHA supports using either one or two FibreBridges per site.

For redundancy, up to two storage enclosures can be connected to each FibreBridge. When two storage enclosures are used, they must be cascaded from the enclosure that is connected to the FibreBridge(s). The illustrations in this section show how the SAS cables are connected between the SAS ports on the FibreBridges and the SAS ports on the storage enclosure.

One FibreBridge and One Storage Enclosure

Figure 4-1 shows the SAS port connections when a single FibreBridge is connected to a single storage enclosure. In this example, both SAS ports on the FibreBridge unit are connected to SAS ports on the storage enclosure, providing a redundant link between the devices.

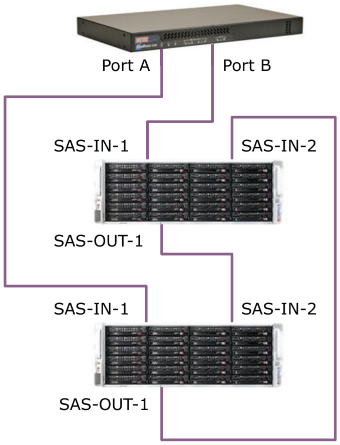
Figure 4-1: SAS Port Connections for One FibreBridge and One Storage Enclosure



One FibreBridge and Two Storage Enclosures

When two storage enclosures are connected to a FibreBridge, as shown in Figure 4-2, both storage enclosures must have one direct SAS uplink to the FibreBridge. Each storage enclosure must also cascade through the other to provide a second path.

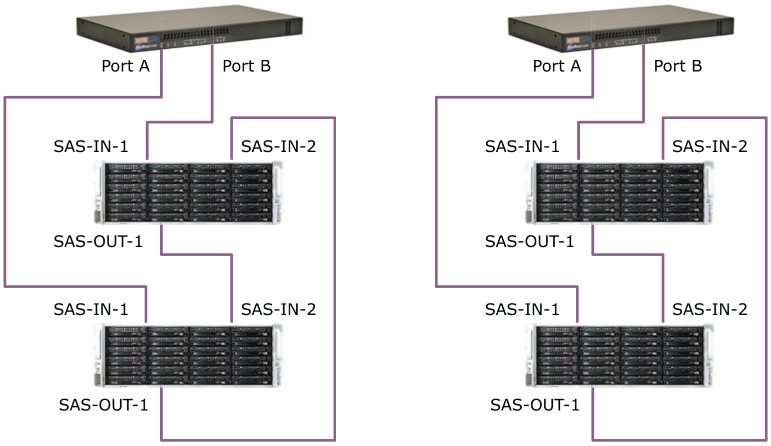
Figure 4-2: SAS Port Connections for One FibreBridge and Two Storage Enclosures



Two FibreBridges With Two Storage Enclosures Each

If MetroHA is deployed in a four-FibreBridge configuration, with two FibreBridges at each site, a given storage enclosure or pair of storage enclosures must connect through only one FibreBridge, not through both FibreBridges. Figure 4-3 shows a configuration where two FibreBridges are deployed at a site, with two storage enclosures connected to each FibreBridge.

Figure 4-3: SAS Port Connections for Two FibreBridges and Two Storage Enclosures



Global Headquarters
451 El Camino Real, Suite 201
Santa Clara, CA 95050
USA

3000-nxsmetro-000001-A