NETAPP TECHNICAL REPORT

MetroCluster Design and Implementation Guide

Jim Lanson, NetApp
TR-3548

ABSTRACT

This document is intended to serve as a guide for architecting and deploying MetroCluster in a customer environment. This design and implementation guide describes the basic MetroCluster architecture, considerations for deployment, and best practices. As always, please refer to the latest technical publications on the NOW™ (NetApp on the Web) site for specific updates on processes, Data ONTAP® command syntax, and the latest requirements, issues, and limitations. This document is intended for field personnel who require assistance in architecting and deploying a MetroCluster solution.

Table of Contents

1 INTRODUCTION

1.1 INTENDED AUDIENCE

The information in this document is intended for field personnel and administrators who are responsible for architecting and deploying successful MetroCluster high-availability and disaster-recovery configurations. A brief overview of MetroCluster is presented in order to establish baseline knowledge before discussing implementation planning, installation, configuration, and operation.

1.2 SCOPE

This document covers the following MetroCluster configurations:

- Stretched or nonswitched MetroCluster (NetApp® storage)
- Fabric or switched MetroCluster (NetApp storage)

Topics that apply only to Stretched MetroCluster refer to Stretch MetroCluster.

Topics that are specific to NetApp storage-based Fabric MetroCluster refer to Fabric MetroCluster.

Topics that apply to all configurations refer simply to MetroCluster.

Other than a short description, V-Series MetroCluster is not covered in this document.

This document refers mostly to Fabric MetroCluster, because Fabric MetroCluster configurations introduce an increased level of complexity in design and implementation. To many, Stretch MetroCluster is simply an active-active configuration with longer cables and mirroring. The introduction of Fibre Channel switches and longer distances requires much more consideration and discussion. The operational sections (on creating mirrors, forced takeover, etc.) apply to both.

1.3 REQUIREMENTS AND ASSUMPTIONS

For the methods and procedures described in this document to be useful to the reader, the following assumptions are made:

The reader has at least basic NetApp administration skills and has administrative access to the storage system via the command-line interface.

The reader has a full understanding of active-active configurations as they apply to the NetApp storage controller environment.

The reader has at least a basic understanding of Fibre Channel switch technology and operation, along with access to the switches via command line.

In the examples in this report, all administrative commands are performed at the storage system or Fibre Channel switch command line.

2 OVERVIEW

2.1 FEATURES

MetroCluster configurations consist of a pair of active-active storage controllers configured with mirrored aggregates and extended distance capabilities to create a high-availability solution. The primary benefits include:

- Higher availability with geographic protection
- Minimal risk of lost data, easier management and recovery, and reduced system downtime
- Quicker recovery when a disaster occurs
- Minimal disruption to users and client applications

2.2 OPERATION

A MetroCluster (either Stretch or Fabric) behaves in most ways just like an active-active configuration. All of the protection provided by core NetApp technology (RAID-DP®, Snapshot™ copies, automatic controller failover) also exists in a MetroCluster configuration (Figure 1). However, MetroCluster adds complete synchronous mirroring along with the ability to perform a complete site failover from a storage perspective with a single command.

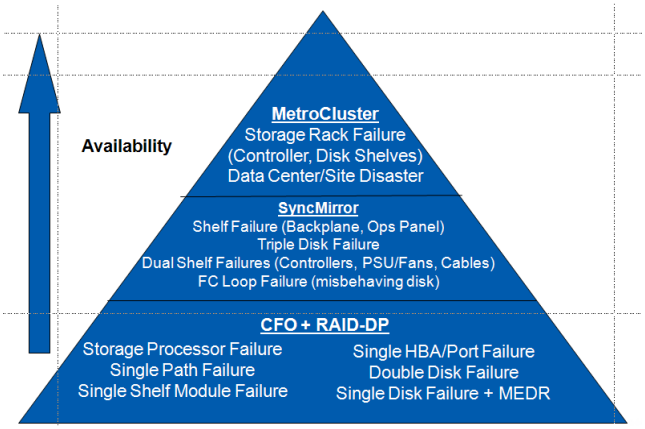


Figure 1) Levels of protection.

2.3 METROCLUSTER TYPES

Stretch MetroCluster (sometimes referred to as nonswitched) is simply an active-active configuration that can extend up to 500m depending on speed and cable type. It also includes synchronous mirroring (SyncMirror®) and the ability to do a site failover with a single command. See Figure 2. Additional resiliency can be provided through the use of multipathing.

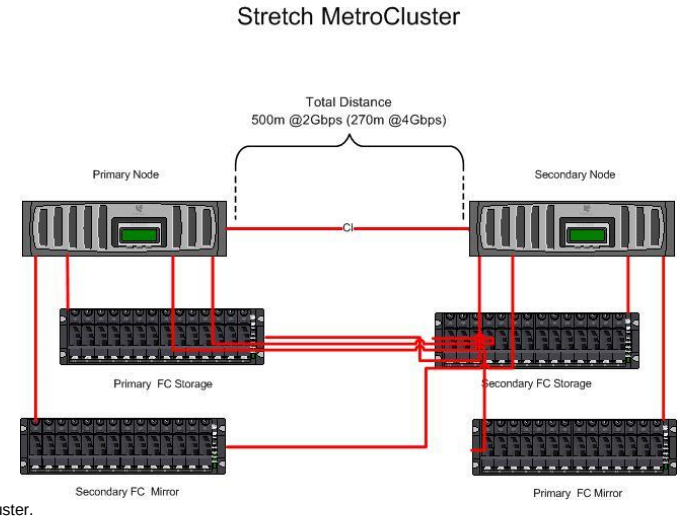


Figure 2) Stretch MetroCluster.

Fabric MetroCluster (also referred to as switched) uses four Fibre Channel switches in a dual-fabric configuration and a separate cluster interconnect card to achieve an even greater distance (up to 100km depending on speed and cable type) between primary and secondary locations. See Figure 3.

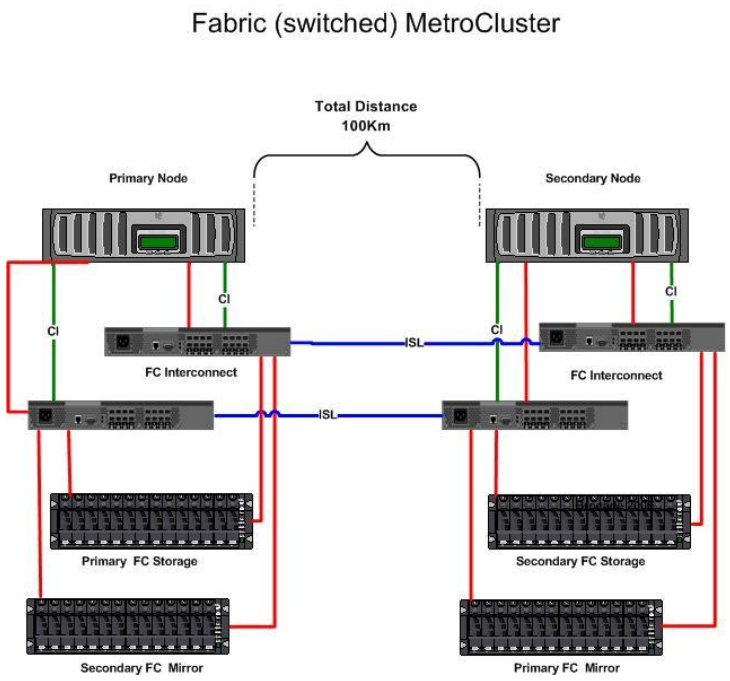


Figure 3) Fabric MetroCluster.

V-Series MetroCluster is simply either of the above configurations with a NetApp V-Series system. Because of the architectural differences between V-Series and a standard active-active configuration, V-Series MetroCluster has additional flexibility when it comes to the maximum number of disk spindles and the Fibre Channel switches. See the V-Series Compatibility Guide on the NOW site for additional information. The implementation of V-Series MetroCluster is outside the scope of this document. See the V-Series documentation on the NOW site.

For more information about MetroCluster, see the *Data ONTAP Cluster Installation and Administration Guide* (for Data ONTAP 7.1) or the *Data ONTAP Active-Active Configuration Guide* (for Data ONTAP 7.2 or later). (For Data ONTAP 7.0 and earlier, see the *Data Protection Online Backup and Recovery Guide* and the *NetApp Hardware Cluster Guide*.)

Note: When referring to the documentation just listed, use the wiring diagrams in this technical report. The diagrams in this document facilitate easier installation and expansion of MetroCluster.

2.4 MIRRORING

NetApp SyncMirror, an integral part of MetroCluster, combines the disk-mirroring protection of RAID 1 with NetApp's industry-leading RAID 4 and RAID-DP technology. In the event of an outage—whether it's due to a disk problem, cable break, or host bus adapter (HBA) failure—SyncMirror can instantly access the mirrored data without any operator intervention or disruption to client applications. SyncMirror maintains a strict physical separation between the two copies of your mirrored data. Each of these copies is referred to as a plex. Each controller's data has its "mirror" at the other site.

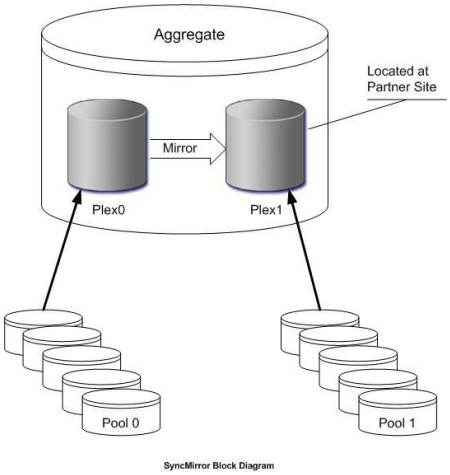


Figure 4) SyncMirror pools and plexes.

When SyncMirror is licensed and hardware ownership is used (discussed in "Disk Ownership," below), spare disks are split into two pools—Pool0 and Pool1. Each plex of a mirror uses disks from these separate pools. When software ownership is used, disks are explicitly assigned to pools by the administrator.

To maximize availability, Pool0 and Pool1 disks need to be on separate loops and use separate HBAs, cables, and shelves.

Before enabling the SyncMirror license, ensure that disks for each pool are located on the appropriate loops that are fault isolated from each other.

- ✓ ***Best Practice:*** Make sure all storage is set up under SyncMirror. While nonmirrored storage is technically permissible, NetApp does not recommend it, since the data on that storage will not be available after a site failover.

2.5 DISK OWNERSHIP

In a MetroCluster configuration in which disk shelves on each side are mirrored to the other side and thus accessible by either controller, disk ownership comes into play. There are two methods of establishing disk ownership: hardware and software. Hardware-based ownership is the default for the 900 series and the FAS3020/3050. All other platforms (V-Series, FAS6000 series, FAS3040/3070, FAS31xx) use software disk ownership. This capability became available in Data ONTAP 6.3.1.

Please remember that while the FAS3020/3050 may be technically able to support software disk ownership, it can't be supported in a Fabric MetroCluster. A brief description of each method follows. For more detail, see "Installation and Configuration" later in this document or the Data ONTAP documentation.

Hardware disk ownership establishes which controller owns which disks by how the shelves are connected. For more information, see the *System Configuration Guide*:

HTTP://NOW.NETAPP.COM/NOW/KNOWLEDGE/DOCS/HARDWARE/NETAPP/SYSCFG/

Table 1) Hardware disk ownership.

Model	Pool Ownership
FAS9xx	Slots 2, 3, 4, 5, and 7 are Pool0
	Slots 8, 9, and 10 are Pool1
	Optional software-based ownership and pool selection (as of Data ONTAP 7.1.1; Stretch MetroCluster only)
FAS3020/3050	0a, 0b, and slots 1 and 2 are Pool0 (slot 1 is usually NVRAM)
	0c, 0d, and slots 3 and 4 are Pool1
	Optional software-based ownership and pool selection (as of Data ONTAP 7.1.1; Stretch MetroCluster only)
FAS3040/3070	Software-based ownership and pool selection
FAS31xx	Software-based ownership and pool selection
FAS60xx	Software-based ownership and pool selection
V-Series	Software-based ownership and pool selection

Note: Physical connectivity is extremely important to provide adequate distribution of disks between the two pools.

Software disk ownership allows greater flexibility in cabling by allowing disk ownership to be assigned through explicit commands. Rather than determining which controller owns which disk by hardware connection, the ownership information is written on each disk. Because disk ownership can be implemented by a system administrator, it is important that the configuration maximizes availability. For more information, see "Installation and Configuration," later in this document.

2.6 FIBRE CHANNEL SAN IN A METROCLUSTER WORLD

For those who are experienced in Fibre Channel technology, storage area networks (SANs) in particular, there are some differences and restrictions worth discussing relative to how Fabric MetroCluster utilizes this technology.

Fabric MetroCluster configurations use Fibre Channel switches as the means to separate the controllers by greater distances. The switches are connected between the controller heads and the disk shelves, and to each other. Each disk spindle or LUN individually logs into a Fibre Channel fabric. Except for the V-Series, the nature of this architecture requires, for performance reasons, that the two fabrics be completely dedicated to Fabric MetroCluster. Extensive testing was performed to provide adequate performance with switches included in a Fabric MetroCluster configuration. For this reason, Fabric MetroCluster requirements prohibit the use of any other model or vendor of Fibre Channel switches than the switches included with the Fabric MetroCluster.

For performance and functional reasons, there is a current disk spindle limit equal to the lesser of either the model limit or 504. Higher spindle count solutions will be available in the future. Keep in mind that this is for Fabric MetroCluster only. The maximum number of disks in a Stretch MetroCluster depends solely on which NetApp platform is deployed. Refer to the *System Configuration Guide* on the NOW site for the specific platform limit.

In a traditional SAN, there is great flexibility in connecting devices to ports as long as the ports are configured correctly and any zoning requirements are met. In a MetroCluster, when using hardware ownership, Data ONTAP expects certain devices to be connected to specific ports or ranges of ports. It is therefore critical that cabling be exactly as described in the installation procedures. Also, no switch-specific functions such as trunking or zoning are currently used in a hardware ownership–based NetApp Fabric MetroCluster (non-V-Series).

2.7 RESTRICTIONS

For further restrictions in using a Fabric MetroCluster configuration, see the *Data ONTAP Cluster Installation and Administration Guide* (for Data ONTAP version 7.1) or the *Data ONTAP Active-Active Configuration Guide* (for Data ONTAP version 7.2 or later). (For Data ONTAP 7.0 and earlier, see the *Data Protection Online Backup and Recovery Guide* and the *NetApp Hardware Cluster Guide*.)

2.8 COMPONENTS

A NetApp Stretch MetroCluster includes the following components:

- Standard active-active pair of FAS900, 3000, 3100 (two single-controller chassis), or 6000 series controllers running Data ONTAP 6.4.1 or later
  (See the MetroCluster Compatibility Matrix on the NOW site for supported models and Data ONTAP releases.)
- A FC/VI cluster adapter (31xx only)
- A syncmirror_local license
- A cluster_remote license
- A cluster license
- Copper/Fibre converters for cluster interconnect (9xx, 30xx, 6xxx only)
- Associated cabling

A NetApp Fabric MetroCluster includes the following components:

- An active-active pair of FAS900, 3000, 3100 (two single controller chassis), or 6000 series controllers running Data ONTAP 6.4.1 or later
  (See the MetroCluster Compatibility Matrix on the NOW site for supported models.)
- Four Brocade Fibre Channel switches with supported firmware supplied by NetApp
  (See the MetroCluster Compatibility Matrix on the NOW site for supported models.)
  There is a pair at each location. Supported models may differ between locations but must be the same at each given location.
- Brocade Extended Distance license (if over 10km)
- Brocade Full-Fabric license
- Brocade Ports-on-Demand (POD) licenses for additional ports
- A VI-MC cluster adapter
- A syncmirror_local license
- A cluster_remote license
- A cluster license
- Associated cabling

2.9 SPECIAL CONSIDERATIONS FOR THE FAS31XX SERIES

In normal Stretch MetroCluster configurations, the cluster interconnect (CI) on the NVRAM cards in each controller is used to provide the path for CI traffic. The new FAS3140 and 3170 offer a new architecture that incorporates a dual-controller design with the cluster interconnect on the backplane. For this reason, the FCVI card that is normally used for CI in a Fabric MetroCluster configuration must also be used for a 31xx Stretch configuration.

It is important to remember that this will not allow one 31xx cluster to connect to another via MetroCluster. The only supported configuration is single controller to single controller (Stretch or Fabric). See Figures 5a and 5b.

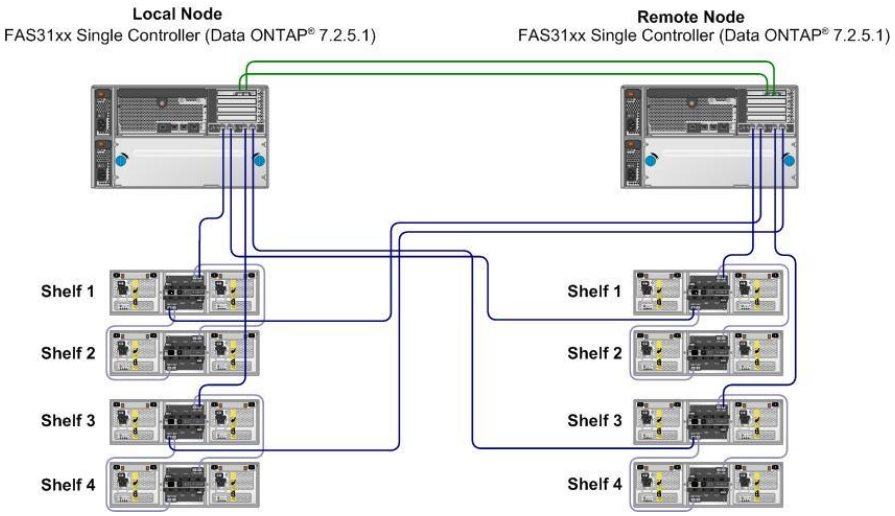


Figure 5a) FAS31xx Stretch MetroCluster.

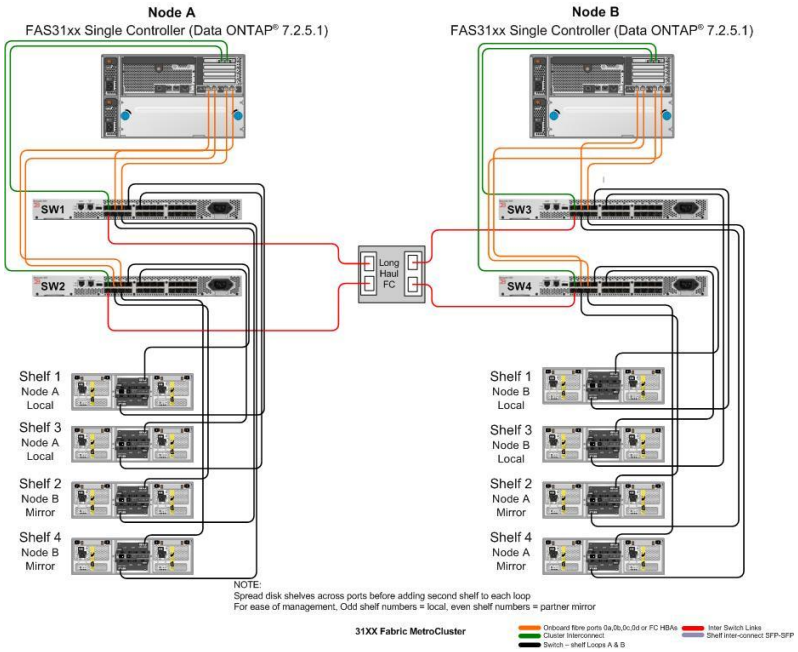


Figure 5b) FAS31xx Fabric MetroCluster.

2.10 SPECIAL CONSIDERATIONS FOR METROCLUSTER AND FAS DEDUPLICATION

1. FAS deduplication functionality is currently supported on the following Data ONTAP releases:
    a. Stretch MetroCluster - Data ONTAP 7.2.5.1 or later, 7.3.1 or later
    b. Fabric MetroCluster – Data ONTAP 7.2.5.1 or later, 7.3.1 or later

    Check the MetroCluster Compatibility Matrix on the NOW site for current information.
2. FAS deduplication is supported on 30xx, 31xx, and 6xxx series platforms.
3. The CPU impact of deduplication generating extra disk write operations will be greater on a MetroCluster because of writing to two plexes. The expected CPU impact on most platforms is less

than 10%. The impact is seen more on lower-end platforms (e.g., 30xx) than on higher-end (6xxx) systems.

4. While in takeover mode, fingerprints of writes to partner flexible volumes will be logged to the change log. The deduplication process will not be run on partner flexible volumes. Upon giveback, data in the change logs will be processed and data will get deduplicated.
5. Also while in takeover mode, logging of fingerprint writes to partner flexible volumes will stop once the change log file is full. This is especially possible if the node is in takeover mode for a long period of time, for example, during a site disaster.
6. A node in takeover mode, in addition to servicing I/Os targeted at partner volumes, will have to handle logging of fingerprints associated with those I/Os to the change log. Workload on a single controller will have to be appropriately adjusted.
7. Only a subset of deduplication commands for partner volumes is available in takeover mode.
8. Deduplication must be licensed on both nodes.

For additional information regarding FAS deduplication, refer to TR-3505, NetApp Deduplication for FAS Deployment and Implementation Guide.

2.11 METROCLUSTER VERSUS STANDARD SYNCHRONOUS REPLICATION

As a high-availability solution, MetroCluster is often compared with other synchronous replication products. Even though it includes SyncMirror, MetroCluster can be differentiated by the following features:

- Low aggregate level RAID mirroring (less performance impact)
- Automatic switchover to remote copy upon failure
- Site failover with a single command
- Simpler to manage than multiple replication relationships
- No extensive scripting required to make data available after failover

2.12 FIBRE CHANNEL SWITCH OVERVIEW

OPERATION OVERVIEW

The NetApp Fabric (switched) MetroCluster configuration uses four Brocade Fibre Channel switches in a dual-fabric configuration to connect two active-active controllers. These switches cannot be combined with any other switch model and must be supplied by NetApp.

SWITCH FABRIC CONFIGURATION

A Fabric MetroCluster contains two switch fabrics. A switch fabric consists of a switch on the primary controller connected to a switch on the remote controller (see Figure 6). The two switches are connected to each other through ISL (Inter-switch Link) cables.

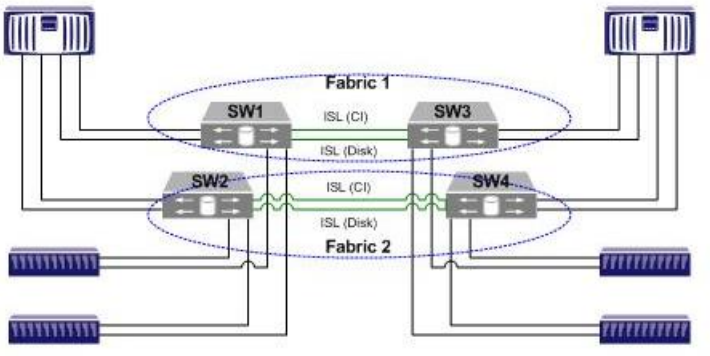


Figure 6) Dual-fabric MetroCluster.

Figure 6 shows a Fabric MetroCluster configuration in which the first fabric begins at switch “SW1” on the primary side and is completed by connecting an ISL cable (or two if using traffic isolation, described later) to the switch “SW3“ on the remote side. The second fabric is created by using switch “SW2” on the primary side, connected through another ISL cable (or two) to the second switch “SW4” on the remote side. The reason for two fabrics is redundancy. The loss of a switch in a fabric or the loss of a fabric will not affect the availability of the Fabric MetroCluster.

Due to the nature of the MetroCluster architecture, Fibre Channel traffic on the switch fabric includes both disk I/O traffic between the controllers and disk shelves and cluster interconnect traffic. As you can probably envision, disk problems can generate excessive traffic and result in a bottleneck for the cluster interconnect. To alleviate this potential condition, several techniques are used.

VIRTUAL CHANNELS (SOFTWARE DISK OWNERSHIP)

All supported models of switches (Brocade 200E, 5000, 300, and 5100) use virtual channels to help separate traffic when the distance between switches is less than 10km. The virtual channels are allocated as follows.

Table 2) Virtual channel assignments.

	32-port Switch	Usage
VC #		
2	0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40	FC/VI, ISL
3	1, 5, 9, 13, 17, 21, 25, 29, 33, 37	FC ports, disk shelves
4	2, 6, 10, 14, 18, 22, 26, 30, 38	FC ports, disk shelves
5	3, 7, 11, 15, 19, 23, 27, 31, 39	FC ports, disk shelves

Consequently, when using software ownership it is important to isolate cluster interconnect (FC/VI) and ISL traffic from storage traffic. When the distance is greater than 10km, all virtual channels are combined into one for buffer credit purposes.

ZONING (SOFTWARE DISK OWNERSHIP)

Effective in Data ONTAP 7.2.4 on MetroCluster systems using software disk ownership, Fibre Channel switch zones are created to help isolate disk traffic from cluster interconnect traffic. Further details can be found in the Brocade Switch Configuration Guide on the NOW site or in the installation section of this document.

TRAFFIC ISOLATION

Effective in Data ONTAP 7.2.6.1 on MetroCluster systems using software disk ownership and the Brocade 5000 or 5100 switches, separate ISLs are can be connected on each fabric to isolate disk traffic from cluster interconnect traffic. This is used in conjunction with the Traffic Isolation feature mentioned below. Further details can be found in the Brocade Switch Configuration Guide on the NOW site or in the installation section of this document.

Effective in Brocade Fabric Operating System (FOS) version 6.0.0b, it is now possible to dedicate interswitch links to certain traffic. The Traffic Isolation feature allows you to control the flow of interswitch traffic by creating a dedicated path for traffic flowing from a specific set of source ports (N_Ports). For example, you might use Traffic Isolation for the following scenarios:

- To dedicate an ISL to high-priority cluster-interconnect traffic such as NVRAM mirroring traffic
- To isolate high-priority traffic from the disruptions that may be caused by N_ports and E_ports handling high volume but low-priority traffic

Traffic Isolation is implemented using a special zone, called a *Traffic Isolation zone* (TI zone). A TI zone indicates the set of ports and ISLs to be used for a specific traffic flow. When a TI zone is activated, the fabric attempts to isolate all inter-switch traffic entering from a member of the zone to only those ISLs that have been included in the zone. The fabric also attempts to exclude traffic not in the TI zone from using ISLs within that TI zone.

Figure 7 shows a fabric with a TI zone consisting of N_Ports “1,8” and “4,6” and E_Ports “1,1,” “3,9,” “3,12,” and “4,7.” The dotted line indicates the dedicated path from Domain 1 to Domain 4.

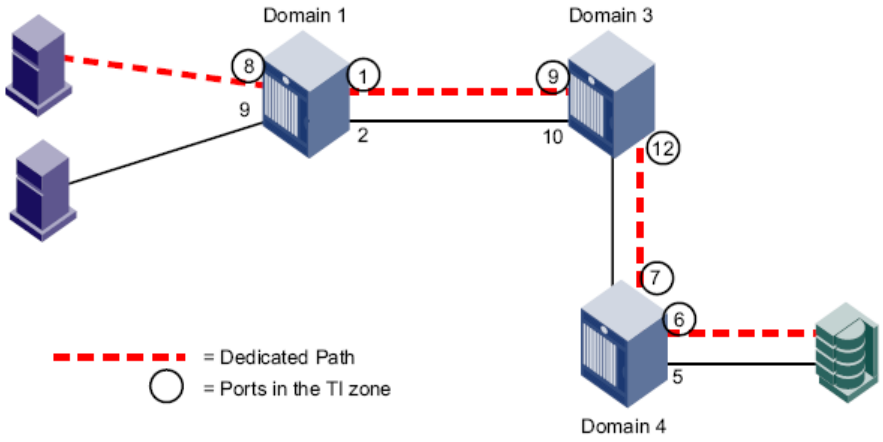


Figure 7) Traffic Isolation zones.

In Figure 7, all traffic entering Domain 1 from port 8 is routed through the ISL on port 1. Similarly, traffic entering Domain 3 from port 9 is routed to the ISL on port 12, and traffic entering Domain 4 from the ISL on port 7 is routed to the device through port 6. Traffic coming from other ports in Domain 1 would *not* use port 1, but would use port 2 instead.

Other traffic is excluded from the dedicated path as long as other equal-cost routes through the fabric exist. For example, if the ISL formed by E_Ports “1,2” and “3,10” failed, *all* traffic between Domains 1 and 3 would use the ISL formed by E_Ports “1,1” and “3,9,” even though that ISL is a dedicated path in a TI zone.

Use the zone command to create and manage TI zones. Refer to the Brocade *Fabric OS Command Reference* for details about the zone command.

TI ZONE FAILOVER

A TI zone can have failover enabled or disabled. If the dedicated path cannot be used and failover is enabled, the TI zone traffic will use a nondedicated path instead. If the dedicated path cannot be used and failover is disabled, the traffic isolation path is broken and traffic for that TI zone is halted until the dedicated

path is fixed. For example, in Figure 7, if the dedicated ISL between Domain 1 and Domain 3 goes off line, then the following occurs, depending on the failover option:

- If failover is enabled for the TI zone, the traffic is routed from Domain 1 to Domain 3 through E_Ports “1,2” and “3,10.”
- If failover is disabled for the TI zone, the traffic is halted until the ISL between Domain 1 and Domain 3 is back on line.

GENERAL RULES FOR TRAFFIC ISOLATION ZONES

Note the following general rules for TI zones:

- A given N_Port can be a member of only a single TI zone. This rule is enforced during zone creation or modification.
- An E_Port can be a member of only a single TI zone. The same checking is done as described for N_Ports.
- If multiple E_Ports are configured that are on the lowest-cost route to a domain, the various source ports for that zone are load-balanced across the specified E_Ports.
- The TI zones appear in the defined zone configuration only and do not appear in the effective zone configuration. A TI zone only provides Traffic Isolation and is not a “regular” zone.
- A TI zone must include E_Ports and N_ports in order to form an end-to-end dedicated path between two N_Ports.
- Each TI zone is interpreted by each switch and each switch considers only the routing required for its local ports. No consideration is given to the overall topology and to whether the TI zones accurately provide dedicated paths through the whole fabric.
- You create and modify TI zones using the zone command. Other zoning commands, such as zoneCreate, aliCreate, and cfgCreate, cannot be used to manage TI zones.

CONFIGURATION RULES FOR TRAFFIC ISOLATION

- Fabric MetroCluster supports Traffic Isolation only on the Brocade 5000, 300, and 5100 switches.
- Ports in a TI zone must belong to switches that run Fabric OS v6.0.0b or later.
- Traffic Isolation is not supported in fabrics with switches running firmware versions earlier than Fabric OS v6.0.0b. However, the existence of a TI zone in such a fabric is backward compatible and does not disrupt fabric operation in switches running earlier firmware versions.

LICENSES

A number of licenses are required for the Fibre Channel switches. When ordered with a Fabric MetroCluster configuration, the switches should include all necessary licenses. For reference, they are:

- Full-Fabric License
- Extended Distance License (if over 10km)
- Brocade Ports-on-Demand (POD) licenses for additional ports

SWITCH PORT ALLOCATION – HARDWARE DISK OWNERSHIP

For MetroCluster systems using the hardware disk ownership model, the Fibre Channel switch ports are divided into banks and pools. Each switch has two banks, dividing the switch into two equal parts. (See Figure 8.) A storage controller connects to a bank through its Fibre Channel ports (HBA or embedded). It then owns the disks that are connected to ports of the opposite bank. If the storage system connects to bank 1, only shelves connected to bank 2 belong to it. Additionally, Pool0 disks are considered “local”—all reads will be satisfied from those disks only. Pool1 disks are “remote”; they will be written to, but cannot be read

from, under normal circumstances. Reads from both plexes may be enabled through an option setting. Together, these two rules limit the usable ports for a certain disk pool to 4 on a 16-port switch and to 2 on an 8-port switch.

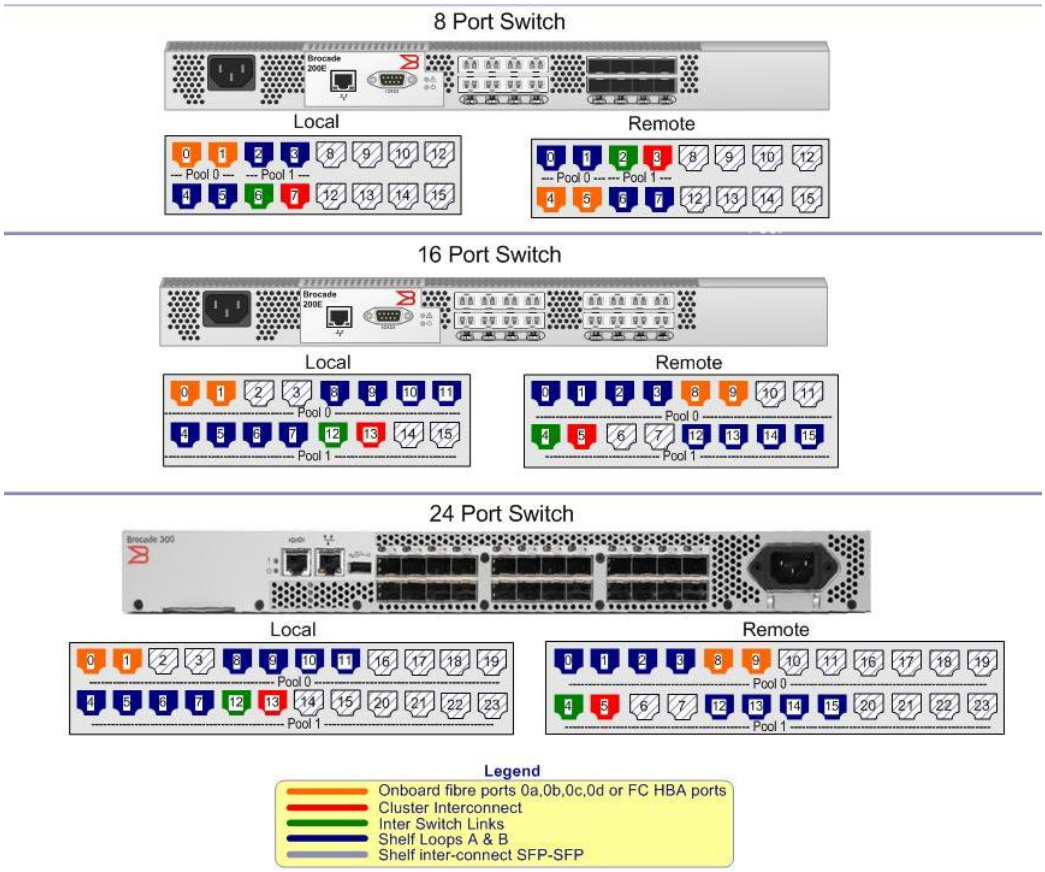


Figure 8) Switch port assignments with hardware ownership.

8-Port Switches

As shown in Figure 8, ports 0-1 and 4-5 belong to Pool0, while ports 2-3 and 6-7 belong to Pool1.

The Brocade 300 24-port switch has ports 0-7 licensed and operational, as shown in Figure 8.

16-Port Switches

As shown in Figure 8, ports 0-3 and 8-11 belong to Pool0, while ports 4-7 and 12-15 belong to Pool1.

The Brocade 300 24-port switch has ports 0-15 licensed and operational, as shown in Figure 8.

32-Port and 40-port Switches

The Brocade 5000 and 5100 are not supported using the hardware disk ownership model.

3 DEPLOYMENT PLANNING

3.1 DATA GATHERING

CUSTOMER REQUIREMENTS

To facilitate a successful MetroCluster installation, the following information should be gathered early on.

ENVIRONMENTAL

Distance between the primary and remote sites. This information is necessary to determine which type of MetroCluster is appropriate, or even whether either version is appropriate. In calculating the effective distance, factors such as cable type, speed, and number of patch panels must be considered. (See Section 3.2, "Distance Considerations.") Although NetApp recommends that dedicated dark fiber be used for a MetroCluster configuration, WDM devices are supported. Refer to the Brocade Compatibility Guide at www.brocade.com for supported devices.

SETUP

Gather the following items before you begin any MetroCluster installation:

- NetApp licenses
- Hostnames and IP addresses for each of the nodes and the Fibre Channel switches
- Brocade switch licenses; this is extremely important since there can be a delay in obtaining these licenses
- Appropriate number and type of cables with appropriate connectors

3.2 DISTANCE CONSIDERATIONS

OVERVIEW

Stretch MetroCluster can support a maximum of 500 meters between nodes at a speed of 2Gbps. Fabric MetroCluster, through the use of Fibre Channel switches, extends this distance to 100km at the same speed. At 4Gbps speeds, these distances are roughly cut in half unless using the Brocade 300, 5000 or 5100, which leaves this maximum distance at 100km. This extended distance capability gives customers greater flexibility in the physical location of their active-active controllers while maintaining the high-availability benefits of clustered failover.

This section describes a number of factors that affect the overall effective distance permissible between the MetroCluster nodes.

- Physical distance
- Number of connections
- Desired speed
- Cable type

PHYSICAL DISTANCE

As stated earlier, the Stretch MetroCluster configuration can extend to a maximum of 500m (2Gbps). This distance is reduced by speed, cable type, and number of connections. A Fabric MetroCluster can extend out to 100km. This distance is affected by the same factors. At a distance of 100km, latency would be around 1ms. Greater distances would obviously result in greater latencies (500km = 5ms), which may be unacceptable to an application.

CABLE TYPE

As shown in Table 3, the cable type affects both distance and speed. Single-mode cable is supported only for the inter-switch links.

Example 1: A customer has 250 meters between sites and wants to run at 4Gbps. The OM-3 cable type is required.

Example 2: A customer currently has a MetroCluster configuration running at 2Gbps with a distance of 300 meters over OM2 cabling and wants to upgrade to 4Gbps speeds. Upgrading the cabling will not help, because OM3 has a maximum of 270 meters. In this case the choices would be:

- Remain at 2Gbps speeds. Customers with the new ESH4 disk shelves could still use them at this distance, as long as the shelf speed is set to 2Gbps.
- Test current optical network infrastructure to make sure that attenuation and latency are acceptable.

Table 3) Cable types and distances.

Fiber Type	Data Rate	Max Distance (M)
OM-2 (50/125UM)	1Gb/s	500
	2Gb/s	300
	4Gb/s	150
OM-3 (50/125UM)	1Gb/s	860
	2Gb/s	500
	4Gb/s	270
OM-3+	1 Gb/s	1100
	2 Gb/s	750
	4 Gb/s	500
OS1 Single Mode (9/125UM)	2Gb/s	10,000*
	4Gb/s	10,000*

*The maximum distance shown here is typically due to the standard 1310nm SFPs. Use of high-power SFPs can extend this dark fiber up to 30km. Using 1550nm high-power SFPs, a distance of 70–100km can be achieved.

This topic is discussed in much greater technical detail in the following technical reports:

- MetroCluster Upgrade Planning Guide (TR-3517)
- Optical Network Installation Guide (TR-3552)

3.3 PHYSICAL LAYOUT

DISK LOOPS

In order to optimize performance, NetApp recommends that instead of installing full loops of disks (two shelves maximum on a Fabric MetroCluster) you install disk shelves one per loop until all disk loop ports are utilized. Add-on shelves can then be added to each shelf. See Figure 9.

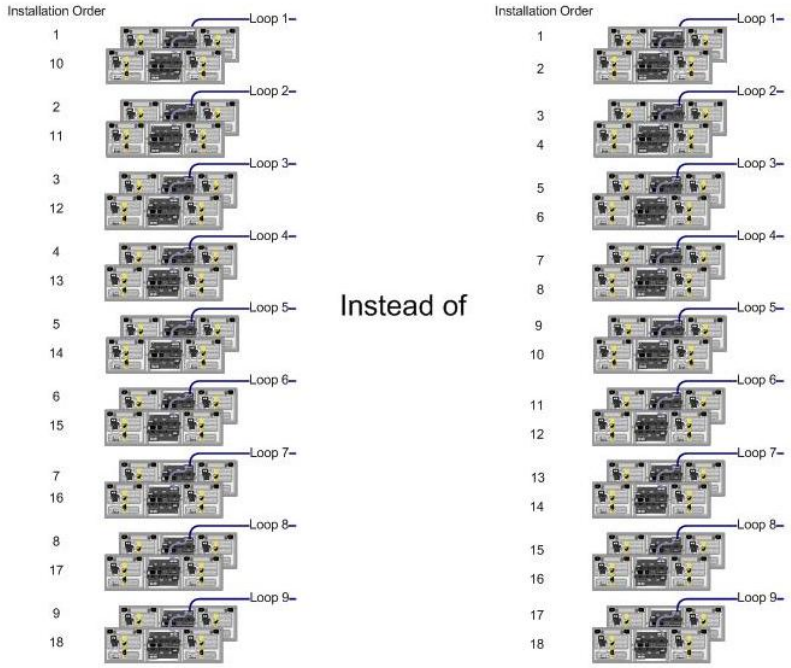


Figure 9) Disk shelf installation order.

CABINETS

Although it is not mandatory for the equipment to be installed as shown in Figures 10a and 10b, this is a physically desirable configuration that facilitates ease of cabling and scalability for expansion. For ease of management and troubleshooting, it is helpful to set shelf IDs based on location. For example, all disk shelves on the primary controller are set to odd numbers (1, 3, 5) while shelves on the remote controller are set to even numbers. This makes identification and location of shelves much easier when using Data ONTAP utilities.

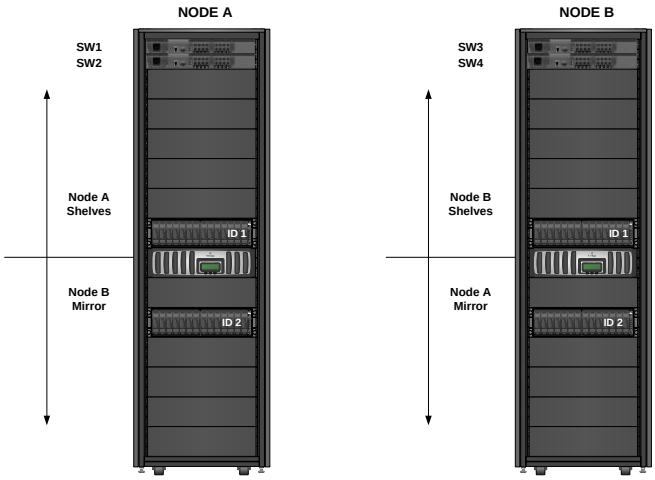


Figure 10a) Typical Fabric MetroCluster rack installation (minimal).

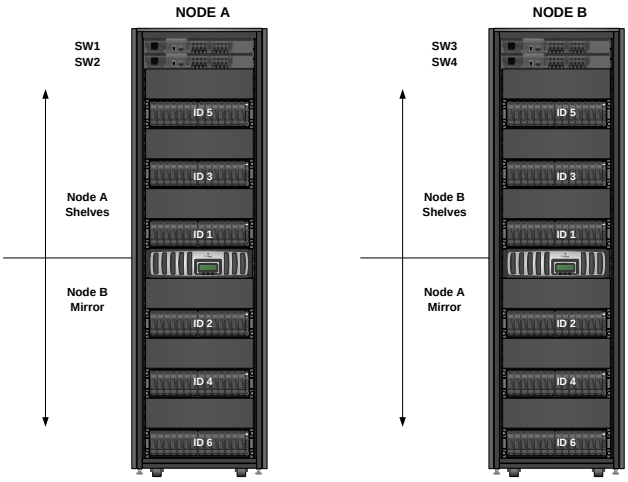


Figure 10b) Typical Fabric MetroCluster rack installation (shown half populated).

3.4 CLUSTER CONFIGURATION CHECKER

The Cluster Configuration Checker is a Perl script that detects errors in the configuration of a pair of active-active NetApp controllers. It can be run as a command from a UNIX® shell or Windows™ prompt, but also doubles as a CGI script that can be executed by a UNIX Web server. The script uses rsh or ssh to communicate with the controllers you're checking, so you must have the appropriate permissions for rsh to run on both controllers in the cluster pair. This script detects and reports the following:

- Services licensed not identical on partner (some services may be unavailable on takeover)
- Options settings not identical on partner (some options may be changed on takeover)
- Network interfaces configured incorrectly (clients will disconnect during takeover)
- FCP cfmode settings not identical on controllers with FCP licensed
- Checks /etc/rc on each controller to see that all interfaces have a failover set

This script is available on the NOW site for download. NetApp recommends that it be run as part of the implementation process.

If the controllers being implemented were part of an active-active configuration, then the configurations are probably compatible. It never hurts to run this utility anyway, just to be certain.

✓ ***Best Practice***: NetApp strongly recommends that the installation planning workbook be completed prior to beginning the installation. A little time spent up front will expedite a successful installation process.

4 INSTALLATION AND CONFIGURATION

As in most successful installations, the key is planning and preparation. The collected information outlined in section 3 is essential for completing the installation. The steps of a MetroCluster installation can be divided according to local or remote site. This section outlines these steps. Refer to the Active-Active Configuration Guide on the NOW site for supplementary information.

4.1 LOCAL SITE

VERIFY CONTROLLER LICENSES

Missing licenses take time to obtain, so the first step should be to verify not only that the licenses are available but that they are installed. Follow the steps under "Storage Controller" and “Fibre Channel Switches" to validate the licenses.

AT THE STORAGE CONTROLLER:

COMMAND	RESULT
telnet controllername (or IPAddr)	Login prompt
Login root	Prompt for password
"password"	Command line prompt
License	Look for the following to be licensed: Cluster syncmirror _local cluster_remote If any of the above are not licensed, perform the following step.
License add "license key"	When adding licenses, remember to add them in the above order.

VERIFY FIBRE CHANNEL SWITCH LICENSES

If this is a brand new switch, the IP address may not be set. The default address is 10.77.77.77. The default user/password is `admin/password`. If network connectivity is not yet available, then any switch setup must be performed using the console port on the switch itself. Refer to the Brocade documentation for more information.

COMMAND	RESULT
telnet "switch IPAddress"	Login prompt
Login admin	Prompt for password
"password"	Command line prompt
licenseshow	Look for the following to be licensed: *brcd200e_whql01:admin> licenseshow* *Web license* *Zoning license* *Fabric Watch license* *Fabric license* Following two licenses only if 16-port switch: *Ports on Demand license - additional 4-port upgrade* *Ports on Demand license - additional 4-port upgrade* Following two licenses only if 32-port switch: *Ports on Demand license - additional 8-port upgrade* *Ports on Demand license - additional 8-port upgrade* *Extended Fabric license* (only if distance between nodes > 10km) If any of the above are not licensed, perform the following step.
Licenseadd "license key"	

SWITCH CONFIGURATION PROCEDURE

To configure a Brocade switch for a Fabric MetroCluster configuration, complete the following steps.

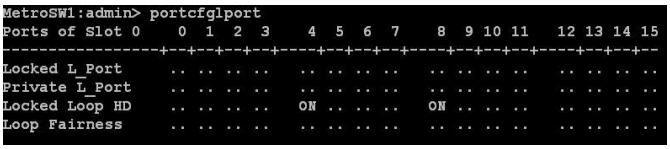
Note: This procedure must be performed on each switch in the MetroCluster configuration.

Command	Result
telnet "switch IPAddress"	Login prompt
Login admin	Prompt for password
"password"	Command line prompt
version	The currently installed switch firmware is displayed. Check the MetroCluster Fabric Switch download page on http://now.netapp.com for the currently supported version of the switch firmware. **Note**: To access this page, go to the Fibre Channel switch link on the Download Software page and select Brocade from the platform list.
	If your switch firmware is not the supported version, complete the steps outlined in section 7.2, "Appendix B: Switch Software Upgrade Procedure."
	Download the switch firmware from http://now.netapp.com and install it as described in the appropriate Brocade switch hardware reference manual.
reboot	Reboot the switch.
Switchdisable	Disable the switch.
cfgclear	Clear any preexisting configuration.
cfgdisable	
cfgsave	
configdefault	Configure the switch with default settings.
configure	Set the switch parameters.
	You should set only the following parameters.
Fabric parameters = **y**	
domain_id =	As a best practice, set the domain ID according to the switch number in the configuration. For example, at the Local site, switches 1 and 2 would have domain IDs SW1 and SW2 and switches 3 and 4 at the remote site would be SW3 and SW4, respectively.

Disable device probing = **1**

```
Virtual Channel parameters (yes, y, no, n): [no]
F-Port login parameters (yes, y, no, n): [no]
Zoning Operation parameters (yes, y, no, n): [no]
RSCN Transmission Mode (yes, y, no, n): [no]
Arbitrated Loop parameters (yes, y, no, n): [no] y
Send FAN frames?: (0..1) [1] 0
Enable CLOSE on OPEN received?: (0..4) [4]
```

`ctrl-d`	Exit the configuration utility.
`fastboot`	Reboot the switch so that the default settings take effect.
`Switchdisable`	Log in to the switch and disable it.

This step for Model 3000 switches (3200, 3800, 3250, 3850)	
`portcfglport <port#>,0,0,1`	Set all ports attached to disk loops to half duplex. **Note:** Perform this command only for the ports to which disks are attached. Refer to the switch-to-disk tables in “Cable Local Node” for the port numbers.
`portcfglporport`	ON is displayed in the Locked Loop HD fields only. 

```
MetroSW1:admin> portcfglport
Ports of Slot 0    0  1  2  3    4  5  6  7    8  9 10 11   12 13 14 15
-----------------+--+--+--+--+----+--+--+--+----+--+--+--+----+--+--+--
Locked L_Port      .. .. .. ..   .. .. .. ..   .. .. .. ..   .. .. .. ..
Private L_Port     .. .. .. ..   .. .. .. ..   .. .. .. ..   .. .. .. ..
Locked Loop HD     .. .. .. ..   ON .. .. ..   ON .. .. ..   .. .. .. ..
Loop Fairness      .. .. .. ..   .. .. .. ..   .. .. .. ..   .. .. .. ..
```

This step for Model 200E, 300, 5100, and 5000 only **Perform ONE of the following two:**	
If running Fabric Operating System version **earlier than 6.1:**	
`portcfglport <port#>,0,0,1`	Set all ports attached to disk loops to half duplex. **Note:** Perform this command only for the ports to which disks are attached. Refer to the switch-to-disk tables in “Cable Local Node” for the port numbers.
`portcfglport <port#>,1,0,1`	Set all ports attached to disk loops to Locked L-Port. **Note:** Perform this command only for the ports to which disks are attached. Refer to the switch-to-disk tables in “Cable Local Node” for the port numbers.
`portcfglport`	ON is displayed in the Locked Loop and HD fields.
If running Fabric Operating System **6.1 or later**	
`portfglport <port#>,1,0,2`	Set all ports attached to disk loops to Locked L-Port and Locked Loop Fairness by entering the command
`portcfglport`	ON is displayed in the Locked Loop and Locked Loop Fairness fields.

`portcfgtrunkport <ISL_port#> 0`	Configure the ISL port. (0 disables trunking.) Trunking must be disabled prior to configuring the long-distance ISL port.

portcfglongdistance <ISLport#>, "L0"	Configure the long-distance ISL port for an ISL length of up to 10km. For information about configuring an ISL port for an ISL length greater than 10km, refer to Appendix C of this document. For additional compatibility information, see the Brocade Fabric Aware Compatibility Matrix: http://www.brocade.com/products/interop_and_compatibility.jsp
portcfgISLMode <ISLport#>, mode	Enable or disable ISL R_RDY mode on a port. Toggling this mode may sometimes be necessary (high error rate on port) when connecting to WDM devices.
switchenable	Enable the switch.
switchname "Loc-SW1"	Set the switch name. Loc_SW1 is the name you gave to the current switch. Use a name that is meaningful, such as location (Loc for Local, Rem for Remote), and the switch number.
Portname <*number*>,"*name*"	**Best practice:** To assist in documenting the installation, names can be assigned to each port. Examples: **Portname 12,"ISL"** Using the `portshow` command, this identifies port 12 as being an ISL port. **Portname 0,"PRI-0a"** This identifies port 0 as being connected to port 0a on storage appliance PRI. **Portname 8,"SHLF1-A-in"** This identifies port 8 as being connected to A-IN on disk shelf 1.
configshow	Verify that the switch settings are correct.

OPTIONAL (REQUIRES FOS 6.0.0B OR LATER) – CREATE A TRAFFIC ISOLATION ZONE

Command	Result
Connect to the switch and log in as admin.	When you create a TI zone, by default failover is enabled and the zone is activated.
Zone –create –t *objectype* [-0 *optlist*] name –p "*portlist*"	Enter the zone –create command where: *Objtype* = The zone object type, which is ti for TI zones. *Optlist* = A list of options for creating the zone and controlling failover mode. A = Activate the zone after it is created d = Deactivate the zone after it is created n = Disable failover mode f = Enable failover mode *name* = The name of the zone to be created

	portlist = The list of the ports to be included in the TI zone. Ports are designated using the "D,I" (domain, index) format. Multiple ports are separated by a semicolon followed by a space.
cfgenable	Enable the appropriate zone configuration to make the change effective.
Zone Create –f ti FCVI –p "domain,port#;……."	EXAMPLE (Figure 11) Zone –create –f ti FCVI –p "1,0;1,4;3,4;3,0" This creates a TI zone with failover enabled. The TI zone called FCVI contains ISL E_ports 1,4 and 3,4 and FCVI N_ports 1,0 and 3,0. If you do not wish to have the FCVI zone immediately activated, use the "–d" switch, then run the zone –activate command
Zone -activate *FCVI*	Will activate the TI zone called *FCVI*
Zone –deactivate *FCVI*	Will deactivate the TI zone called *FCVI*

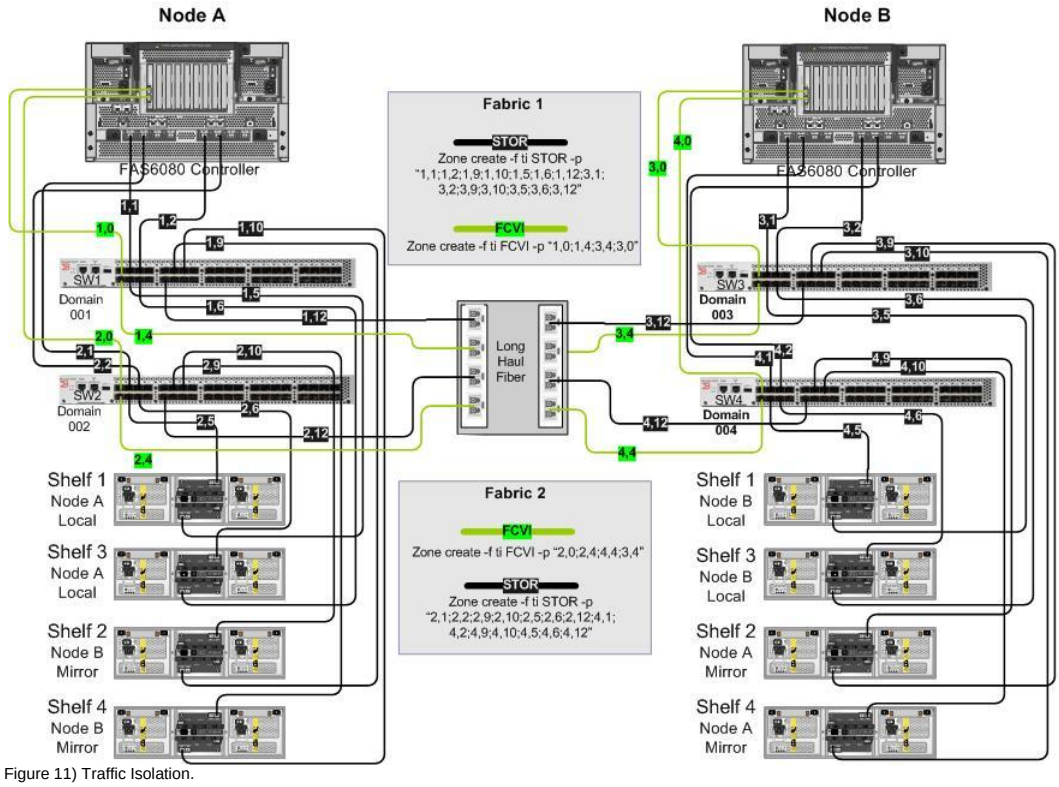


Figure 11) Traffic Isolation.

CABLE LOCAL NODE

Most of the currently shipping platforms use the software disk ownership model previously described.

Figure 12 shows the cabling for a MetroCluster system using the software-based ownership model. Refer to Appendix E for cable diagrams for systems using hardware-based ownership.

While there is greater cabling flexibility when using software disk ownership, it is important to be consistent and to use the virtual channel rules previously described.

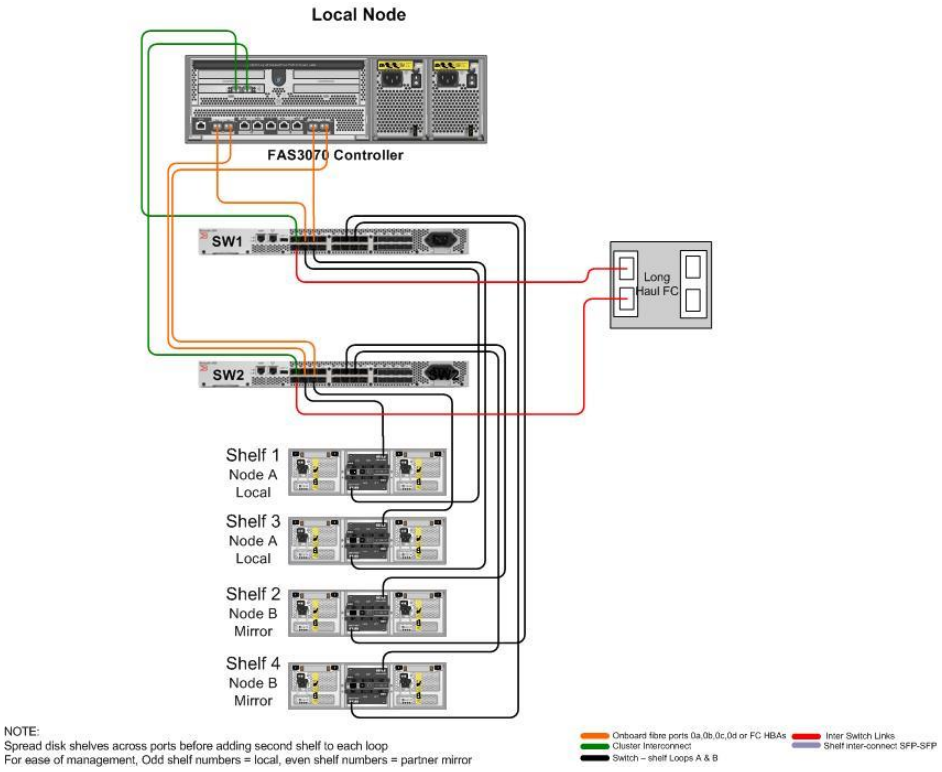


Figure 12) Local node cabling (software-based ownership).

SOFTWARE DISK OWNERSHIP

Software disk ownership is the method used for the FAS3040/3070, FAS31xx, and FAS6000 series MetroCluster systems to assign disk ownership. It allows greater cabling flexibility but requires greater thought in assigning disks. Two important items to remember are:

- Always assign all disks on the same loop to the same system and pool.
- Always assign all loops connected to the same adapter to the same pool.

Use the **`disk show -n`** command to view all disks that do not have assigned owners.

Use the following command to assign the disks that are labeled "Not Owned" to one of the system controllers.

Note: If you are assigning unowned disks to a nonlocal storage system, you must identify the storage system by using either the `-o ownername` or the `-s sysid` parameter or both.

```
disk assign {disk_name|all|-n count|auto} [-p pool] [-o ownername] [-s
sysid] [-c block|zoned] [-f]
```

`disk_name` specifies the disks that you want to assign to the system.

`All` specifies that all the unowned disks are assigned to the system.

`-n count` specifies the number of unassigned disks to be assigned to the system, as specified by count.

`Auto` causes disk autoassignment to be performed.

`-p pool` specifies which SyncMirror pool the disks are assigned to. The value of the pool is either 0 or 1.

Note: "Unassigned" disks are associated with a pool. To assign them to a different pool, use the `-f` option. However, moving individual disks between pools could result in the loss of redundancy and cause disk

autoassignment to be disabled for that loop. For this reason, you should move all disks on that loop to the other pool if possible.

`-o` *`ownername`* specifies the system that the disks are assigned to.

`-s` *`sysid`* specifies the system that the disks are assigned to.

`-c` specifies the checksum type (either block or zoned) for a LUN in V-Series systems.

`-f` must be specified if a system already owns the disk or if you want to assign a disk to a different pool. Enter **`man disk`** or refer to the Data ONTAP installation and administration guide for details.

VERIFY CONNECTIONS

Confirm that the disks are visible and have dual paths by entering the following command on the console:

```
storage show disk –p
```

The output shows the disks connected to the switches, the port to which they are connected, and the disk and module to which they belong, as shown in the following example:

```
ha16*> storage show disk -p
PRIMARY  PORT     SECONDARY PORT     SHELF   BAY
-------  ----     ---------  ----     -----   --
switch3:0.40 A    switch4:0.40 B     5       0
switch4:0.41 B    switch3:0.41 A     5       1
switch3:12.52 B   switch4:12.52 A    6       4
```

If redundant paths are not shown for each disk, recheck the cabling.

Note: On a MetroCluster system using hardware disk ownership, when performing recabling or expansions it is important to triple-check that the A-loop and B-loop of a shelf run to the very same port number on the two local switches. This is especially important when, for example, an older 16-port 3850 switch gets replaced by a 300. Older versions of Data ONTAP immediately panic and halt when the links to a shelf are not connected to the same port number. This affects both nodes of a Fabric MetroCluster system simultaneously, leading to a 100% outage for *both* nodes.

SET UP MIRRORS

To enable highly available access to data, all data on each node must be mirrored to the other node using SyncMirror. Although it is possible to have data on one node that is not mirrored on the other, NetApp does *not* recommend it. Keep in mind that mirrors can exist only between like drive types (FC to FC or ATA to ATA in the case of Stretch MetroCluster).

With the SyncMirror license installed, disks are divided into pools. When a mirror is created, Data ONTAP pulls disks from Pool0 for the Local data and from Pool1 for the mirror. The selection of which disks to use for the mirror can be left up to Data ONTAP or chosen specifically. For detailed information, see the *Data ONTAP Online Backup and Recovery Guide*.

It is important to verify the correct number of disks in each pool before creating the mirrored aggregate or traditional volume.

Any of these commands can be used to verify the number of drives in each pool:

```
Sysconfig –r –gives the broadest information
Aggr status -r
Vol status -r
```

Once the pools are verified, mirrors can be created by one of the following:

```
{ aggr | vol } create { aggrname | volname} –m ndisks[@disk-size]
```

For example, the command **`aggr create aggrA -m 6`** creates a mirrored aggregate called aggrA with 6 drives (3 for plex 0, 3 for plex 1).

Or an existing aggregate can be mirrored using the following:

```
{ aggr | vol } mirror { aggrname | volname }
```

In both of the above cases, Data ONTAP is allowed to choose the specific drives to be used. Optionally, the user can choose the drives. Further information can be found in the *Data ONTAP Online Backup and Recovery Guide.*

4.2 REMOTE SITE

VERIFY LICENSES

Follow the same procedures as performed at the Local site.

CONFIGURE FIBRE CHANNEL SWITCHES

Follow the same procedures as performed at the local site.

CABLE REMOTE CONTROLLER

Figure 13 shows the cabling for a MetroCluster system using the software-based ownership model. Refer to Appendix E for cable diagrams for systems using hardware-based ownership.

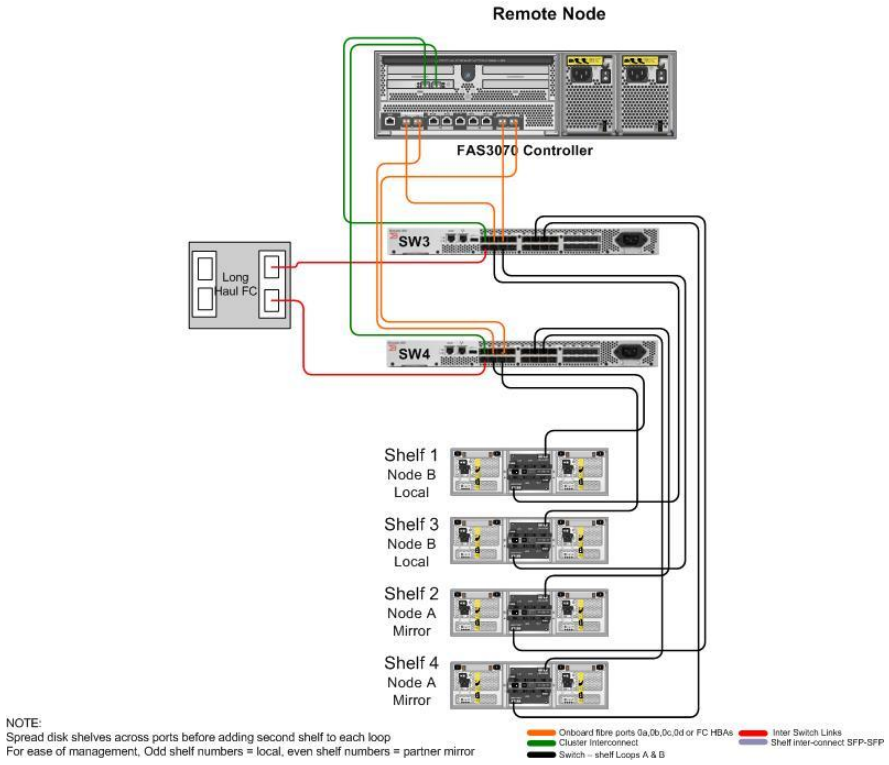


Figure 13) Remote node (software-based ownership).

SOFTWARE DISK OWNERSHIP

Follow the same procedures as performed at the local site.

VERIFY CONNECTIONS

Follow the same procedures as performed at the local site.

SET UP MIRRORS

Follow the same procedures as performed at the local site.

4.3 TESTING

VERIFY THAT EACH SITE CAN SEE THE PROPER DISKS

TEST FAILOVER

- Local to remote
- Remote to local

TEST GIVEBACK

- Local to remote
- Remote to local

4.4 ADDING MORE SHELVES

Additional disk shelves can be added to a MetroCluster configuration according to the procedures defined in the *Data ONTAP Active-Active Configuration Guide*. The following guidelines apply:

- ATA shelves are not supported on a Fabric MetroCluster.
- Make sure that additional shelves are added at the partner for mirroring.
- When adding shelves to new ports, the very same port on the other switch must be used. Otherwise, both nodes may not boot (with invalid disk configuration), or they may panic and halt at once, if this is performed online.
- Only like shelves can be mirrored; for example, FC to FC or ATA to ATA (Stretch MetroCluster only).
- Limit of two shelves per switch port on a Fabric MetroCluster.
- Total limit for a Fabric MetroCluster is the lesser of the platform limit or 504 disks. For a Stretch MetroCluster, the limit is that of the specific platform.
- When adding shelves to a Fabric MetroCluster, be sure to configure the associated switch ports to which the new shelves are connected.
- Keep in mind that each loop must have shelves of the same speed (2 or 4Gbps) and have the same type of controller (all LRC or all ESH/ESH2/ESH4).
- ESH4 disk shelves have a shorter maximum distance at 4Gbps.

As shown in Figures 14a and 14b, additional shelves can be added with a maximum of two shelves per loop (Fabric MetroCluster).

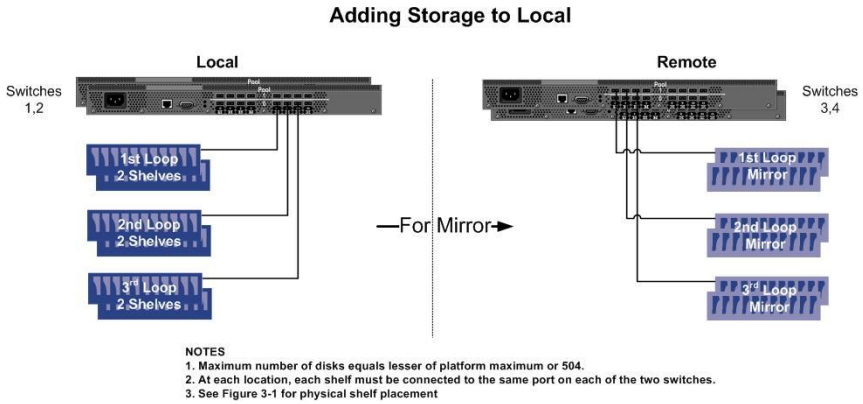


Figure 14a) Adding disk shelves to local node.

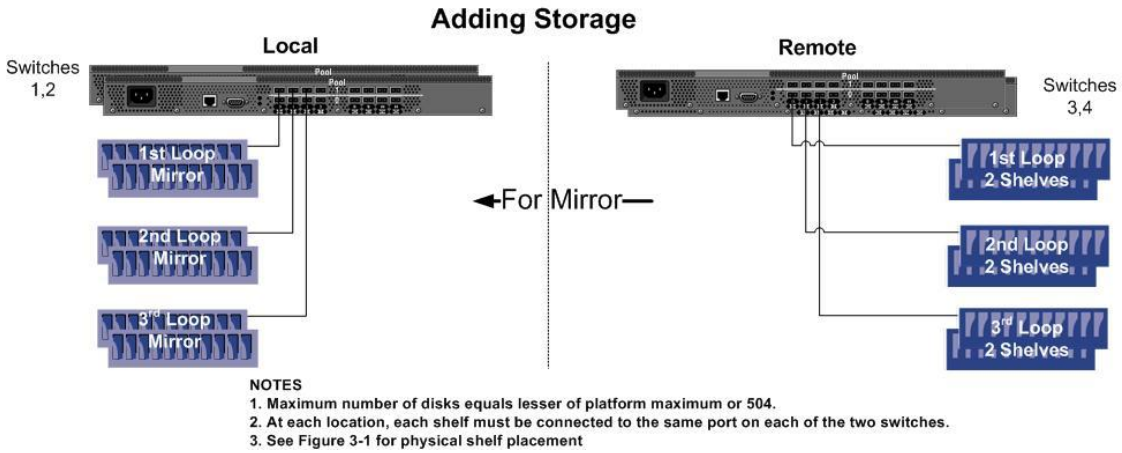


Figure 14b) Adding disk shelves to remote node.

4.5 BEST PRACTICES AND RECOMMENDATIONS

The following recommendations can enable a timely and successful installation.

- ✓ Complete installation planning workbook before attempting installation.
- ✓ When configuring the switches in a Fabric MetroCluster, configure all the ports to be used for disk loops at once. That way they won't have to be done as shelves are added.
- ✓ Use up the disk loop switch ports before doing any daisy-chaining.
- ✓ Make sure that, for each disk loop, the port number on the switch that the A loop is connected to is identical to the switch port number to which the B loop is connected.
- ✓ Make sure the A ports of the cluster interconnect card of both nodes are connected to the same fabric.
- ✓ Rather than set all shelves on disk loops to ID 1 or 2, set shelf IDs to indicate pool or location.
- ✓ Make sure ISL distance settings are correct for the actual distance.
- ✓ Mirror all storage from one site to the next.
- ✓ Make sure all ISL speeds are set correctly. Avoid having ISLs set to 2Gbps when the rest of the system is running at 4Gbps.

CABLING, CABLING, CABLING

Many people with SAN experience and knowledge of Fibre Channel switches assume that they can take advantage of the inherent flexibility that these switches typically provide. In the case of a Fabric MetroCluster configuration using hardware disk ownership, Data ONTAP expects specific devices on specific ports. Therefore it is important to follow the cabling diagram exactly. Full cabling diagrams are included in Appendices E and F.

CONFIGURATION OF STORAGE CONTROLLER FC PORTS

Depending on any prior use of the NetApp storage controllers, it may happen that the onboard Fibre Channel ports are not configured correctly. This affects the controller's ability to see all of the necessary disk shelves. To verify the configuration:

Verify that the onboard ports are configured correctly:

```
fcadmin config
Local
Adapter Type State Status
-------------------------------------------------------------
0a initiator UNDEFINED online
0b initiator UNDEFINED online
0c initiator UNDEFINED online
0d initiator CONFIGURED online
```

They should all show up as *initiator*. For any that do not, perform the following.

Ensure that the Fibre Channel ports are off line.

```
fcadmin config -d <adapter>
```

```
Note: Adapter refers to the port ID (for example, fcadmin config -d 0c
took port 0c offline). Also, more than one port can be specified.
```

Set the onboard port to target or initiator mode:

```
fcadmin config -t initiator <adapter>
```

Reboot the controller and check again:

```
fcadmin config
```

This verifies that the onboard Fibre Channel ports are on line.

CONFIGURATION OF SWITCHES

If the Brocade 200E switch is being used in the 8-port configuration, make sure that no SFPs are plugged into ports 8–15. Even though those ports may not be licensed or enabled, if Data ONTAP sees SFPs in those ports it treats the switch as a 16-port switch and expects to see devices connected accordingly.

5 SITE FAILOVER AND RECOVERY

5.1 SITE FAILOVER

There are several situations that could necessitate a site takeover. They include:

- Complete environmental failure (air conditioning, power, etc.)
- Geographic disaster (earthquake, fire, flood, etc.)

Upon determining that one of the sites has failed, the administrator must execute a specific command on the surviving node to initiate a site takeover. The command is:

```
cf forcetakeover -d
```

The takeover is not automatic because there may be cases in an active-active configuration in which the network between sites is down and each site is still fully functional. In this case a forced takeover might not be desirable.

It is important to remember that this is the case only when a complete site is lost. In the case of a failed controller at one of the sites, a normal cluster failover occurs. Due to the operation of SyncMirror, there is also added protection from multidrive or complete shelf failures.

When a storage controller fails in an active-active configuration, the partner detects the failure and automatically (if enabled) performs a takeover of the data-serving responsibilities from the failed controller. Part of this process relies on the surviving controller being able to read information from the disks on the failed controller. If this quorum of disks is not available, then automatic takeover won't be performed.

In a MetroCluster configuration, manually executing a single command will allow a takeover to occur in spite of the lack of a quorum of disks.

This "forced takeover" process breaks the mirrored relationships in order to bring the failed controller's volumes on line. This results in the following:

- Volumes have a new file system ID (FSID) in order to avoid conflict with the original volumes.
- LUNS (iSCSI or FCP) have a new serial number (in part derived from the FSID).
- Previous NFS mounts are stale and will need to be remounted.

LUNS are off line in order to make sure that only the desired LUNs are brought on line after the site failure.

Effective in Data ONTAP 7.2.4, there is an option to preserve the original FSID, which allows LUNs to retain their original serial number and the NFS mounts to be brought on line automatically. This option is called `cf.takeover.change_fsid`. If set to off (0) the original FSID will be preserved. This will be covered further in the Implementation section of this document.

SPLIT BRAIN SCENARIO

The `cf forcetakeover` command previously described allows the surviving site to take over the failed site's responsibilities without a quorum of disks available at the failed site (normally required). Once the problem at the failed site is resolved, the administrator must follow certain procedures, including restricting booting of the previously failed node. If access is not restricted, a split brain scenario may occur. This is the result of the controller at the failed site coming back up not knowing that there is a takeover situation. It begins servicing data requests while the remote site also continues to serve requests. The result is the possibility of data corruption. You can restrict access to the previously failed site controller in the following ways:

1) Turn off power to the previously failed node (disk shelves should be left on).
2) Disconnect the cluster interconnect and Fibre Channel adapter cables of the node at the surviving site.
    a) Use network management procedures to enable the storage systems at the disaster site to be isolated from the external public network.

b) Use any application-specified method that either prevents the application from restarting at the disaster site or prevents the application clients from accessing the application servers at the disaster site. Methods can include turning off the application server, removing an application server from the network, or any other method that prevents the application server from running applications.

RECOVERY PROCESS

Although a complete site failover can be performed with a single command, there are cases in which another step or two may be necessary before data is accessible at the surviving site.

If using a release of Data ONTAP earlier than 7.2.4 or if using 7.2.4 or later and the option setting `cf.takeover.change_fsid =1`:

- ***NFS volumes*** must be remounted. For more information about mounting volumes, see the File Access and Protocols Management Guide.
- ***iSCSI and Fibre Channel LUNS*** may need to be rescanned by the application if the application (i.e., VMware®) relies on the LUN serial number. When a new FSID is assigned, the LUN serial number changes.

iSCSI and Fibre Channel LUNS After a forced takeover, all LUNs that were being served from the failed site are served by the surviving site. However, each of these LUNs must be brought on line. For example:

```
lun online /vol/vol1/lun0 /vol/vol1/lun1
```

The reason they are off line is to avoid any LUN ID conflict. For example, suppose that two LUNs with the ID of 0 are mapped to the same igroup, but one of these LUNs was off line before the disaster. If the LUN that was previously off line came on line first, the second LUN would not be accessible because two LUNs with the same ID mapped to the same host cannot be brought on line.

Both of the steps described above could be automated with some clever scripting. The main challenge for the script is to understand the state of the LUNs before the site failure so that the script brings on line only LUNs that were on line prior to the site failure.

GIVEBACK PROCESS

After all the problems causing the site failure have been resolved and you have ensured that the controller at the failed site is off line, it is time to prepare for the giveback so the sites can return to their normal operation. The following procedure can be used to resynchronize the mirrors and perform the giveback.

COMMAND	RESULT
Turn on power to the disk shelves and FC switches at the disaster site.	
`aggr status -r`	Validate that you can access the remote storage. If remote shelves don't show up, check connectivity.
`partner`	Go into partner mode on the surviving node.
`aggr status –r`	Determine which aggregates are at the surviving site and which aggregates are at the disaster site by entering the following command: Aggregates at the disaster site show plexes that are in a failed state with an out-of-date status. Aggregates at the surviving site show plexes as on line.

If aggregates at the disaster site are on line, take them off line by entering the following command for each online aggregate: **aggr offline *disaster_aggr***	*disaster_aggr* is the name of the aggregate at the disaster site. **Note:** An error message appears if the aggregate is already off line.
Recreate the mirrored aggregates by entering the following command for each aggregate that was split: **aggr mirror aggr_name -v *disaster_aggr***	*aggr_name* is the aggregate on the surviving site's node. *disaster_aggr* is the aggregate on the disaster site's node. The *aggr_name* aggregate rejoins the *disaster_aggr* aggregate to reestablish the MetroCluster configuration. **Caution:** Make sure that resynchronization is complete on each aggregate before attempting the following step.
partner	Return to the command prompt of the remote node.
Enter the following command at the partner node: **cf giveback**	The node at the disaster site reboots.

6 RELEVANT DOCUMENTATION

6.1 METROCLUSTER DOCUMENTATION

The following documents can be found in the technical library:

- TR-3517: MetroCluster Upgrade Planning Guide
- TR-3660: An Automated MetroCluster Site Failover Solution
- TR-3606: High Availability and Disaster Recovery for VMware Using SnapMirror and MetroCluster
- TR-3552: Optical Network Installation Guide
- TR-3412: Example Proof of Concept for Microsoft Exchange Running with Fibre Channel Protocol on a NetApp MetroCluster

The following documents can be found on the NOW site:

- Active-Active Configuration Guide
- Data Protection Online Backup and Recovery Guide (Chapter 8, SyncMirror)
- MetroCluster Compatibility Matrix

BROCADE SWITCH DOCUMENTATION

http://now.netapp.com/NOW/knowledge/docs/brocade/relbroc30_40/

7 APPENDICES

7.1 APPENDIX A: SWITCH PORT ASSIGNMENTS (HARDWARE OWNERSHIP)

SWITCH 1 (local top)					SWITCH 2 (local bottom)				
Port	**Type**	**9xx**	**3020/ 3050**	**Pool**	**Port**	**Type**	**9xx**	**3020/ 3050**	**Pool**
0	FC Port or HBA	5a	0a	0	0	FC Port or HBA	5b	0b	0
1	FC Port or HBA	8a	0c	0	1	FC Port or HBA	8b	0d	0
2				0	2				0
3				0	3				0
4	Remote shelf B loop 1			1	4	Remote shelf A loop 1			1
5	Remote shelf B loop 1			1	5	Remote shelf A loop 1			1
6	Remote shelf B loop 1			1	6	Remote shelf A loop 1			1
7	Remote shelf B loop 1			1	7	Remote shelf A loop 1			1
8	Local shelf B loop 1			0	8	Local shelf A loop 1			0
9	Local shelf B loop 2			0	9	Local shelf A loop 2			0
10	Local shelf B loop 3			0	10	Local shelf A loop 3			0
11	Local shelf B loop 4			0	11	Local shelf A loop 4			0
12	Cluster Interconnect	10a	1a	1	12	Cluster Interconnect			1
13	ISL	To Switch 3			13	ISL	To Switch 4		
14				1	14				1
15				1	15				1
SWITCH 3					**SWITCH 4**				
Port	**Type**	**9xx**	**3020/ 3050**	**Pool**	**Port**	**Type**	**9xx**	**3020/ 3050**	**Pool**
0	FC Port or HBA	5a	0a	0	0	FC Port or HBA	5b	0b	0
1	FC Port or HBA	8a	0c	0	1	FC Port or HBA	8b	0d	0
2				0	2				0
3				0	3				0
4	Remote shelf B loop 1			1	4	Remote shelf A loop 1			1
5	Remote shelf B loop 1			1	5	Remote shelf A loop 1			1
6	Remote shelf B loop 1			1	6	Remote shelf A loop 1			1
7	Remote shelf B loop 1			1	7	Remote shelf A loop 1			1
8	Local shelf B loop 1			0	8	Local shelf A loop 1			0
9	Local shelf B loop 2			0	9	Local shelf A loop 2			0
10	Local shelf B loop 3			0	10	Local shelf A loop 3			0
11	Local shelf B loop 4			0	11	Local shelf A loop 4			0
12	Cluster Interconnect	10a	1a	1	12	Cluster Interconnect			1
13	ISL	From Switch 1			13	ISL	From Switch 2		
14				1	14				
15				1	15				

7.2 APPENDIX B: SWITCH SOFTWARE UPGRADE PROCEDURE

In order to achieve an optimum MetroCluster environment, it may occasionally be necessary to upgrade the firmware on the Fibre Channel switches. To determine the proper revision level, consult the *MetroCluster Compatibility Matrix* on the NOW site.

The instructions in this section can also be found at

http://now.netapp.com/NOW/download/software/sanswitch/fcp/Brocade/mc_ontap641_fabric_200_download.shtml.

NOTE: When upgrading firmware make sure all paths are available first.

To download the Brocade 5.1.0 firmware to a workstation and then install it on a Brocade 200E switch, complete the following steps:

Ensure that the workstation recognizes the switch on the network (using the ping utility, for example).

From the browser on the workstation, right-click this installation package (240.5MB) link.

Specify the directory to which you want to download the firmware installation package.

After downloading is complete, access the switch from the workstation using a telnet session.

Install the firmware on the switch by issuing the firmware download command. For the syntax and a description of the firmware download command, see the Brocade Fabric OS Reference Guide.

Reboot the switch to activate the new firmware.

7.3 APPENDIX C: FIBRE CHANNEL SWITCH ISL DISTANCE SETTINGS

Level	Distance @ 1Gbps	Distance @ 2Gbps	Distance @ 4 Gbps	Earliest FOS Release	License Required
L0	10km	5km	2km	All	No
LE	10km	10km	10km	3.x, 4.x	No
L0.5	25km	25km	25km	3.1.0, 4.1.0, 4.x, 5.x	Yes
L1	50km	50km	50km	All	Yes
L2	100km	100km	100km	All	Yes
LD*	Auto to 500 km	Auto to 250 km	Auto to 100km	3.1.0, 4.1.0, 4.4.0, 5.x	Yes
LS**	500km	250km	100km	5.1.0	Yes
* LD ... Fabric calculates distance.					
** LS ... Administrator specifies distance.					

7.4 APPENDIX D: FABRIC METROCLUSTER WORKSHEET

The Fabric MetroCluster worksheet is a planning aid on which you can record specific cabling information about your Fabric MetroCluster system. Use this information during configuration procedures.

CUSTOMER INFORMATION:

Customer Name			
Address			
City		**State/Country**	
Main Phone			
Local Contact			
Primary Phone Number			
Requested Install Date			

Site A Installation Location (if different from above)

Address			
City		**State/Country**	
Main Phone			
Primary Contact			
Primary Phone Number			

Site B Installation Location (if different from above)

Address			
City		**State/Country**	
Main Phone			
Primary Contact			
Primary Phone Number			

GENERAL

As with most projects, the key to a successful deployment of a NetApp MetroCluster system is planning, preparation, and documentation. In order to facilitate these activities, certain data must be gathered early in the process.

PHYSICAL

1. What is the total distance between the two locations?

2. Please describe the fiber infrastructure between locations:

    a. Fiber type

    b. Number of unused pairs

    c. Number of patch panels

3. Is this infrastructure already in place?

4. If not, when will it be complete?

5. Is this an upgrade to a MetroCluster system?

6. Please outline any data to be migrated to/from either of the nodes.

7. Is there or will there be a mix of 2Gbps and 4Gbps disk shelves?

8. If 4Gbps operation is desired, have the appropriate HBAs been ordered? (Most onboard ports are only capable of 2Gbps operation.)

9. What type of disks (FC, ATA, etc.) currently exist (if this is an upgrade) or will exist? Note: ATA is not supported on a Fabric MetroCluster configuration.

INSTALL CHECKLIST:

DESCRIPTION	Yes or No
Appropriate power is installed/ready for Site A	
Appropriate power is installed/ready for Site B	
Distance between FAS at Site A and Switch 1 – Please list measured distance ***(This distance will determine minimum cable lengths needed)*	
Distance between shelves at Site A and Switch 1 – Please list measured distance ***(This distance will determine minimum cable lengths needed)*	
Distance between FAS at Site A and Switch 2 – Please list measured distance ***(This distance will determine minimum cable lengths needed)*	
Distance between shelves at Site A and Switch 2 – Please list measured distance ***(This distance will determine minimum cable lengths needed)*	
Distance between FAS at Site B and Switch 3 – Please list measured distance ***(This distance will determine minimum cable lengths needed)*	
Distance between shelves at Site B and Switch 3 – Please list measured distance ***(This distance will determine minimum cable lengths needed)*	
Distance between FAS at Site B and Switch 4 – Please list measured distance ***(This distance will determine minimum cable lengths needed below)*	
Distance between shelves at Site B and Switch 4 – Please list measured distance ***(This distance will determine minimum cable lengths needed)*	
Distance between Switch 1 and Switch 3 – Please list measured distance	
Distance between Switch 2 and Switch 4 – Please list measured distance	
Determine connection type at your Demarc. ****Cables used between switches and Demarc are the responsibility of the customer****	
Serial console (laptops) to be used to connect to the switches and appliances for installation, configuration, and testing. Laptops should have HyperTerm or TeraTerm applications installed. A minimum of 2 serial consoles is needed, 1 per site for the duration of the project. Requesting 2 per site if possible. Please advise the number of serial consoles available.	
Complete NetApp System Configuration Worksheet prior to project start date	
Appropriate number of Ethernet network connections are available/ready	
Both systems need access to ALL the same networks/subnets/SAN for failover capability	
Complete data layout design to include naming conventions for aggregates, volumes, and qtrees	
UNIX system administrator(s) with domain privileges dedicated for this project	
UNIX system administrator(s) with domain privileges dedicated for this project for any work beyond the normal 8-hour workday	
Windows system administrator(s) with domain privileges dedicated for this project	
Windows system administrator(s) with domain privileges dedicated for this project for any work beyond the normal 8-hour workday	
Network administrator(s) with domain privileges dedicated for this project	
Network administrator(s) with domain privileges dedicated for this project for any work beyond the normal 8-hour workday	
Total number of cables:	
****NOTE: Any cable distances that are greater than the NetApp- supplied cable lengths are the responsibility of the customer.***	

STORAGE CONTROLLER INFORMATION:

	Local	Remote
Model		
Data ONTAP Version		
IP Address		
Host Name		
Licenses		

BROCADE SWITCH INFORMATION:

Fibre Channel Switch Inventory				
	Switch 1	Switch 2	Switch 3	Switch 4
Domain ID				
Switch Name				
Switch Model				
Firmware Version				
Licenses				
No. Ports				
Physical Location				
Hostname				
IP Address				

BROCADE SWITCH-ATTACHED DEVICES:

Port	Switch 1	Switch 2	Switch 3	Switch 4
0				
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				

DISK SHELF LAYOUT:

Local Site					
	Shelf ID	Owner System	Pool	Firmware Version	Physical Location
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					

9

Remote Site					
	Shelf ID	Owner System	Pool	Firmware Version	Physical Location
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					

7.5 APPENDIX E: FABRIC METROCLUSTER (HARDWARE OWNERSHIP)

Because of the architecture of MetroCluster using a hardware ownership model, it is important to install cabling exactly as described. Although a Fibre Channel switch is quite flexible, Data ONTAP has specific requirements. Most installation problems are caused by cabling problems or misconfiguration of switches.

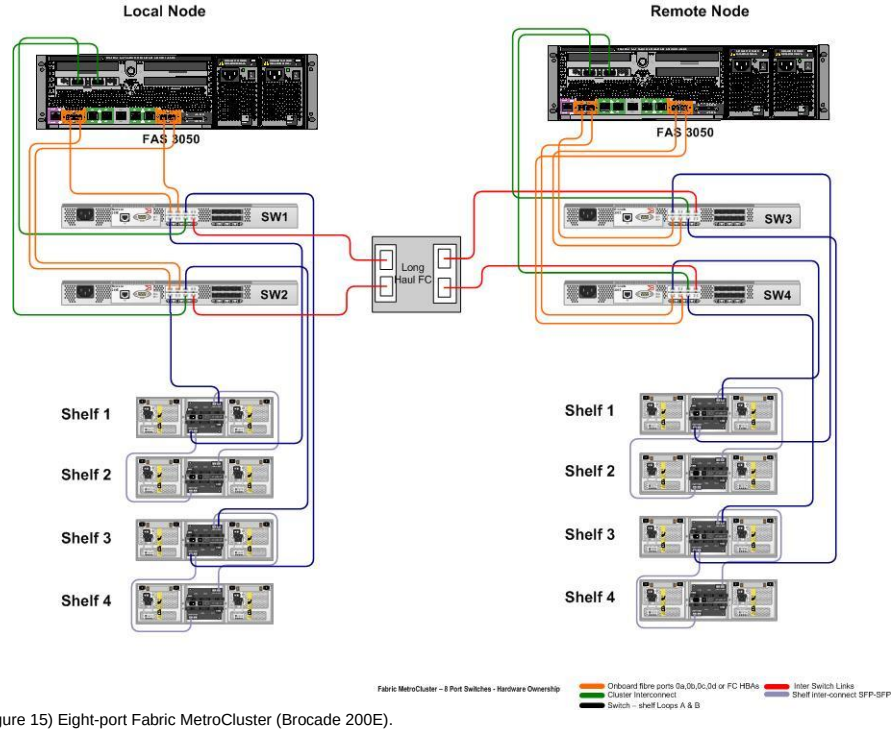


Figure 15) Eight-port Fabric MetroCluster (Brocade 200E).

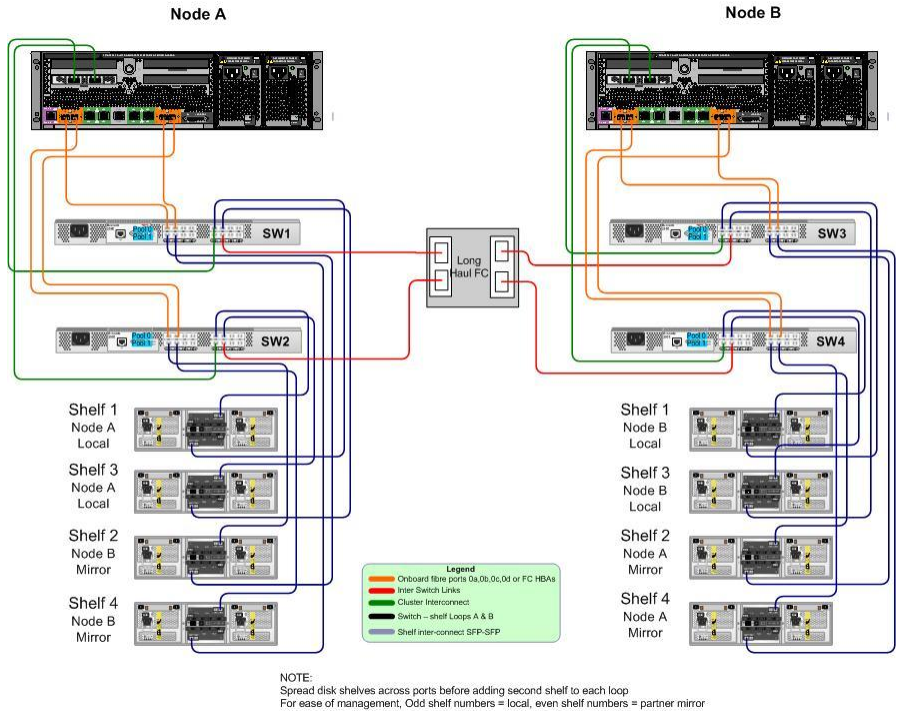


Figure 16) Sixteen-port Fabric MetroCluster (Brocade 200E).

If the installation includes more than two shelves at each site, refer to Section 4.4, "Adding More Shelves."

7.6 APPENDIX F: FABRIC METROCLUSTER (40-PORT SWITCHES)

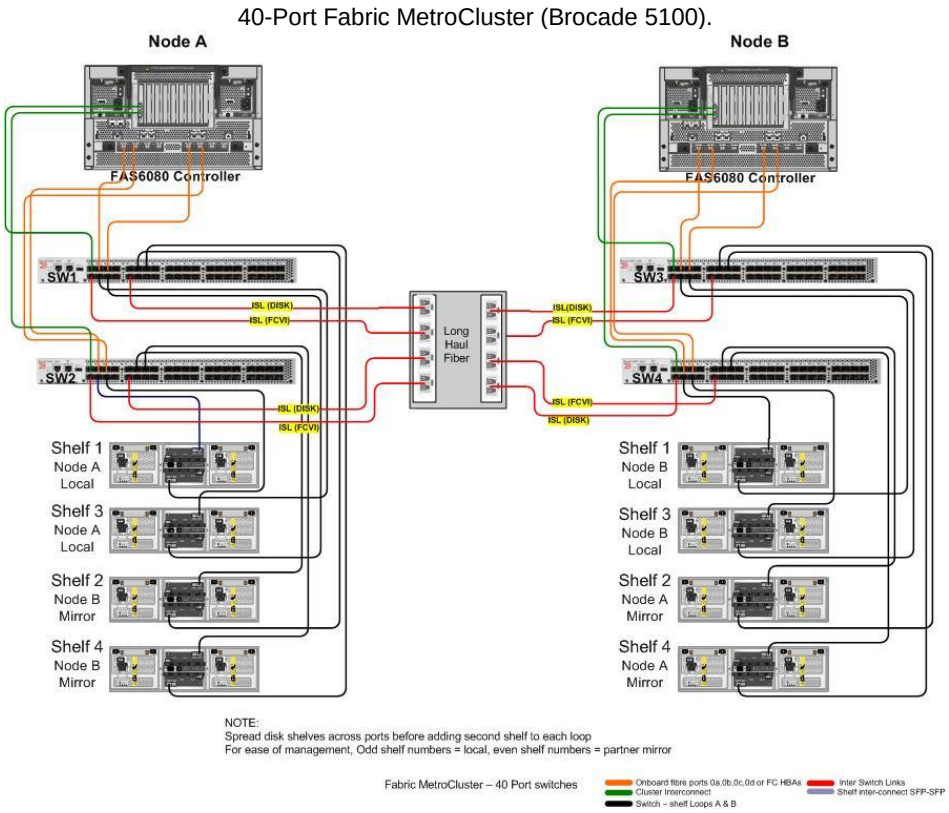


Figure 17) Forty-port Fabric MetroCluster (Brocade 5100).

7.7 APPENDIX G: 31XX METROCLUSTER

STRETCH METROCLUSTER

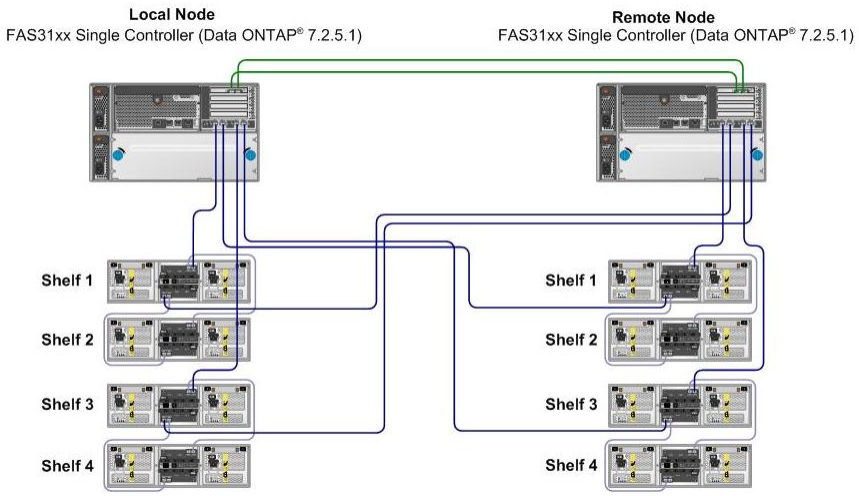


Figure 18) Stretch MetroCluster

FABRIC METROCLUSTER

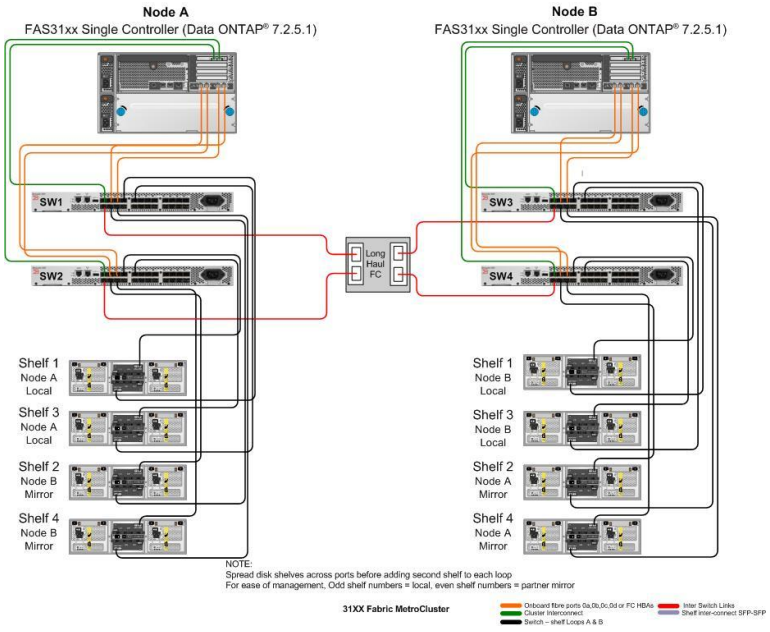


Figure19) 24-port Fabric MetroCluster (Brocade 300).