Technical Report

MetroCluster Version 8.2.1 Best Practices for Implementation

Charlotte Brooks, NetApp
May 2014 | TR-3548

Abstract

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

TABLE OF CONTENTS

1 Introduction 6

1.1 Intended Audience 6

1.2 Scope 6

1.3 Prerequisites and Assumptions 6

2 MetroCluster Overview 6

2.1 Components Required for MetroCluster 9

Stretch MetroCluster 9

Fabric MetroCluster 10

2.2 MetroCluster Dual-Chassis and Twin Configurations 12

2.3 MetroCluster with Nonmirrored Aggregates 13

3 MetroCluster Failure Handling 14

3.1 Controller Failure 14

3.2 Shelf Failure 15

3.3 Switch Failure 16

3.4 Complete Site Failure 17

3.5 Link Loss Between Sites 18

3.6 Rolling Failures 19

4 SAS-Based MetroCluster Configurations 20

4.1 FibreBridge Details 20

4.2 Stretch MetroCluster with FibreBridge 21

4.3 Fabric MetroCluster with FibreBridge 22

4.4 FibreBridge Setup and Configuration 23

4.5 Shelf Mixing Rules for Stretch MetroCluster 23

4.6 Stretch MetroCluster with SAS Optical Cables 23

5 Stretch MetroCluster Considerations 26

5.1 Stretch MetroCluster and MPHA 26

5.2 Sufficient Initiator Ports or Initiator HBAs 26

5.3 Stretch MetroCluster Mixed Disk Pool Configurations 26

5.4 Stretch MetroCluster with FAS3210 and FAS8020 26

5.5 Stretch MetroCluster Spindle Limits 26

5.6 Stretch MetroCluster Distances 26

6 Fabric MetroCluster Hardware and Software Components 27

6.1 Components 27

 © 2014 NetApp, Inc. All Rights Reserved.

6.2 Choice of Fabric Switch 28

7 Fabric MetroCluster Using Brocade Switches 28

7.1 Brocade: Traffic Isolation Zones 28

7.2 Brocade: Mixing Switch Models 30

8 Fabric MetroCluster Using Cisco Switches 30

8.1 Cisco: Port Groups....30

8.2 Cisco 9710 31

8.3 Cisco: VSANs 32

8.4 Cisco: Cabinet Integration....32

8.5 Cisco: Mixing Switch Models....32

9 Fabric MetroCluster Considerations 32

9.1 Fabric MetroCluster Traffic Flow Considerations 32

9.2 Fabric MetroCluster Distance Considerations....33

9.3 Fabric MetroCluster ISL Considerations 33

9.4 Fabric MetroCluster Latency Considerations 33

9.5 Fabric MetroCluster Using DS14FC Shelves....34

9.6 Fabric MetroCluster Using SAS Shelves 34

9.7 SAS, SATA, and SSD Mixed Shelf Configuration Considerations 36

9.8 Fabric MetroCluster Using Shared Switches....36

9.9 Fabric MetroCluster Using DS14FC and SAS Shelves (Mixed Shelf Configurations)....36

10 MetroCluster Sizing and Performance 37

10.1 Round-Trip Time Between Sites 37

10.2 Aggregate Mirroring (SyncMirror) Performance 37

10.3 Takeover and Giveback Times 37

10.4 FibreBridge 6500N Performance 38

10.5 Flash Pool Performance 40

10.6 Sizing MetroCluster Interswitch Links (ISLs)....41

10.7 FC-VI Speed 43

10.8 Unidirectional Port Mirroring....43

10.9 ISL Encryption....43

11 Configuring MetroCluster 43

11.1 Stretch MetroCluster Configuration....44

11.2 Fabric MetroCluster Configuration 45

12 Site Failover and Recovery.... 48

 © 2014 NetApp, Inc. All Rights Reserved.

12.1 Site Failover 48
12.2 Split-Brain Scenario 49
12.3 Recovery Process 49
12.4 Giveback Process 50

13 Monitoring MetroCluster and MetroCluster Tools 51

13.1 Fabric MetroCluster Data Collector (FMC_DC) 51
13.2 MetroCluster Tie-Breaker 51
13.3 OnCommand Site Recovery 52
13.4 MetroCluster vCenter Easy Button 52

14 Nondisruptive Operation 53

14.1 Nondisruptive Shelf Replacement (NDSR) for Fabric MetroCluster 53
14.2 Nondisruptive Hardware Changes 56
14.3 Nondisruptive Operation for SAS Shelves 56

15 MetroCluster Conversions 56

15.1 HA Pair to Stretch MetroCluster 56
15.2 HA Pair to Fabric MetroCluster 58
15.3 Stretch MetroCluster to Fabric MetroCluster 59
15.4 Controller-Only Upgrades 61
15.5 MetroCluster to HA Pair 62
15.6 Relocating a MetroCluster Configuration 62

16 MetroCluster Interoperability 62

16.1 Fabric MetroCluster and the Front End 63
16.2 Fabric MetroCluster and FCP Clients 63
16.3 V-Series/FlexArray Virtualization MetroCluster 63
16.4 MetroCluster and SnapMirror 63
16.5 MetroCluster and Storage Efficiency 63

17 Solutions on MetroCluster 64

Appendix 65

Cabling Diagrams 65

 © 2014 NetApp, Inc. All Rights Reserved.

LIST OF FIGURES

Figure 1) NetApp HA pair failover. 7
Figure 2) NetApp stretch MetroCluster. Point-to-point connections between MetroCluster halves using long cables. 8
Figure 3) NetApp fabric MetroCluster. Four fabric switches are used to implement two independent fabrics (for fabric redundancy). 8
Figure 4) NetApp stretch MetroCluster detailed graphic. FAS32xx single controller on each site with DS14 shelves. ..10
Figure 5) NetApp fabric MetroCluster detailed graphic. FAS62xx single controller on each site with FibreBridges and Brocade 6510 switches. 12
Figure 6) Controller failure: automatic failover to surviving controller. 14
Figure 7) Shelf (or plex or aggregate) failure: automatic and seamless failover to mirrored plex. 15
Figure 8) Switch failure: automatic failover to redundant fabric. 16
Figure 9) Complete site failure: Enter CFOD command on surviving controller. 17
Figure 10) Link loss between sites: MetroCluster takes no action except to suspend mirroring. 18
Figure 11) Example of rolling failures. 19
Figure 12) FibreBridge 6500N front. 20
Figure 13) FibreBridge 6500N back. 20
Figure 14) SAS-based stretch MetroCluster using FibreBridge. 22
Figure 15) SAS-based fabric MetroCluster using FibreBridge. 23
Figure 16) SAS-based stretch MetroCluster using multimode SAS optical. 24
Figure 17) SAS-based stretch MetroCluster using single-mode SAS optical. 25
Figure 18) Brocade TI zones. 29
Figure 19) Port chunking with Cisco 9710 switch. 32
Figure 20) FibreBridge IOPS and throughput under various read/write workloads. 38
Figure 21) Read latency: FibreBridge compared to direct attached. 39
Figure 22) Write latency: FibreBridge compared to direct attached. 39
Figure 23) Read latency: FibreBridge compared to direct attached. 40
Figure 24) Write latency: FibreBridge compared to direct attached. 40
Figure 25) FAS32xx stretch MetroCluster using DS14 shelves. 65
Figure 26) FAS32xx stretch MetroCluster using SAS shelves with FibreBridge. 66
Figure 27) FAS32xx stretch MetroCluster using DS14 and SAS shelves (with FibreBridge). 67
Figure 28) FAS62xx fabric MetroCluster using DS14FC shelves. 68
Figure 29) FAS62xx fabric MetroCluster using SAS shelves with FibreBridge. 69
Figure 30) FAS62xx fabric MetroCluster using DS14FC and SAS shelves (with FibreBridge). 70
Figure 31) FAS62XX fabric MetroCluster with Cisco 9148. 71
Figure 32) FAS62XX fabric MetroCluster with Cisco 9222i. 72
Figure 33) FAS62XX fabric MetroCluster with Cisco 9222i plus 4/44 FC module with 4Gbps ISLs. 73
Figure 34) FAS62XX fabric MetroCluster with Cisco 9222i plus 4/44 FC module with 8Gbps ISLs. 74

 © 2014 NetApp, Inc. All Rights Reserved.

1 Introduction

1.1 Intended Audience

The information in this document is for field personnel and administrators who are responsible for architecting and deploying MetroCluster continuous availability and disaster recovery solutions.

1.2 Scope

This document covers the following 7-Mode MetroCluster configurations for Data ONTAP 8.2.x releases:

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

Topics that apply only to stretch 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 of it, V-Series/FlexArray virtualization MetroCluster specifics are not covered in this document. However, any references to FAS controllers apply to the equivalent V-Series/FlexArray virtualization models unless otherwise specified.

1.3 Prerequisites 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.
- The reader has a full understanding of HA pair controller 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.
- This technical report is being written in the context of Data ONTAP operating in 7-Mode. MetroCluster is not supported in clustered Data ONTAP at the time of writing.

2 MetroCluster Overview

MetroCluster extends high availability, providing additional layers of protection.

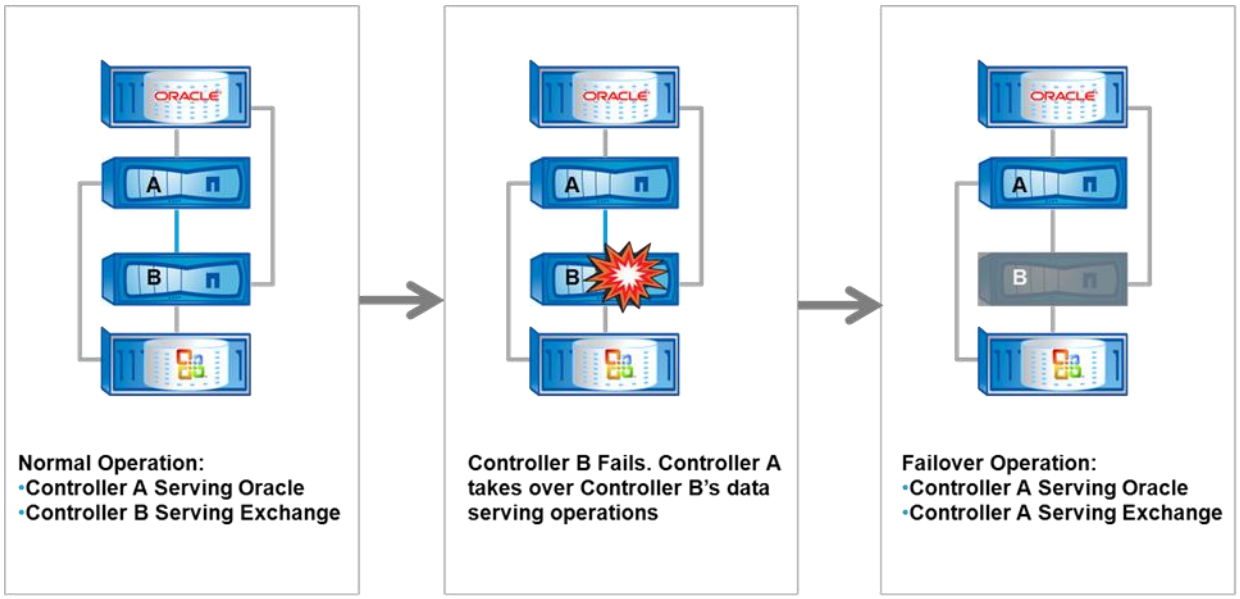
NetApp highly available (HA) pairs couple two controllers protecting against single controller failures. NetApp disk shelves have built-in physical and software redundancies such as dual power supplies and RAID-DP® (double parity) technology. NetApp HA pairs and shelves protect against many data center failures but cannot guarantee extreme levels of availability.

MetroCluster layers additional protection onto existing NetApp HA pairs. MetroCluster enables extreme levels of availability. Synchronous data mirroring enables zero data loss, and automatic failover enables nearly 100% uptime. Thus, MetroCluster enables a zero recovery point objective and a near-zero recovery time objective.

NetApp HA leverages takeover functionality, otherwise known as cluster failover (CFO), to protect against controller failures. On the failure of a NetApp controller, the surviving controller takes over the failed controller's data-serving operations while continuing its own data-serving operations (Figure 1). Controllers in a NetApp HA pair use the cluster interconnect to monitor partner health and to mirror NVLOG information composed of recent writes not propagated to disk.

 © 2014 NetApp, Inc. All Rights Reserved.

Figure 1) NetApp HA pair failover.



MetroCluster leverages NetApp HA CFO functionality to automatically protect against controller failures. Additionally, MetroCluster layers local SyncMirror® technology, cluster failover on disaster (CFOD), hardware redundancy, and geographical separation to achieve extreme levels of availability.

SyncMirror synchronously mirrors data across the two halves of the MetroCluster configuration by writing data to two plexes: the local plex (on the local shelf) actively serving data and the remote plex (on the remote shelf) normally not serving data. On local shelf failure, the remote shelf seamlessly takes over data-serving operations. No data loss occurs because of synchronous mirroring.

CFOD, as distinct from CFO, protects against complete site disasters by:

- Initiating a controller failover to the surviving controller
- Serving the failed controller's data by activating the data mirror
- Continuing to serve its own data

Hardware is redundant for all MetroCluster components. Controllers, storage, cables, switches (fabric MetroCluster), bridges, and adapters are all redundant.

Geographical separation is implemented by physically separating controllers and storage, creating two MetroCluster halves. For distances under 500m (campus distances), long cables are used to create stretch MetroCluster configurations (Figure 2). For distances over 500m but under 200km/~125 miles (metro distances), a fabric is implemented across the two geographies, creating a fabric MetroCluster configuration (Figure 3).

 © 2014 NetApp, Inc. All Rights Reserved.

Figure 2) Stretch MetroCluster. Point-to-point connections using long cables.

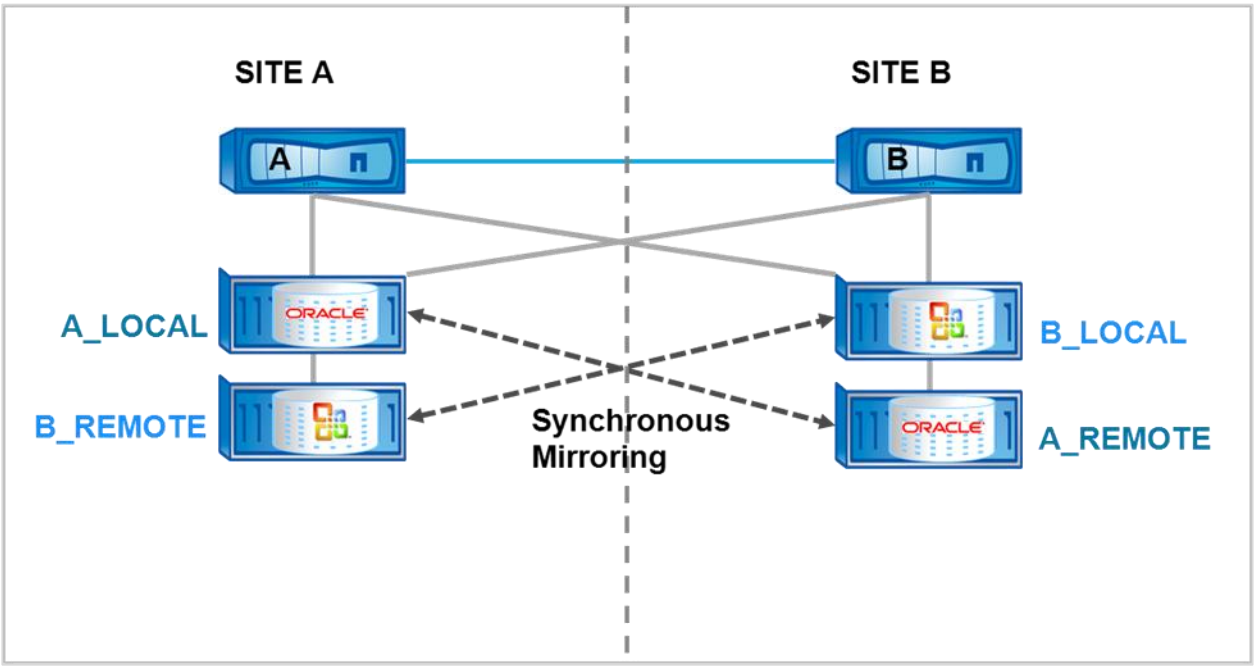
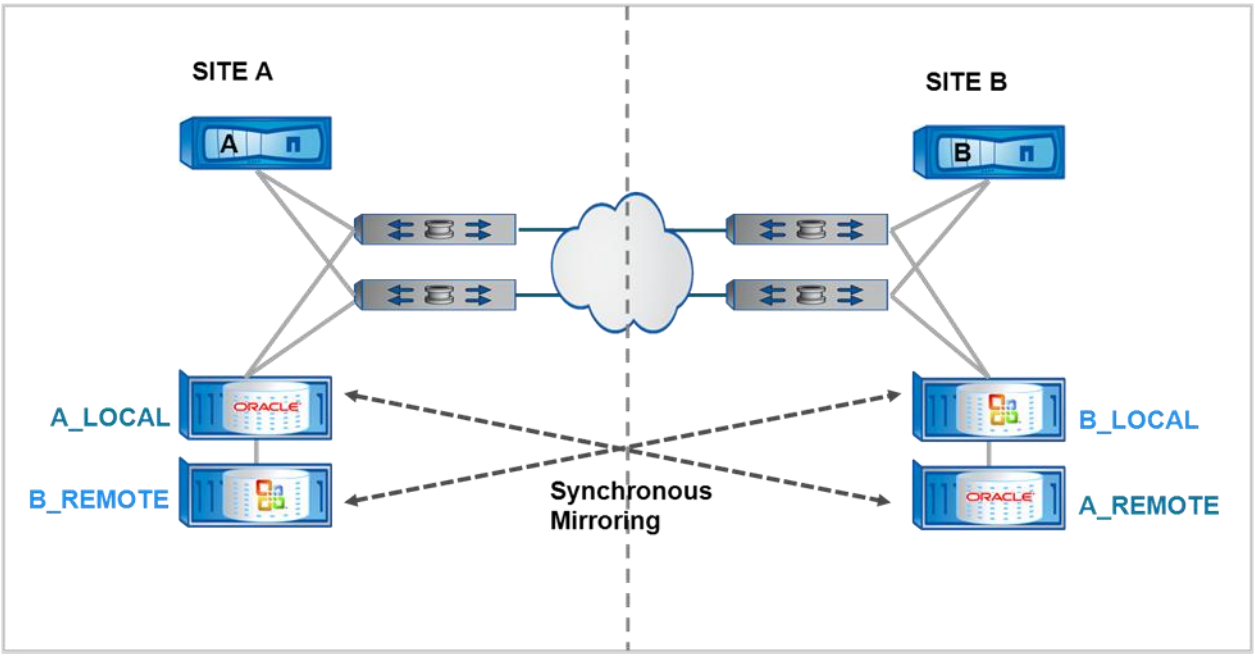


Figure 3) Fabric MetroCluster. Four fabric switches used in two independent fabrics.



 © 2014 NetApp, Inc. All Rights Reserved.

2.1 Components Required for MetroCluster

Stretch MetroCluster

A NetApp stretch MetroCluster configuration includes the following components. A sample configuration is shown in Figure 4.

Hardware

- Standard HA pair of controllers running a compatible version of Data ONTAP (see the Interoperability Matrix on the NetApp Support site for supported models and specific Data ONTAP releases).
- FC-VI cluster adapter (one per controller) is required for the cluster interconnect on FAS31xx, 32xx, 62xx, and 80x0 controllers. FC-VI adapters are used and required only with MetroCluster ; standard HA pairs do not use these adapters. The NVRAM InfiniBand interconnect is supported for the cluster interconnect only on earlier models than those mentioned.
- Sufficient initiator ports (storage adapters).
- Extra disk shelves to accommodate the mirrored data.
- SAS copper cables, SAS optical cables, or FibreBridges: two per stack of SAS shelves.

Data ONTAP Licenses

For versions of Data ONTAP earlier than 8.2, enable the following licenses (included as part of the base software package starting in Data ONTAP 7.3):

- cluster_remote license for the site failover on disaster functionality
- syncmirror_local license for synchronous mirroring across sites
- cluster license for controller failover functionality

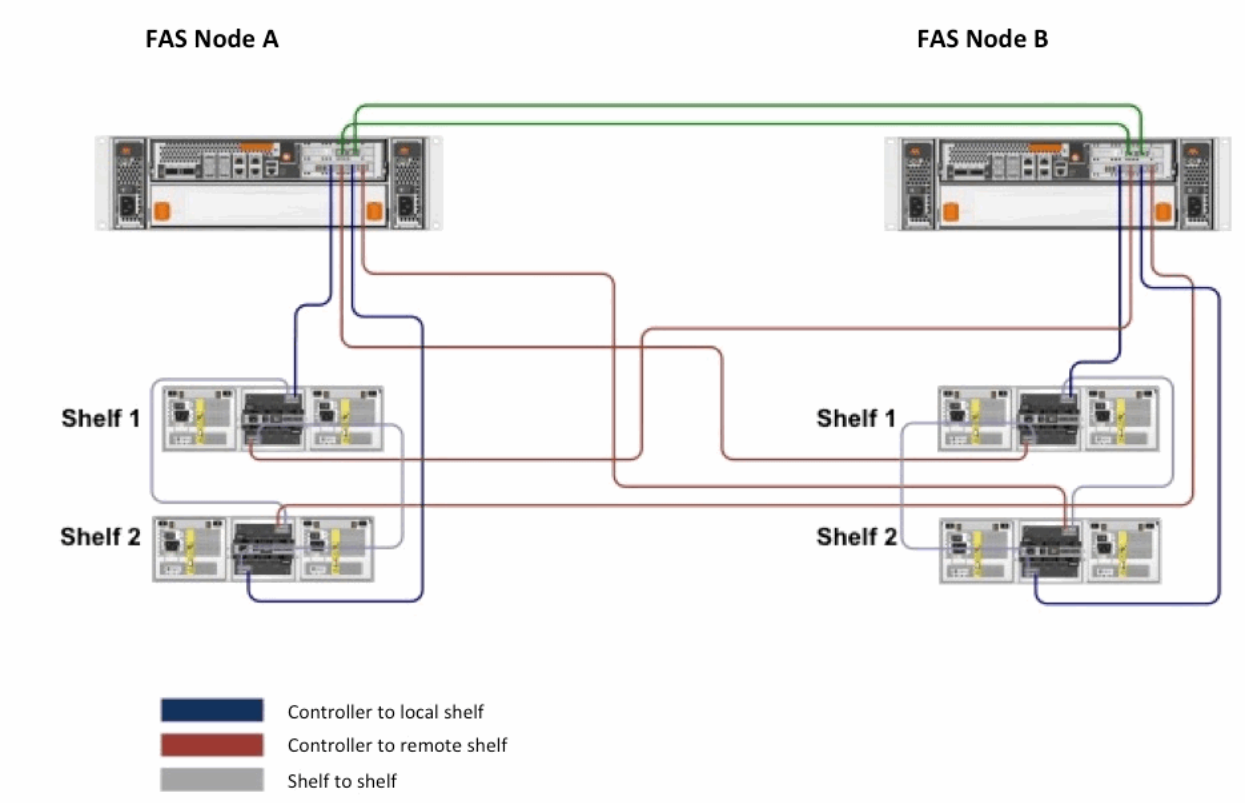
Starting in Data ONTAP 8.2, required licenses are replaced with options. Features specific to MetroCluster are enabled during configuration; for configuration steps, see the High-Availability and MetroCluster Configuration Guide.

The following Data ONTAP options must be enabled on both nodes:

- cf.mode: You must set this option to ha.
- cf.remote_syncmirror.enable: Set the option to on.

 © 2014 NetApp, Inc. All Rights Reserved.

Figure 4) Stretch MetroCluster, FAS32xx single controller with DS14 shelves.



Fabric MetroCluster

A NetApp fabric MetroCluster configuration includes the following components. A sample configuration is shown in Figure 5.

Hardware

- An HA pair of controllers running a compatible version of Data ONTAP (see the Interoperability Matrix on the NetApp Support site for supported models).
- FC-VI cluster adapter (one per controller) for the cluster interconnect.
- Extra disk shelves to accommodate the mirrored data.
- Fibre Channel initiator ports: four per controller.
- If using SAS shelves, FibreBridges: two per stack of SAS shelves.
- Four Fibre Channel switches from the same vendor (Brocade and Cisco® fabric switches are supported) with supported firmware: a pair of switches is installed at each location (or switch with two line cards in each location for the Cisco 9710). See the Interoperability Matrix on the NetApp Support site for supported models. The switches in a fabric **must** be the same model. It is supported, though not recommended, to use different switches (from the same vendor) in the two fabrics. The switches must be dedicated to MetroCluster; sharing with components other than MetroCluster is not permitted. Existing switches compliant with these requirements can be used, providing they were supplied by NetApp. Note that if the Cisco 9710 is used, only one switch with two blades (line cards) is required at each site.
- Associated cabling.
- Dedicated native Fibre Channel links between sites.

 © 2014 NetApp, Inc. All Rights Reserved.

Switch Licenses

For Brocade fabric:

On each switch:

- Brocade extended distance license. Required for intersite distances >7km.
- Brocade ports-on-demand (POD) licenses. Required to scale switch with additional ports.
- BNA Pro+ license enables ASUP™ support for the Brocade switches. One license of BNA Pro+ is included with every new fabric MetroCluster order when Brocade switches are selected.

For Cisco fabric:

On each switch:

- Cisco ENTERPRISE_PKG license to maximize buffer-to-buffer credits and provide QoS for the VSANs. This license is required when Cisco switches are selected.
- Cisco PORT_ACTIVATION_PKG license to scale switch with additional ports and to allow activation and deactivation of switch ports. This license is required for the 9148 switch only if more than the default licensed ports are being used. It is not required for 9222i or 9710 switches because all ports are enabled by default.
- The Cisco FM_Server_PKG license allows simultaneously fabric management and management of switches through a web browser. It also provides performance management features. This license is optional. It is only required if Fabric Manager features are used.

Data ONTAP Licenses

For versions of Data ONTAP earlier than 8.2, enable the following licenses (included as part of the base software package starting in Data ONTAP 7.3):

- cluster_remote license for the site failover on disaster functionality
- syncmirror_local license for synchronous mirroring across sites
- cluster license for controller failover functionality

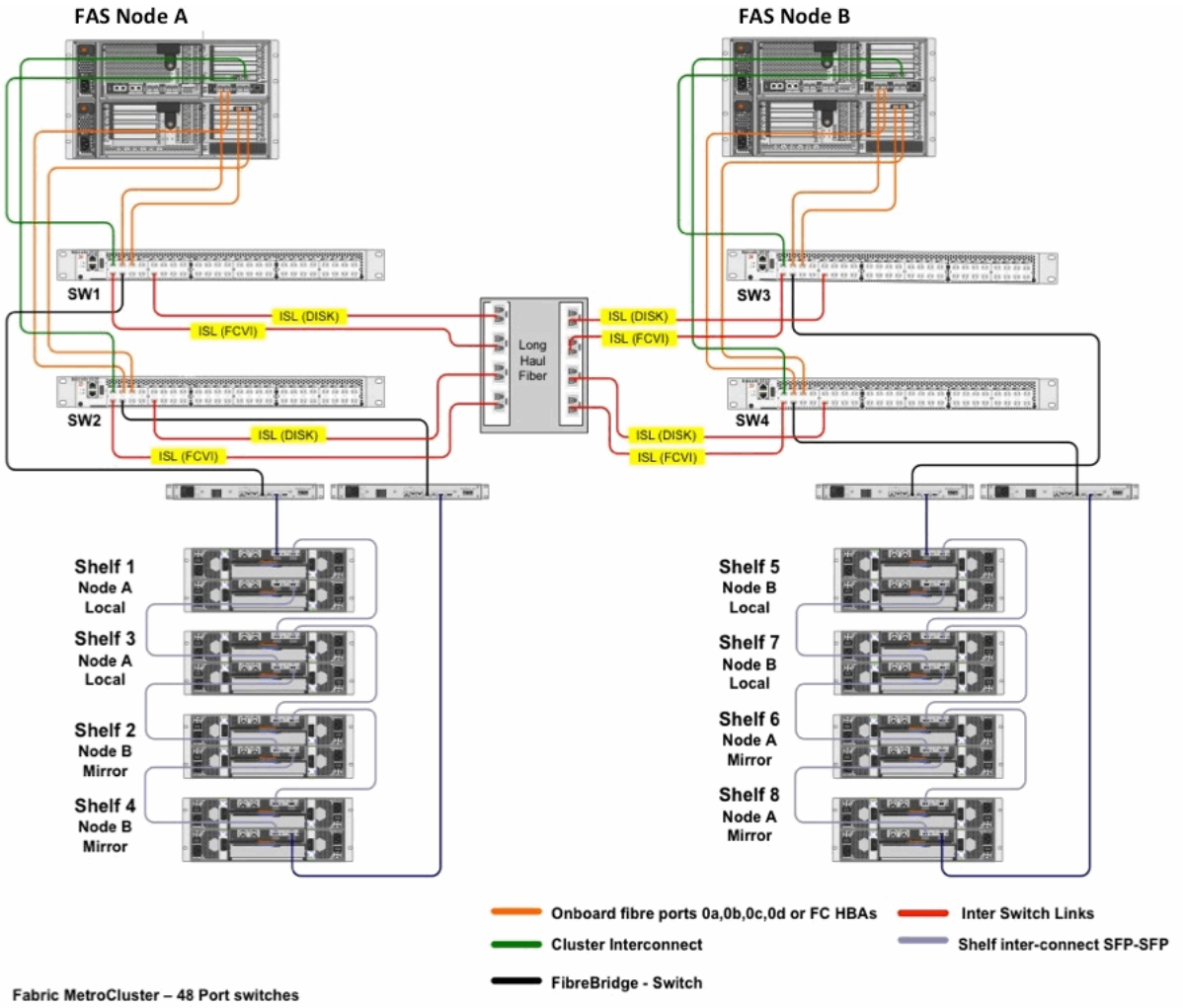
Starting in Data ONTAP 8.2, required licenses are replaced with options. Features specific to MetroCluster are enabled during configuration; for configuration steps see the High-Availability and MetroCluster Configuration Guide.

The following Data ONTAP options must be enabled on both nodes:

- cf.mode: You must set this option to ha.
- cf.remote_syncmirror.enable: Set the option to on.

 © 2014 NetApp, Inc. All Rights Reserved.

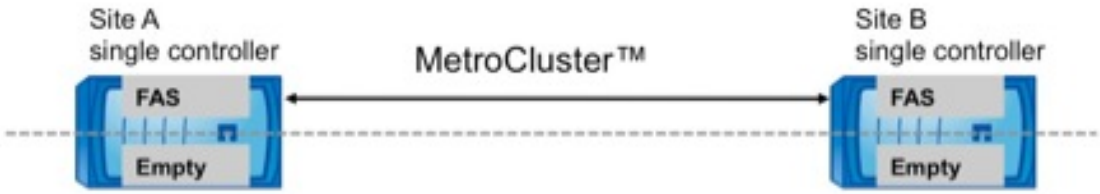
Figure 5) Fabric MetroCluster, FAS62xx single controller with FibreBridges and Brocade 6510 switches.



2.2 MetroCluster Dual-Chassis and Twin Configurations

Dual-chassis and twin configurations are supported in both stretch and fabric MetroCluster.

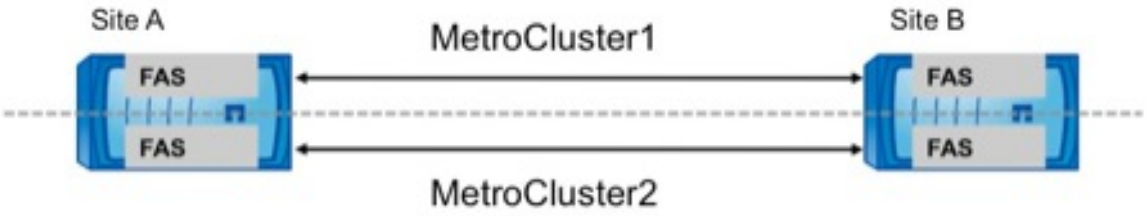
- **Dual-chassis MetroCluster configuration**. Two single controllers in individual chassis are connected:



- **Twin MetroCluster configuration**. Four controllers in 2 chassis form two independent MetroCluster configurations. Controller A at site A connects to controller A at site B; controller B at site A connects to controller B at site B. This configuration is supported with Data ONTAP release 7.3.3 and higher. This does **not** give you both local and site failover, because the controllers in each chassis are not

 © 2014 NetApp, Inc. All Rights Reserved.

connected to each other. It simply allows two MetroCluster configurations (instead of one) to share the chassis.



Note: If multiple MetroCluster configurations are implemented, either in a twin MetroCluster configuration or by implementing a pair of dual-chassis MetroCluster configurations, switch sharing between two MetroCluster configurations is supported for selected switches. This is currently possible with Brocade 5100 or 6510 and Cisco 9710; see section 9.8 for more information.

2.3 MetroCluster with Nonmirrored Aggregates

NetApp recommends mirroring all aggregates in a stretch or fabric MetroCluster configuration; however, some customers might choose not to mirror a subset of their aggregates. There are inherent risks associated with including nonmirrored aggregates in MetroCluster configurations:

- Controller panic on multidisk failure in a nonmirrored aggregate can affect overall MetroCluster availability.
- Nonmirrored aggregate data will be unavailable in the event of a site failure.

Note: All SAS storage, including nonmirrored aggregates, must utilize the FibreBridge in a fabric MetroCluster configuration; directly attached storage is not supported.

 © 2014 NetApp, Inc. All Rights Reserved.

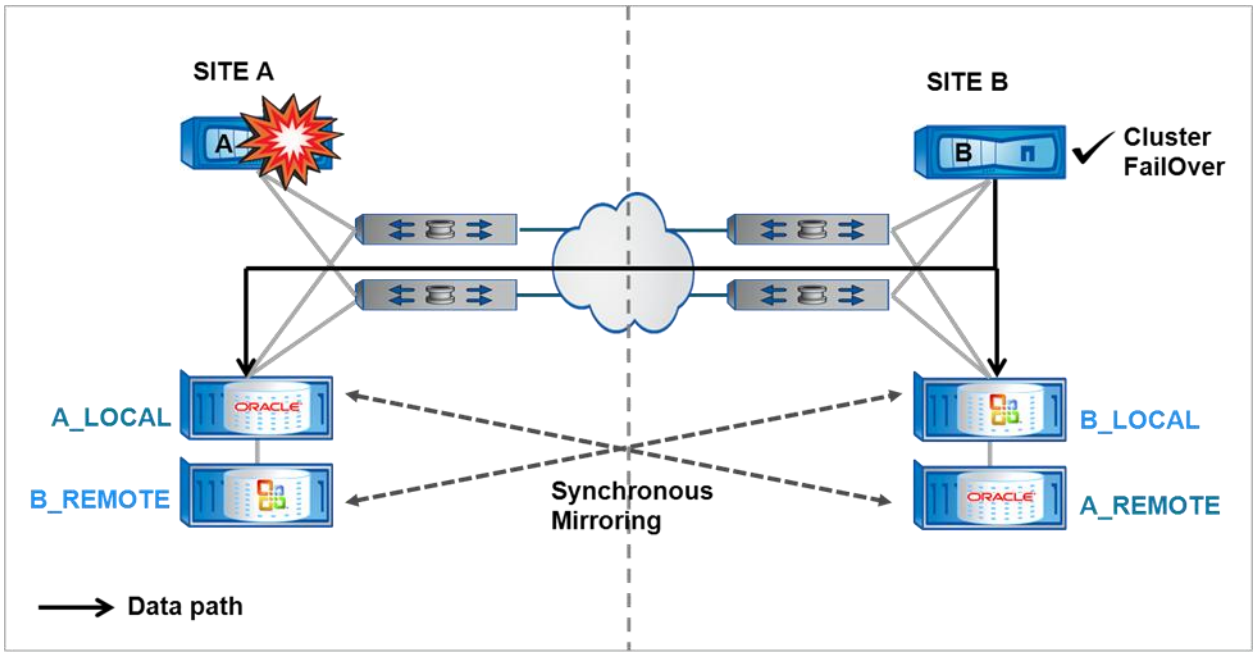
3 MetroCluster Failure Handling

MetroCluster components—cluster failover, SyncMirror, cluster failover in case of disaster, and geographical separation—work individually or in tandem to protect against failures. The following graphics illustrate failure scenarios and the components responsible for protection. A fabric MetroCluster configuration is demonstrated in the following figure; the same concepts apply to stretch MetroCluster.

3.1 Controller Failure

In Figure 6, normal HA (CFO) causes controller B to take over controller A's data-serving operations. Failover is automatic.

Figure 6) Controller failure: automatic failover to surviving controller.

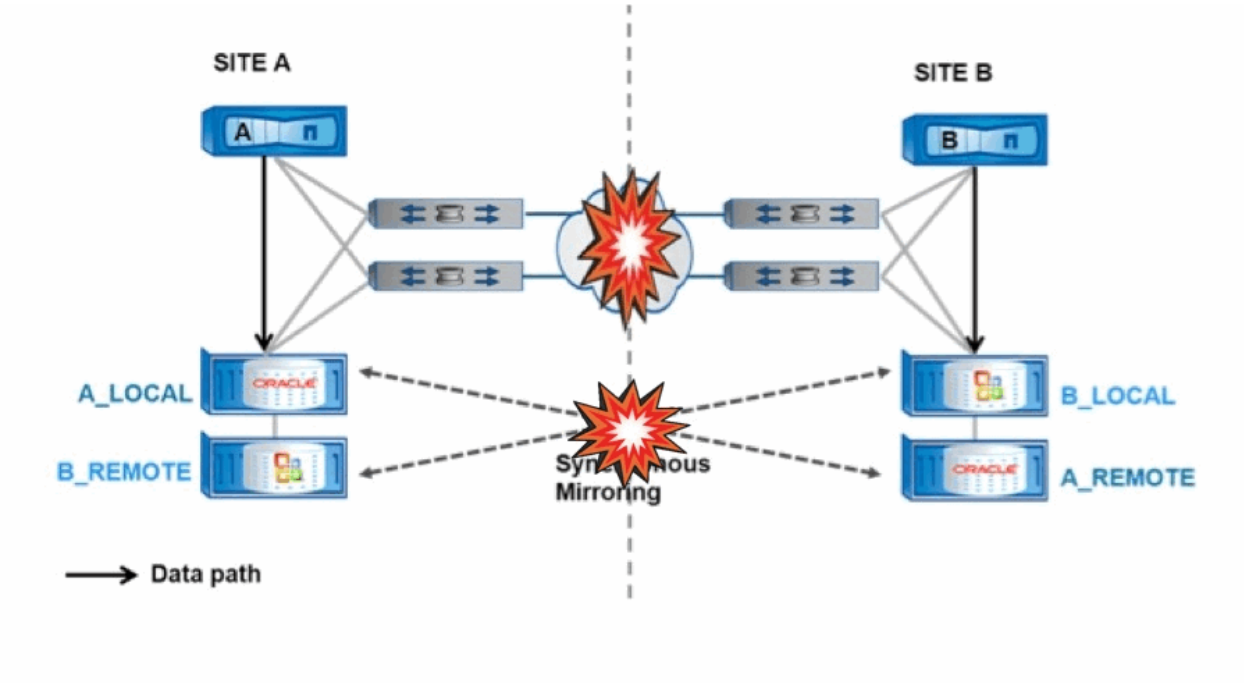


 © 2014 NetApp, Inc. All Rights Reserved.

3.2 Shelf Failure

In normal MetroCluster operation, reads are serviced from the local plex, and writes are sent to both the local and remote plex. If a shelf on the site A controller fails, as shown in Figure 7, the reads will be redirected to the remote plex, plex A_REMOTE in this instance. Writes will also be sent only to the remote plex until the shelf is repaired or replaced. The process of directing reads and writes as necessary is seamless and automatic.

Figure 7) Shelf (or plex or aggregate) failure: automatic and seamless failover to mirrored plex.

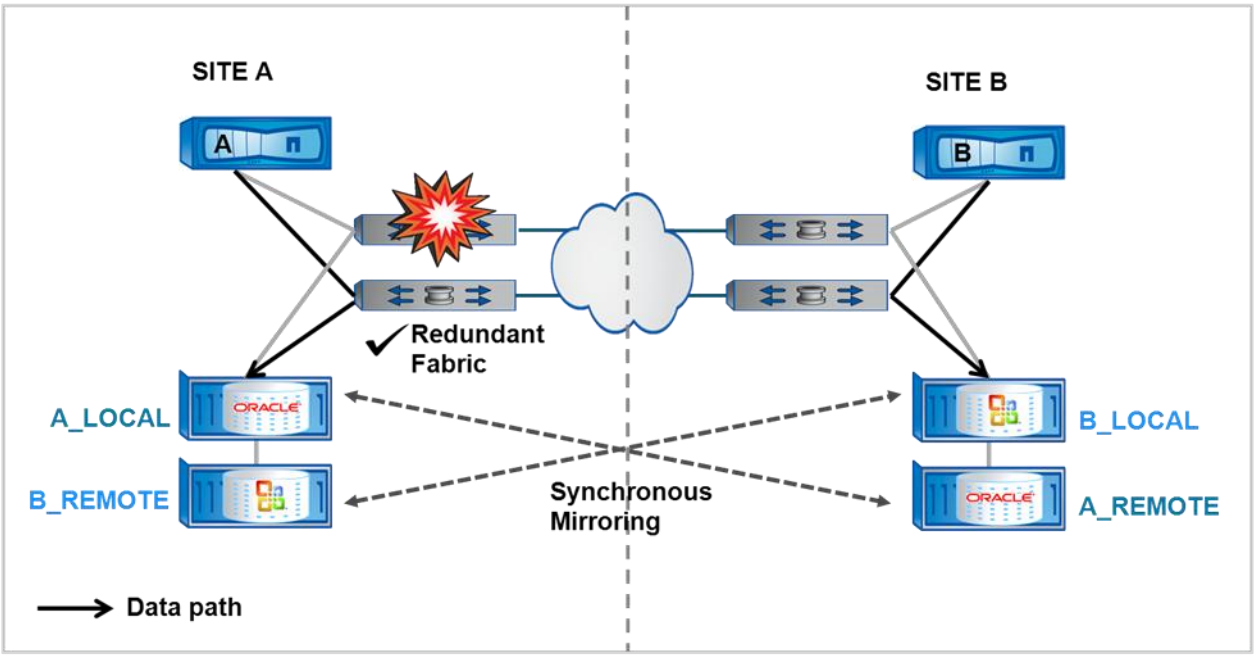


 © 2014 NetApp, Inc. All Rights Reserved.

3.3 Switch Failure

Fabric MetroCluster configurations use four fabric switches (either Brocade or Cisco) per MetroCluster configuration, creating two fabrics between sites. Failure of one fabric or a switch simply causes failover to the surviving fabric. Failover is automatic.

Figure 8) Switch failure: automatic failover to redundant fabric.

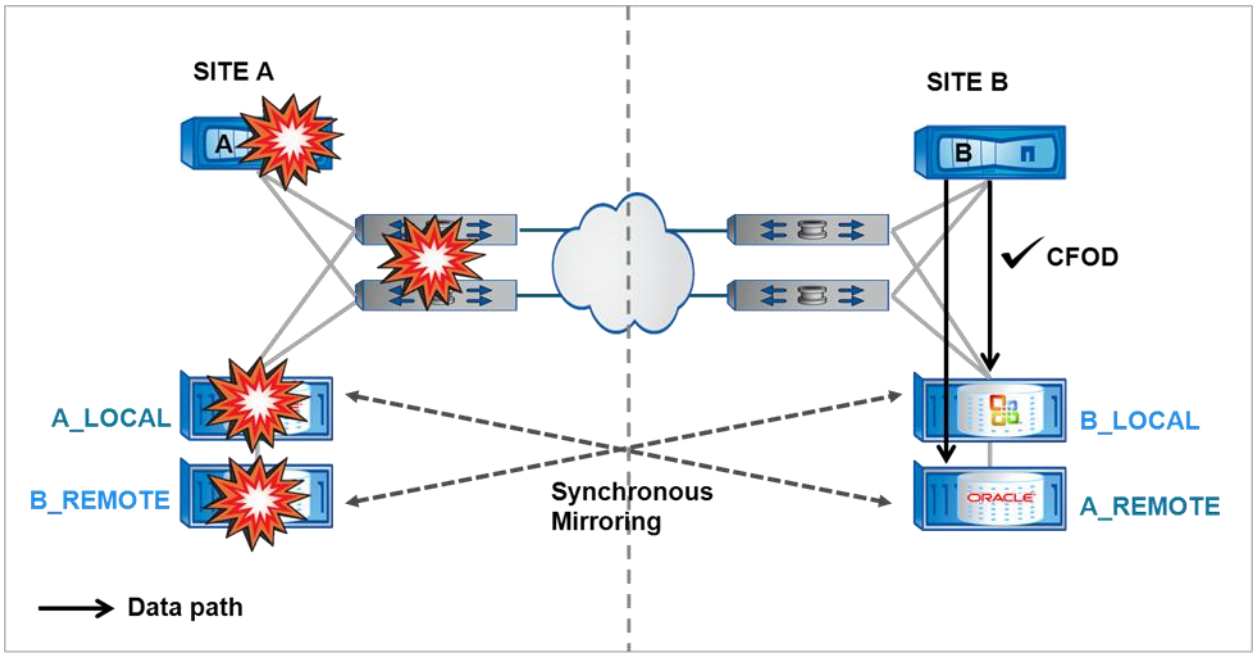


 © 2014 NetApp, Inc. All Rights Reserved.

3.4 Complete Site Failure

Failure of an entire site requires an administrator to initiate a takeover from the surviving site using a single CFOD command. Requiring manual takeover eliminates ambiguity between “site disaster” and “loss of connectivity between sites.” Normally, MetroCluster cannot differentiate between these two scenarios; however, witness software called the MetroCluster tie-breaker (MCTB) can be implemented to differentiate between them. See the section “Monitoring MetroCluster and MetroCluster Tools.”

Figure 9) Complete site failure: Enter CFOD command on surviving controller.

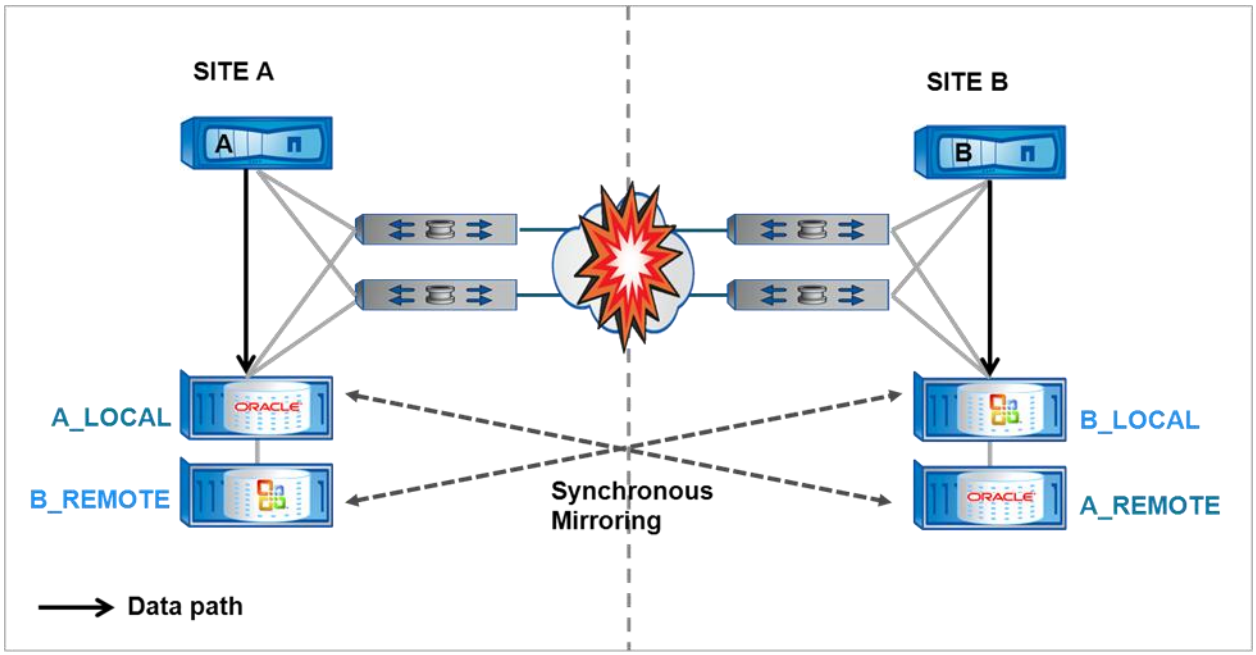


 © 2014 NetApp, Inc. All Rights Reserved.

3.5 Link Loss Between Sites

MetroCluster takes no action if both links are lost between sites. Each controller continues to serve data normally, but the mirrors are not written to because access to them is lost.

Figure 10) Link loss between sites: MetroCluster takes no action except to suspend mirroring.

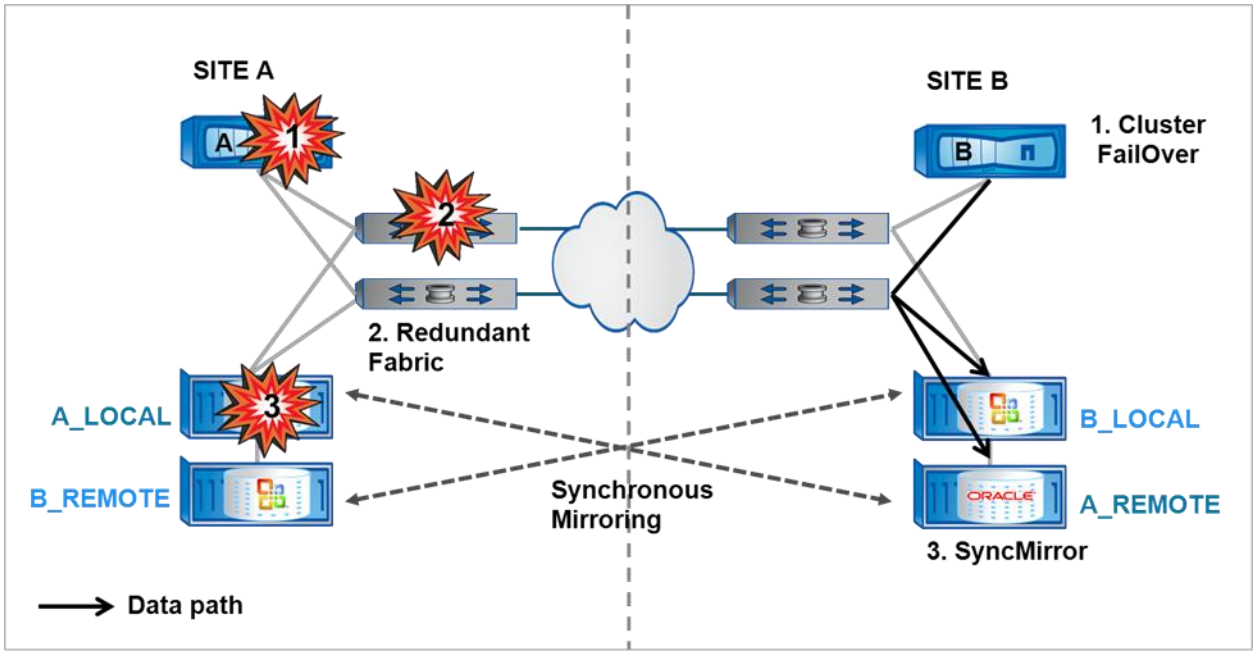


 © 2014 NetApp, Inc. All Rights Reserved.

3.6 Rolling Failures

The previous failure scenarios covered instances of single component failure and complete site disaster. Instances in which multiple components fail sequentially are called *rolling failures*. For example, failure of a controller followed by the failure of a fabric followed by the failure of a shelf results in CFO, fabric redundancy, and SyncMirror sequentially protecting against downtime and data loss.

Figure 11) Example of rolling failures.



Conceptually, all the preceding failure scenarios apply to all MetroCluster configurations:

- DS14-based stretch MetroCluster configurations (except switch/fabric failures)
- DS14 FC-based fabric MetroCluster configurations
- SAS-based stretch MetroCluster configurations (except switch/fabric failures)
- SAS-based fabric MetroCluster configurations

 © 2014 NetApp, Inc. All Rights Reserved.

4 SAS-Based MetroCluster Configurations

MetroCluster supports SAS shelves. The shelves attach to the MetroCluster configuration using SAS copper cables, SAS optical cables, or FibreBridges:

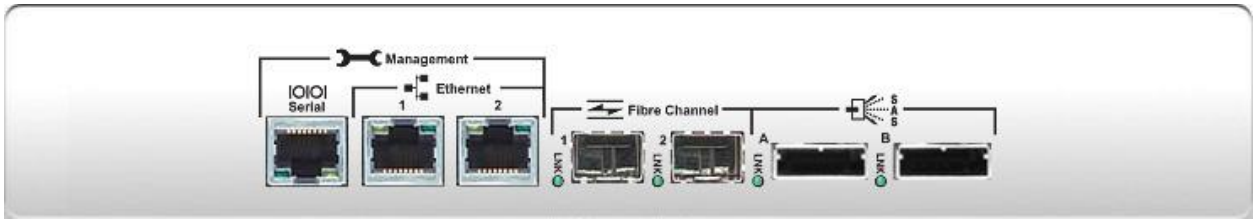
- Fabric MetroCluster configurations create storage area network (SAN) fabrics across sites; therefore, the storage must understand the Fibre Channel protocol. The 6500N FibreBridge performs protocol conversion from SAS to FC, enabling SAS disks to appear as LUNs in a MetroCluster fabric.
- Stretch MetroCluster configurations achieve campus distances (max 500m) using cables of sufficient length and optionally patch panels. The cables extend from controller to controller and from controller to shelf across campus distances. SAS cables are available in copper and optical mediums to reach these distances. FibreBridges may also be used.

4.1 FibreBridge Details

Figure 12 and Figure 13 show the FibreBridge front and back, respectively.

Figure 12) FibreBridge 6500N front.

Figure 13) FibreBridge 6500N back.



FibreBridge 6500N physical characteristics are:

- 1U of rack space
- 2 8Gb Fibre Channel Small Form Factor Pluggable Plus (SFP+) ports
- 2X4 6Gb SAS QSFP+ ports (only 1 port used)
- 2 Ethernet ports for management
- 1 serial port for initial configuration

FibreBridge 6500N configuration:

- 2 FibreBridges required per disk stack (for redundancy and performance)

 © 2014 NetApp, Inc. All Rights Reserved.

- Data ONTAP 8.1 or higher required
- 10 shelves maximum per stack, unless SSD drives are present, in which case the limit is 2 shelves of SSD drives per FibreBridge stack and a maximum of 4 shelves overall (See section 9.7 for more information on mixed shelf requirements.)
- Existing DS14-based MetroCluster configurations can be expanded with SAS shelves using FibreBridge

FibreBridge 6500N performance:

- 63k IOPS maximum for 4kB reads
- 52k IOPS maximum for 4kB writes
- 1.1GB/sec throughput maximum, using both FC ports and 128kB reads and writes

For more information on FibreBridge performance, see section 10.4.

4.2 Stretch MetroCluster with FibreBridge

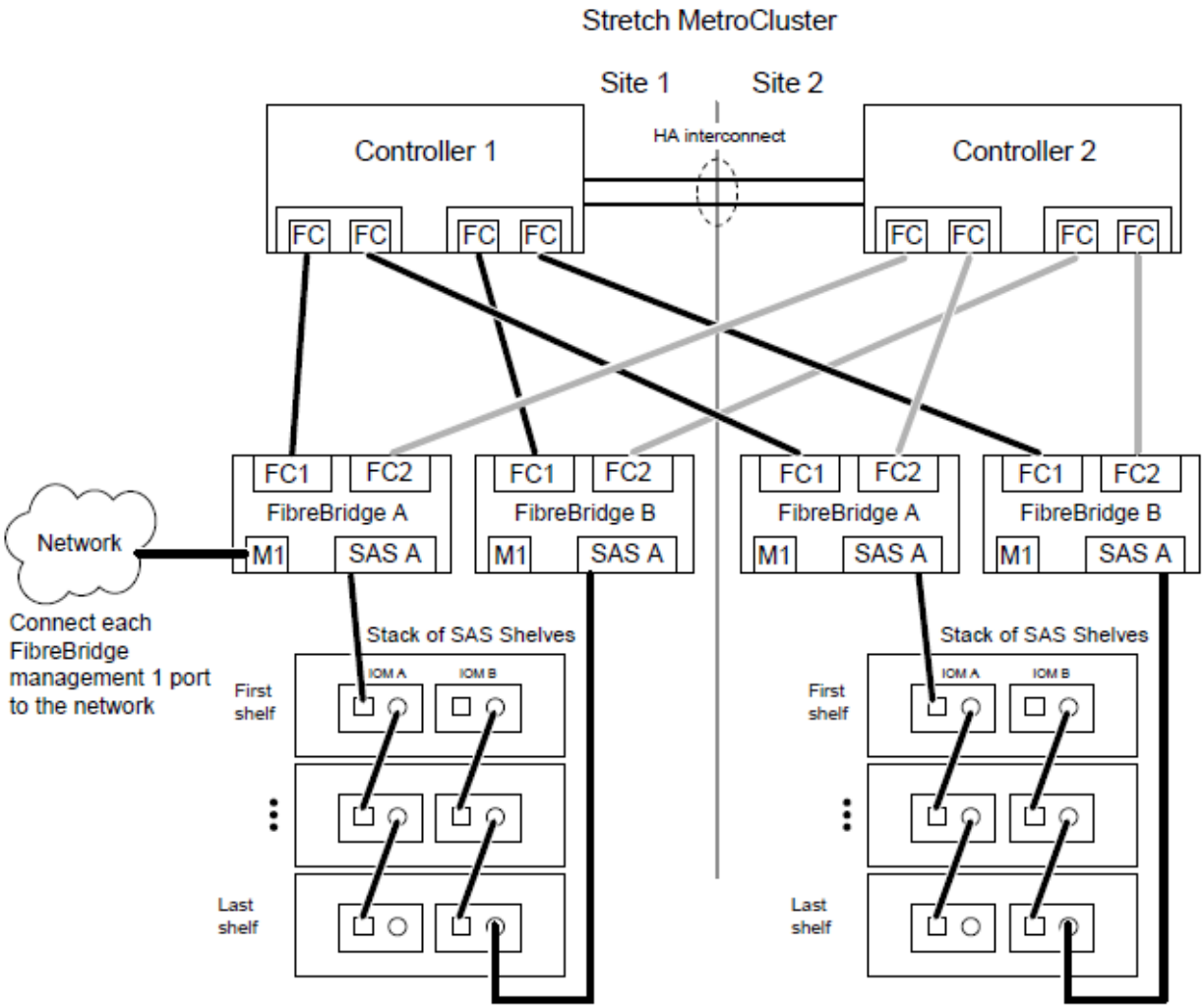
In stretch MetroCluster, a minimum of four FibreBridges per configuration are required, two on each site.

Existing DS14 stretch MetroCluster configurations can be expanded with SAS shelves using FibreBridge.

Figure 14 shows the cabling diagram. Each of the required four FC ports on each controller connects to an FC port on each bridge.

 © 2014 NetApp, Inc. All Rights Reserved.

Figure 14) SAS-based stretch MetroCluster using FibreBridge.



4.3 Fabric MetroCluster with FibreBridge

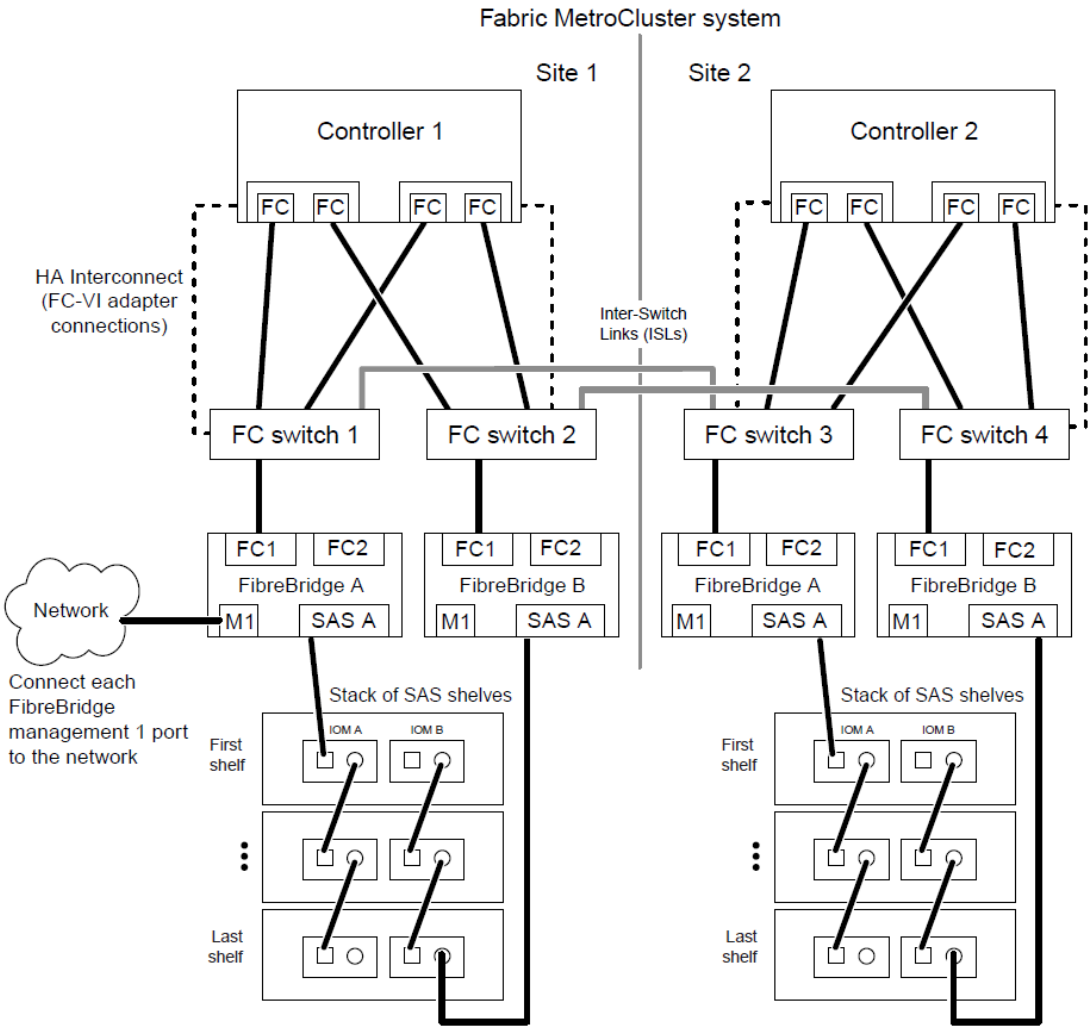
In fabric MetroCluster, a minimum of four FibreBridges per configuration are required, two on each site.

Existing DS14 stretch MetroCluster configurations can be expanded with SAS shelves using FibreBridge.

Figure 15 shows the cabling diagram. Two of the required four FC ports on each controller connect to each FC switch. Each FibreBridge also connects to a switch.

 © 2014 NetApp, Inc. All Rights Reserved.

Figure 15) SAS-based fabric MetroCluster using FibreBridge.



4.4 FibreBridge Setup and Configuration

Refer to Configuring a MetroCluster system with SAS disk shelves and FibreBridge 6500N bridges on the NetApp Support site for step-by-step configuration information.

For manageability and supportability purposes, a unique shelf ID is required for SAS shelves.

4.5 Shelf Mixing Rules for Stretch MetroCluster

Refer to section 9.7 for information on mixing SAS, SATA, and SSD storage in MetroCluster.

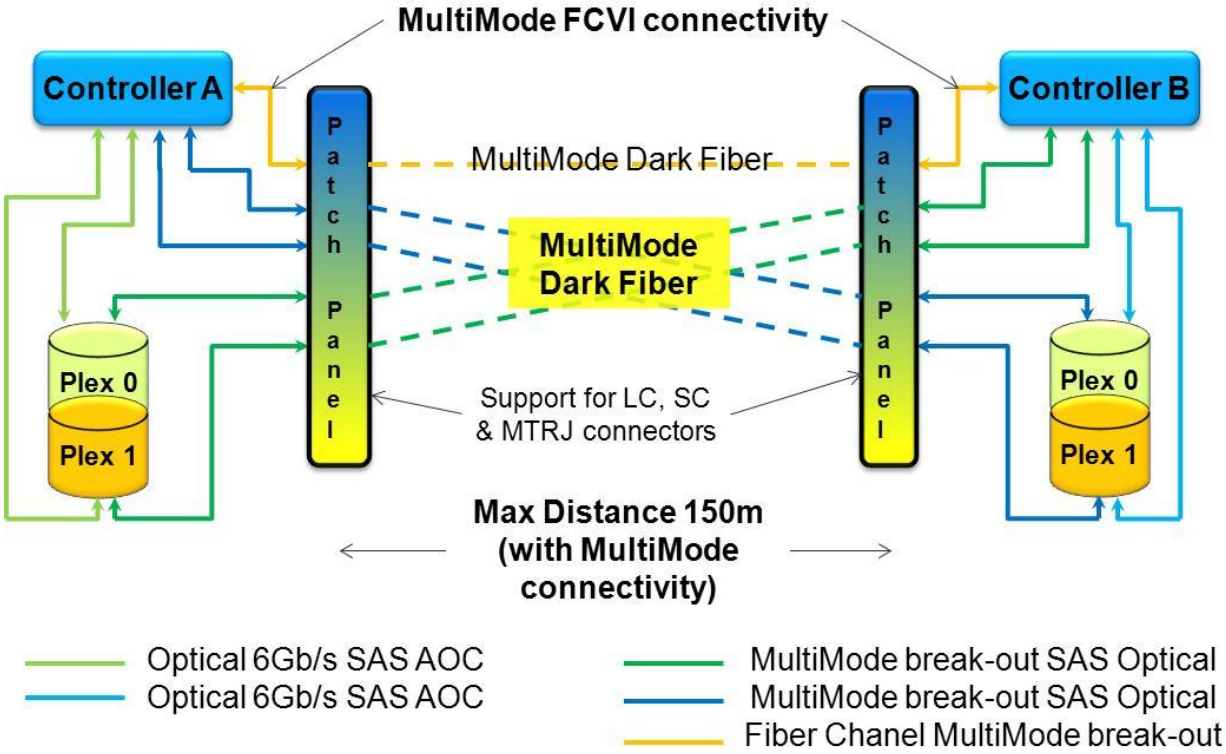
4.6 Stretch MetroCluster with SAS Optical Cables

Connections are made using active optical cables with existing QSFP connectors on supported SAS HBAs. Optical SAS cables are available in both multimode and single-mode fiber types and are compatible with existing optical patch panel infrastructure. Figure 16 and Figure 17 show connections with multimode and single-mode fiber, respectively.

 © 2014 NetApp, Inc. All Rights Reserved.

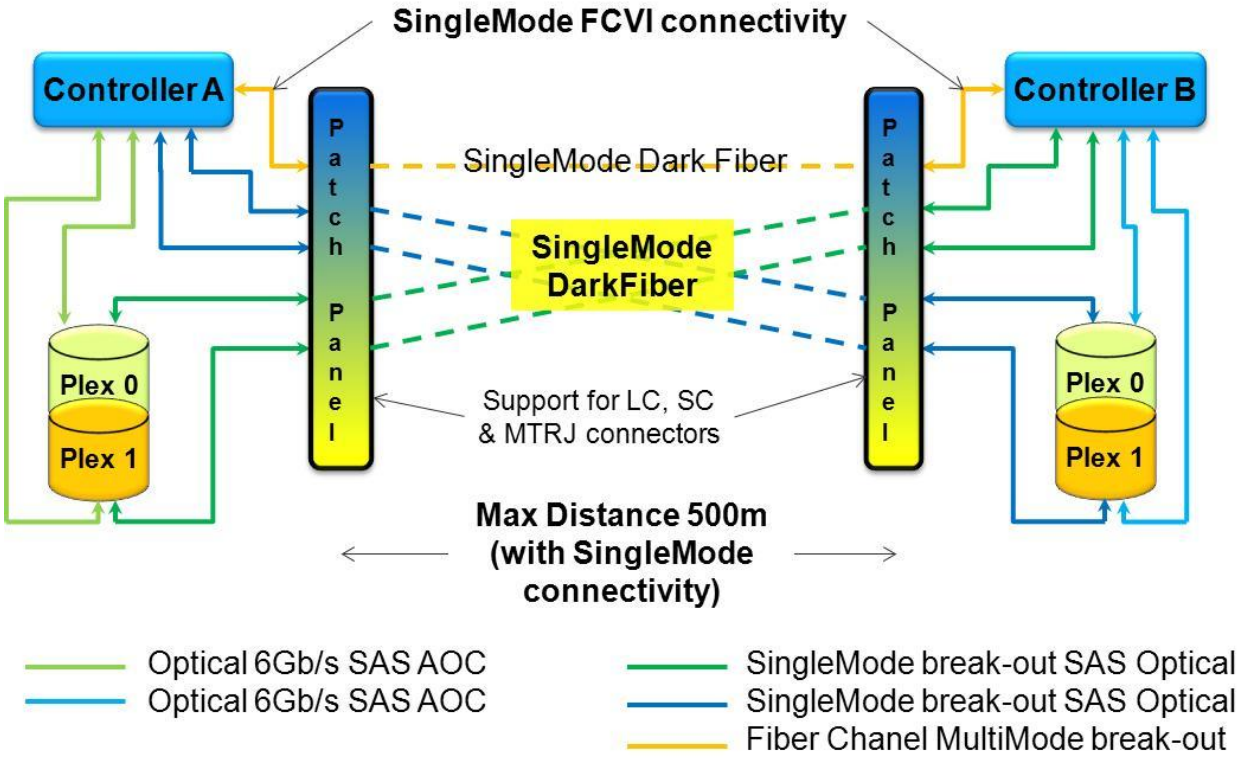
Data ONTAP version 8.1.3 and 8.2.1 or later is required. Optical SAS is NOT supported with Data ONTAP 8.2.

Figure 16) SAS-based stretch MetroCluster using multimode SAS optical.



 © 2014 NetApp, Inc. All Rights Reserved.

Figure 17) SAS-based stretch MetroCluster using single-mode SAS optical.



 © 2014 NetApp, Inc. All Rights Reserved.

5 Stretch MetroCluster Considerations

A stretch MetroCluster configuration provides continuous availability across data centers in adjacent buildings or across floors or across campus distances. A maximum of 500m between controllers is supported.

Keep the following planning points in mind when deploying a stretch MetroCluster configuration.

5.1 Stretch MetroCluster and MPHA

All stretch MetroCluster configurations follow the multipath HA (MPHA) cabling rules of HA NetApp pairs.

5.2 Sufficient Initiator Ports or Initiator HBAs

Controllers in a stretch MetroCluster (non-FibreBridge) configuration attach directly to the local and remote storage, following MPHA cabling rules. As a result, a minimal stretch MetroCluster configuration with a single stack (or loop) on either site requires four initiator ports per controller. Additionally, the FC-VI cluster interconnect adapter consumes one controller expansion slot, reducing the number of empty slots for additional FC initiator HBAs.

5.3 Stretch MetroCluster Mixed Disk Pool Configurations

A minimal stretch MetroCluster configuration with a single stack (or loop) on either site requires four initiator ports per controller. This configuration results in each stack including a mix of two disk pools owned by the different controllers:

Site A Stack/Loop = Controller1_pool0 and Controller2_pool1
Site B Stack/Loop = Controller2_pool0 and Controller1_pool1

Replacing a failed disk in a configuration might require user intervention to manually assign a new disk to the correct controller and pool.

It is highly recommended that each individual shelf should be assigned to only node and pool. Although it is possible to share a shelf between two pools, this is not recommended.

5.4 Stretch MetroCluster with FAS3210 and FAS8020

FAS3210 and FAS8020 controllers have limited on board ports and expansion slots: two initiator onboard ports and two expansion slots. The FC-VI cluster interconnect adapter consumes one expansion slot, leaving one empty slot for an initiator HBA, because a minimum of four initiators are required for stretch MetroCluster configurations.

Note: In this configuration, there are no free expansion slots for other cards. See KB article 1012811 for details.

5.5 Stretch MetroCluster Spindle Limits

Stretch MetroCluster configurations adhere to platform spindle limits. However, because each controller attaches to all storage, the platform spindle limit applies to the entire configuration. For instance, if the spindle limit for FAS80x0 is *n*, then even though there are two controllers, the spindle limit for a FAS80x0 stretch MetroCluster configuration remains *n*.

5.6 Stretch MetroCluster Distances

The maximum distance between the two controllers in stretch MetroCluster is 500m. Distance is a function of speed and fiber cable type, as shown in the following two tables of maximum distances according to FC industry standards.

 © 2014 NetApp, Inc. All Rights Reserved.

Table 1) Stretch MetroCluster maximum distances 4 and 8Gbps SFP.

Maximum Distance Supported with Stretch MetroCluster, 4 and 8Gbps SFP				
Speed (Gbps)	Maximum Distance (m)			
	4Gbps SFP		8Gbps SFP	
	OM2	OM3, OM3+/OM4	OM2	OM3, OM3+/OM4
2	300	500	300	500
4	150	270	150	270
8	NA	NA	50	150
16	NA	NA	NA	NA

Table 2) Stretch MetroCluster maximum distance, 16Gbps SFP.

Maximum Distance Supported with Stretch MetroCluster, 16Gbps SFP				
Speed (Gbps)	Maximum Distance (m)			
	16Gbps SW SFP			16Gbps LW SFP
	OM2	OM3,	OM3+/OM 4	Single-Mode (SM) Fiber
2	NA	NA	NA	NA
4	150	270	270	500
8	50	150	170	500
16	35	100	125	500

Theoretically, distances greater than 500m could be possible at lower speeds. **However, stretch MetroCluster is qualified up to a maximum of 500m only. For distances greater than 500m, a fabric MetroCluster configuration is necessary.**

See the Interoperability Matrix on the NetApp Support site for the latest information.

6 Fabric MetroCluster Hardware and Software Components

A fabric MetroCluster configuration provides continuous availability across sites for distances up to 200km, or 125 miles.

6.1 Components

Section 2.1 provided requirements for the hardware and software components in a fabric MetroCluster configuration. Here we examine some components in more detail:

- **Cluster interconnect**. FC-VI adapters are required for all fabric MetroCluster configurations, one per controller.

 © 2014 NetApp, Inc. All Rights Reserved.

- **Switches**. Fabric MetroCluster implements two fabrics (one for redundancy) across sites. Each fabric consists of two switches (one on each site), so therefore four switches per MetroCluster configuration. With the Cisco 9710 director-class switch, one 9710 is installed in each site, with two line cards or blades to provide the two fabrics. The controllers and storage connect to the switches directly (controllers do not directly attach to storage as in configurations other than MetroCluster), and the switches cannot be shared by traffic other than MetroCluster. Switch sharing between MetroCluster configurations is supported with certain switches; see section 9.8.

- **Storage adapters/HBAs**. Four FC initiator ports are required per controller; two FC initiators connect to each fabric (two fabrics = four FC initiator ports).

- **Storage**. Make sure there is sufficient storage for the data mirrors.

- **FibreBridges**. If using SAS shelves, two FibreBridges are required per stack of SAS shelves.

Intersite Infrastructure

- **Customers buy dark fiber**. Direct connections using long-wave SFPs. NetApp does not source long-wave SFPs for large distances >30km.
- **Customers lease metrowide transport services from a service provider**. Typically provisioned by dense wavelength division multiplexer/time division multiplexer/optical add drop multiplexer (DWDM/TDM/OADM) devices. Make sure the device is supported by the fabric switch vendor (Brocade or Cisco).
- **Dedicated bandwidth between sites (mandatory)**. One interswitch link (ISL) per fabric, or two ISLs if using the traffic isolation (TI) feature.

See section 9.3 for guidance regarding fabric MetroCluster ISL requirements. Always refer to the Interoperability Matrix on the NetApp Support site for the latest information on components and compatibility.

6.2 Choice of Fabric Switch

Before Data ONTAP 8.1.1, Fabric MetroCluster supported only Brocade switch fabrics. From Data ONTAP 8.1.1 onward, Cisco switch fabrics are also supported. Section 2.1 contains more guidance on switch requirements.

Only supported switches, supplied by NetApp, as listed in the Interoperability Matrix may be used with fabric MetroCluster.

7 Fabric MetroCluster Using Brocade Switches

Keep in mind the following when deploying a fabric MetroCluster configuration with Brocade switches.

7.1 Brocade: Traffic Isolation Zones

The TI zone feature of Brocade switches (FOS 6.0.0b or later) allows control of the flow of interswitch traffic by creating a dedicated path for traffic flowing from a specific set of source ports. In the case of a fabric MetroCluster configuration, the traffic isolation feature can be used to dedicate an ISL to high-priority cluster interconnect traffic.

Why Traffic Isolation Zones?

Customers can benefit from using two ISLs per fabric rather than one to separate high-priority cluster interconnect traffic from other traffic, preventing contention on the back-end fabric and for additional

 © 2014 NetApp, Inc. All Rights Reserved.

bandwidth in some cases. The TI feature is used to enable this separation. The TI feature provides better resiliency and performance but requires more fiber between sites.

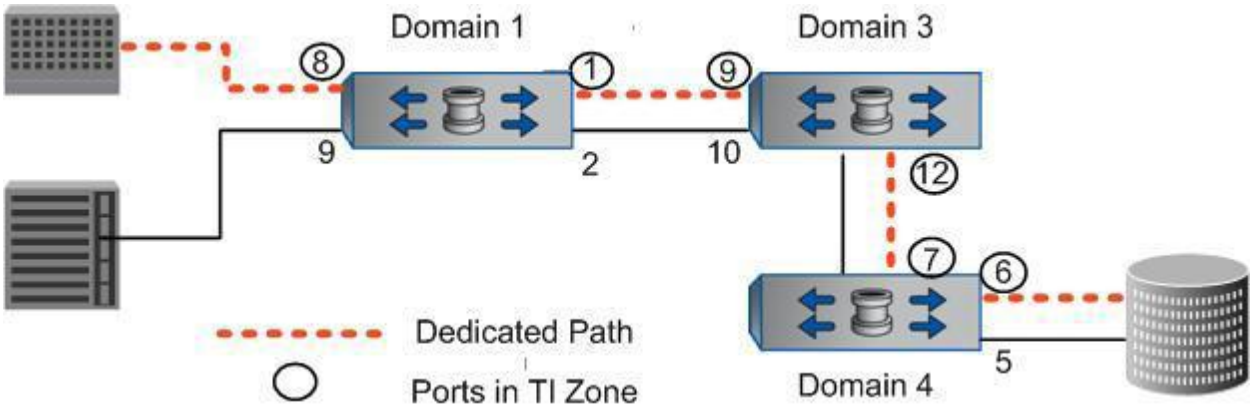
Implementation

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 interswitch 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.

TI Traffic Flow

In Figure 18, 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.

Figure 18) Brocade TI zones.



Configuration Rules for Traffic Isolation

- Fabric MetroCluster supports traffic isolation only on the Brocade 300, 5100, and 6510 switches.
- Ports in a TI zone must belong to switches running fabric OS v6.0.0b or later.
- Note that as stated later, all four switches in a MetroCluster fabric must run the same FabOS version.

Traffic Isolation Zone Failover

A TI zone can have failover enabled or disabled.

Generally speaking:

- If failover is enabled and if the dedicated path cannot be used, then the TI zone traffic will use a nondedicated path.
- If failover is disabled and if the dedicated path cannot be used, then traffic for that TI zone is halted until the dedicated path is fixed.

In the context of a fabric MetroCluster configuration:

- If failover has been disabled and the dedicated path cannot be used, traffic is never halted, but rather traffic flows through the redundant fabric.

 © 2014 NetApp, Inc. All Rights Reserved.

See the Fabric-Attached MetroCluster Systems: Brocade Switch Configuration Guide on the NetApp Support site for more information on setting up TI zones and the recommendations for TI zone failover.

7.2 Brocade: Mixing Switch Models

NetApp highly recommends that all four switches are the same Brocade model, but that is not required as long as all the following rules are complied with:

- All switches support the same maximum port speed. For example, all 8Gbps switches.
- Both switches on each site are licensed for the same number of ports.
- All switches connected to the same MetroCluster configuration run the same Brocade fabric operating system (FOS) version except for a brief period during nondisruptive upgrade (NDU).
- The switches on a site are the same model. Different models can be used in the two sites.

8 Fabric MetroCluster Using Cisco Switches

Keep in mind the following points when deploying a fabric MetroCluster configuration with Cisco switches.

8.1 Cisco: Port Groups

Both the Cisco MDS 9148 and the Cisco MDS 9222i utilize the concept of port groups. A certain number of buffer-to-buffer credits are allocated to each port group, which are then shared between the ports within the group. Bandwidth allocation takes place at the port group level.

See the Fabric-Attached MetroCluster Systems Cisco Switch Configuration Guide on the NetApp Support site for step-by-step information on switch configuration. Also, see the appendix for cabling diagrams.

Cisco MDS 9148: 48-Port 8Gbps

- Each port group includes four ports: 1–4, 5–8, and so on up to 41–44 and 45–48.
- 128 buffer-to-buffer credits are allocated per port group.

To maximize bandwidth:

Because bandwidth allocation takes place at the port group level, utilize one port per port group for controller, storage, and ISL connectivity. There are a total of 12 port groups.

To maximize distance:

To maximize distance between sites, the maximum number of buffer-to-buffer credits must be allocated to the interswitch links (ISLs). Therefore, we assign just one buffer-to-buffer credit (the minimum) to non-ISL ports in the ISL port group, saving the rest of the buffer-to-buffer credits for the ISL port.

For example, say we use:

- Port group 41–44 with port 41 for one ISL
- Port group 45–48 with port 45 for the other ISL

Then assign just one buffer-to-buffer credit to ports 42, 43, and 44 and ports 46, 47, and 48. Doing so will leave 125 buffer-to-buffer credits each to port 41 and port 45 (the ISL ports).

Cisco MDS 9222i: 18-Port 4Gbps

- Each port group includes six ports: 1–6, 7–12, and 13–18.
- 4,509 buffer-to-buffer credits are allocated per port group.

To maximize bandwidth:

 © 2014 NetApp, Inc. All Rights Reserved.

Because bandwidth allocation takes place at the port group level and there are only three port groups, port group utilization is prioritized based on the nature of the traffic. The higher the priority of the traffic, the less the port group is shared.

Cluster interconnect traffic and ISL traffic are considered high-priority traffic; storage traffic is of relatively lower priority. Port groups are shared accordingly:

- FC-VI (cluster interconnect) uses one port group: no sharing.
- ISLs share one port group: two ports share a port group.
- FC initiator and storage connections: three or more ports share a port group.

To maximize distance:

Because the MDS 9222i has more than sufficient (4,059) buffer-to-buffer credits per port group, no special configuration is necessary for the ISL ports, unlike the MDS 9148.

Cisco MDS 9222i with FC Module: 18-Port 4Gbps with 4/44-Port FC Module

18-port 4Gbps:

- Each port group includes six ports: 1–6, 7–12, and 13–18.
- 4,509 buffer-to-buffer credits are allocated per port group.

4/44-port FC module:

- Each port group includes 12 ports: 1–12, 13–24, 25–36, 37–48.
- Four of the 48 ports work at 8Gbps.

When using the MDS 9222i with the 4/44-port FC module, two configurations are possible:

- Maximize controller-to-storage bandwidth by using the four 8Gbps ports on the FC module for the FC initiator and FC-VI (cluster interconnect) connections, one port per port group.
- Maximize ISL bandwidth by using the four 8Gbps ports on the FC module for the ISLs and FC-VI (cluster interconnect) connections, one port per port group.

Maximize distance:

Because the MDS 9222i has more than sufficient (4,059) buffer-to-buffer credits per port group, no special configuration is necessary for the ISL ports (as opposed to the MDS 9148).

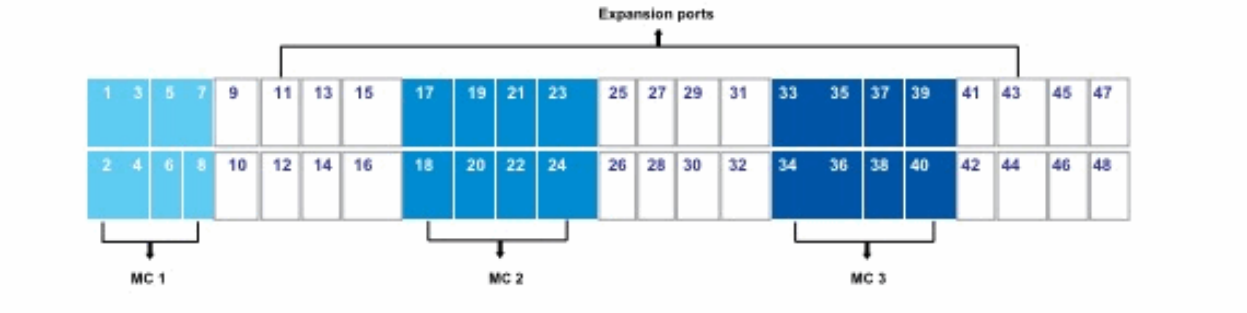
8.2 Cisco 9710

The Cisco 9710 is a director-class switch. When used with MetroCluster, each site will have one 9710 with two blades (line cards) installed at each site: one blade for each fabric. A single 48-port module or blade can connect to a maximum of six fabric MetroCluster configurations; however, it is recommended to attach only up to three. ISLs cannot be shared between MetroCluster configurations; each must have dedicated ISLs. Each fabric MetroCluster configuration must use eight consecutive ports on the module. A group of eight ports is called a port chunk. A port chunk of eight ports is required even if the configuration requires fewer than eight ports. It is recommended to leave a vacant port chunk adjacent to each port chunk that is attaching a MetroCluster configuration, in order to allow for future expansion, because adjacent ports must be used for each MetroCluster configuration.

The recommended port chunk configuration is shown in Figure 19 for three MetroCluster configurations. One MetroCluster configuration uses the port chunk consisting of ports 1 through 8. The second port chunk, ports 9 through 16, is vacant for expansion. Similarly, a second MetroCluster configuration uses port chunk with ports 17 through 24, with expansion ports 25 through 32.

 © 2014 NetApp, Inc. All Rights Reserved.

Figure 19) Port chunking with Cisco 9710 switch.



8.3 Cisco: VSANs

Two virtual SANs (VSANs) are implemented on the Cisco fabric switches, one for cluster interconnect traffic and another for storage traffic. VSANs are used in conjunction with Cisco's quality of service (QoS) feature to prioritize one VSAN traffic type over the other when traversing the ISLs. Cluster interconnect traffic is considered higher priority than storage traffic, and thus the VSAN corresponding to cluster interconnect traffic receives a higher priority.

See the Fabric-Attached MetroCluster Systems Cisco Switch Configuration Guide on the NetApp Support site for step-by-step information on switch configuration.

8.4 Cisco: Cabinet Integration

The MDS 9148 and the MDS 9222i are not cabinet integrated. This means that when a fabric MetroCluster configuration with Cisco switches ships from the factory, the Cisco switches ship separately and not within the same rack housing the controller, FibreBridge, and disk shelves. The MDS 9148 is cabinet qualified and can be installed in the NetApp rack. The MDS 9222i can be installed in an adjacent third-party cabinet or two-post telco. Plan data center rack space accordingly.

See the release notes for information on the current state of Cisco switch cabinet integration.

8.5 Cisco: Mixing Switch Models

NetApp highly recommends that all four switches (or 2 9710) be the same Cisco model, but this is not required as long as these requirements are met:

- All switches support the same maximum port speed.
- The switches on each site are licensed for the same number of ports.
- All switches connected to the same MetroCluster configuration run the same NX-OS firmware version, except for a brief period during NDU operation.
- The switches on a site are the same model. Different models can be used in the two sites.

9 Fabric MetroCluster Considerations

Keep in mind the following points when deploying a fabric MetroCluster configuration.

9.1 Fabric MetroCluster Traffic Flow Considerations

Fabric MetroCluster configurations create two types of traffic:

- High-priority cluster interconnect or FC-VI traffic (NVLOG mirroring, heartbeat).
- Relatively lower priority disk mirroring traffic: In normal mode (nonfailover), disk traffic flows through both fabrics, approximately 50% per fabric.

 © 2014 NetApp, Inc. All Rights Reserved.

9.2 Fabric MetroCluster Distance Considerations

The maximum theoretical distance attained by a fabric MetroCluster configuration is a function of:

- ISL speed
- Number of ISLs per fabric (one or two)
- Fabric switch buffer-to-buffer credits

Although theoretical distances greater than 200km (~125 miles) are possible, fabric MetroCluster supports a maximum of 200km between sites with Data ONTAP 8.2 and SAS-based storage.

See the Interoperability Matrix on the NetApp Support site for the latest information on fabric MetroCluster supported distances.

9.3 Fabric MetroCluster ISL Considerations

The supported ISL speeds are 1Gbps, 2Gbps, 4Gbps, 8Gbps, and 16Gbps. The ISL connection type can be either:

- **Native FCP over dark fiber** (dedicated fiber cabling; no protocol)

OR

- **xWDM/TDM devices:**
  - Only xWDM/TDM devices supported by fabric switch vendor
  - No specifications for the physical cable between sites; however, they must:
    - Have dedicated wavelength (lambda)
    - Be supported by the fabric switch vendor for the distance, switch type, and OS version
  - Use the switch LS setting to configure distance on the fabric switches (do not use the LD setting)

The following are not supported:

- Non-FC native framing/signaling
- Metro-E or TDM (SONET/SDH)
- FCR (native FC routing) or FCIP extensions

ISLs of different lengths for different fabrics are supported provided each ISL conforms to the preceding considerations. ISLs on an individual fabric must be the same length.

Achieving metro distances requires specialized long-wave SFPs:

- **If dedicated fiber cabling is being used**. Long-wave SFPs for the fabric switches are necessary.
- **If active xWDM/TDM devices are being used between sites**. The required SFPs should match the cabling (single-mode cabling requires long-wave SFPs) between switches and xWDM/TDM gear, as well as the type of SFPs installed in the xDWDM gear.

See the Interoperability Matrix on the NetApp Support site for more information on SFPs. **Note**: Currently NetApp does not source long-wave SFPs capable of greater than 30km distances. To achieve distances that are greater than 30km, external sourcing is necessary. Make sure the SFPs are supported by the fabric switch model being used.

9.4 Fabric MetroCluster Latency Considerations

A dedicated fiber link has a round-trip time (RTT) of approximately 1ms for every 100km (~60 miles). Additional nonsignificant latency might be introduced by devices (for example, multiplexers) en route.

Generally speaking, as the distance between sites increases (assuming 100km = 1ms link latency):

- Storage response time increases by the link latency. For example, if storage has a response time of 1.5ms for local access, then over 100km the response time increases by 1ms to 2.5ms.

 © 2014 NetApp, Inc. All Rights Reserved.

- Applications, in contrast, respond differently to the increase in storage response time. Some applications' response time increases by approximately link latency, while other applications' response time increases by greater than link latency. For example, applicationA response time with local storage access might be 5ms and over 100km is 6ms; ApplicationB response time with local storage access might also be 5ms, but over 100km is 10ms. Be cognizant of application behavior with latency-sensitive applications.

9.5 Fabric MetroCluster Using DS14FC Shelves

Prior to Data ONTAP 8.1, fabric MetroCluster supported only DS14 FC shelves. In Data ONTAP 8.1 and later, DS14 shelves are supported with Brocade switches only. When using these shelves:

- Every DS14 FC disk must log into the fabric in FC loop mode.
- There is a maximum of two DS14 FC shelves per disk loop (28 disk logins per switch port).

There is an imposed DS14 FC spindle limit:

- Data ONTAP 7.2.4 to Data ONTAP 7.3.2: lesser of platform limit or 504
- Data ONTAP 7.3.2 to Data ONTAP 7.3.4: lesser of platform limit or 672
- Data ONTAP 7.3.5 or later: lesser of platform limit or 840

As of May 2012, DS14 FC shelves are available as add-on only. They are no longer configurable.

Check with the DS14 product team to confirm future availability of DS14 FC shelves.

9.6 Fabric MetroCluster Using SAS Shelves

From Data ONTAP 8.1 onward, fabric MetroCluster supports SAS shelves using the SAS-to-FC FibreBridge 6500N.

Two bridges are required per stack of SAS shelves (for redundancy and performance), up to a maximum of 10 SAS shelves per stack or 4 SAS shelves if SSD drives are present (see section 9.7 for more information). At a minimum, fabric MetroCluster requires four bridges, two per stack, with one stack on either site.

The platform spindle limits apply to SAS-based fabric MetroCluster configurations. Because each controller sees all storage, the platform spindle limit applies to the entire configuration. For instance, if the spindle limit for FAS32xx is *n*, then, although there are two controllers, the spindle limit for a FAS32xx fabric MetroCluster configuration remains *n*.

 © 2014 NetApp, Inc. All Rights Reserved.

The following table provides a comparison between DS14 FC and SAS-based fabric MetroCluster configurations.

Table 3) FC and SAS MetroCluster configuration.

	DS14 Fabric MetroCluster	SAS Fabric MetroCluster
Storage	DS14 FC only (FC disks required) DS14 FC availability post May 2012: Add-on only Note: Check with the DS14 storage team for the latest information on DS14 shelf availability No support for DS14 Flash Pool™	DS4243, DS4246, or DS2246 using the FibreBridge 2 FibreBridges per stack Support for Flash Pool starting in Data ONTAP 8.1.1
Loop/stack depth	Max: 2 shelves per loop	Max: 10 shelves per stack, except if SSD drives are present, in which case the limit is 4 shelves with no more than 2 shelves (48 drives) of SSD.
Connections	Controllers (FC initiators, FC-VI) connect to switches Shelves connect to switches (loop mode)	Controllers (FC initiators, FCVI) connect to switches FibreBridges connect to switches (point-to-point mode) Shelves connect to FibreBridges
Distances	Max: 100km See the Interoperability Matrix on the NetApp Support site for the latest information	Max: 200km See the Interoperability Matrix on the NetApp Support site for the latest information
Switch support	Dedicated Brocade switches supplied by NetApp	Dedicated Brocade or Cisco switches supplied by NetApp
See the Interoperability Matrix on the NetApp Support site for the latest information		
Shared Brocade 5100 and 6510 switches, Cisco 9710.	Not allowed	Brocade 5100 or 6510 switches can be shared between 2 MetroCluster configurations Cisco 9710 can be shared between MetroCluster configurations

 © 2014 NetApp, Inc. All Rights Reserved.

9.7 SAS, SATA, and SSD Mixed Shelf Configuration Considerations

In general, the same disk support and mixing guidelines for shelves in a stack apply to MetroCluster as for configurations other than MetroCluster. DS4243, DS4246, and DS2246 shelves may be mixed in a stack provided they comply with the specifications in the section "Shelf Intermixing" in the Storage Subsystem Technical FAQ on the Field Portal. Refer to this document for more information. All-SSD aggregates are supported in Data ONTAP 8.1.3 and 8.2.1. The guidelines are the same regardless of whether SAS cables or the FibreBridge are used to attach the storage, with the following exceptions:

- A maximum of 10 shelves per stack can be configured, except for FibreBridge attached, which includes SSD. In that case, the stack limit is 4 shelves.
- A maximum of 2 SSD shelves (48 drives) can be present in any individual stack using FibreBridges.
- DS2246, DS4246, and DS4243 can be mixed in the same stack with FibreBridge or SAS copper attachment. Only one transition between IOM3 and IOM6 per stack is allowed.
- IOM3 (DS4243) shelves are **not supported** with optical SAS attachment. Only DS2246 and DS4246 shelves are supported, with only one transition allowed.

9.8 Fabric MetroCluster Using Shared Switches

Shared switches are supported in configurations including twin MetroCluster and a pair of dual-chassis MetroCluster configurations. Consider sharing when each MetroCluster configuration consumes less than 50% of the switch resources and make sure sufficient ports are available on the switches. SAS-based storage only is supported.

In Data ONTAP 8.1 and later, four Brocade 5100 or 6510 switches can be shared between two independent MetroCluster pairs, with ISL sharing - two ISLs are required between each pair of switches. In Data ONTAP 8.2.1, the Cisco 9710 also supports switch sharing for up to six MetroCluster configurations on each line card. Dedicated ISLs are required for each MetroCluster configuration.

Refer to the Interoperability Matrix for the latest support information on controllers and storage, and refer to the "Fabric MetroCluster Brocade Switch Configuration Guide" and the "Fabric MetroCluster Cisco Switch Configuration Guide" on the NetApp Support site for information on setting up shared switches.

9.9 Fabric MetroCluster Using DS14FC and SAS Shelves (Mixed Shelf Configurations)

DS14 FC and SAS shelves can coexist in a fabric MetroCluster configuration. The SAS shelves connect into the fabric switches using the SAS-to-FC FibreBridge; the DS14 FC shelves connect directly to the fabric switches.

Spindle limit considerations:

- Fabric MetroCluster imposed spindle limits DO NOT exist for SAS disks.
- Fabric MetroCluster imposed spindle limits DO exist for DS14 FC disks.
- Total spindles (DS14FC plus SAS) should not exceed the platform limit. For example, the FAS6240 platform limit is 1,440; DS14 FC drives should not exceed 840 (imposed spindle limit); and the total (DS14 FC plus SAS) should not exceed 1,440 (platform limit).

Other considerations:

- Expand an existing DS14 FC fabric MetroCluster configuration by adding stacks of SAS shelves using the FibreBridge.
- Existing DS14 FC aggregates cannot be expanded using SATA disks.
- Aggregate mirroring must be between the same compatible drives (see the following caveat).

Disk drive mixing considerations:

 © 2014 NetApp, Inc. All Rights Reserved.

NetApp recommends NOT mixing "old and new" compatible drive technologies. In the context of fabric MetroCluster, this implies that:

- NetApp does not recommend creating an aggregate with a mix of DS14MK4 FC disks and SAS disks, but it is supported.
- NetApp does not recommend expanding existing DS14 MK4 FC aggregates with SAS disks, but it is supported.
- NetApp does not recommend creating SyncMirror relationships between DS14 MK4 FC disks and SAS disks, but it is supported.

Be aware of the fact that mixing FC and SAS disks will cause change management issues in the future when FC disks reach their end of support.

10 MetroCluster Sizing and Performance

Performance considerations for MetroCluster include the following areas:

- Round-trip time (RTT) between sites
- Aggregate mirroring (RAID SyncMirror) performance
- Takeover and giveback times
- FibreBridge 6500 performance
- Flash Pool
- ISLs
- Speed of FC-VI link
- Unidirectional port mirroring
- ISL encryption

10.1 Round-Trip Time Between Sites

Dedicated fiber link has an RTT of 1ms for every 100km (~60 miles). Additional latency might be introduced by devices (for example, multiplexers) along the way.

Generally speaking, applications running on MetroCluster have insignificant increases in response time. However, as distances increase between sites, some applications show proportionally greater increases in response time. See section 9.4 for more information.

Note: The maximum supported distance between sites for a fabric MetroCluster configuration is 200km using SAS-based storage with the FibreBridge.

10.2 Aggregate Mirroring (SyncMirror) Performance

In stretch MetroCluster where there is no significant distance between sites, RAID SyncMirror has an insignificant write performance impact, and read performance can potentially increase by 20% to 50%, provided reads are enabled from both plexes.

In fabric MetroCluster, write performance decreases by 6% to 10%. Over shorter distances, read performance increases by 20% to 50%, provided reads are enabled from both plexes. As distance increases, the gain in read performance decreases.

10.3 Takeover and Giveback Times

- CFO functionality protecting against controller failures: Takeover/giveback times range from 30 to 120 seconds, depending on controller load conditions.
- CFOD functionality protecting against complete site disasters: Takeover/giveback times range from 30 to 120 seconds, depending on controller load conditions.
- On average, takeover/giveback times of 60 seconds are observed.

 © 2014 NetApp, Inc. All Rights Reserved.

10.4 FibreBridge 6500N Performance

The FibreBridge can achieve significant IOPS and throughput without significant latency. To demonstrate this, a number of performance tests are documented here.

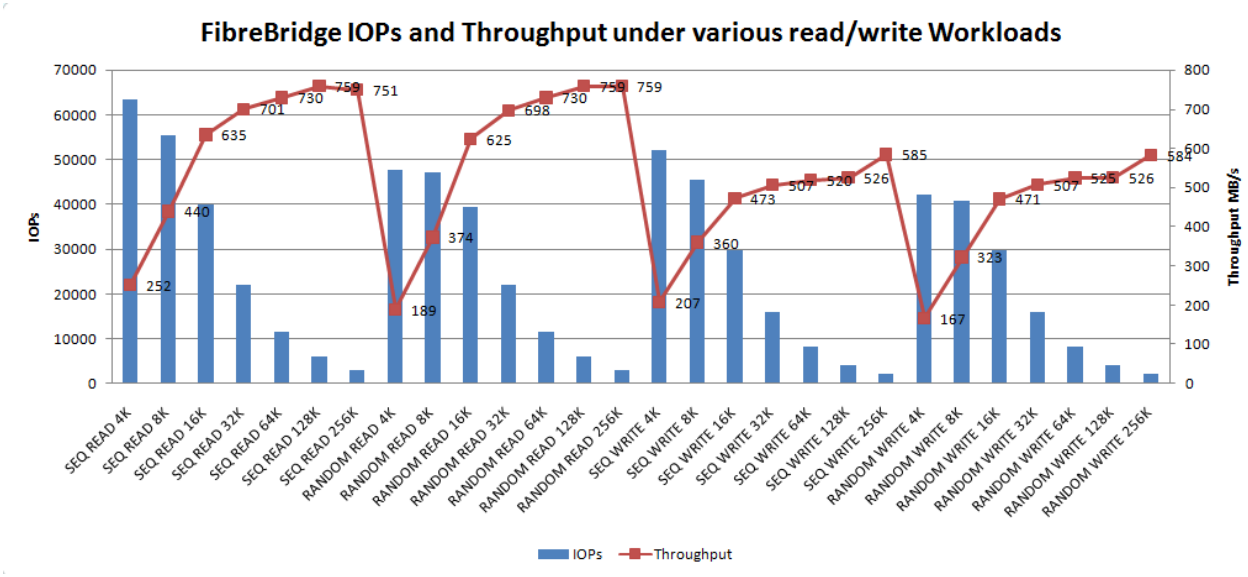
IOPS and Throughput

Using one FC port, the FibreBridge can achieve a maximum of:

- 63,000 IOPS at 4KB reads
- 53,000 IOPS at 4KB writes

Figure 20 details the IOPS and throughput numbers for a single FibreBridge under various load conditions. Because in MetroCluster two FibreBridges are used for each stack of shelves, the net available throughput (on average) is twice the following numbers.

Figure 20) FibreBridge IOPS and throughput under various read/write workloads.



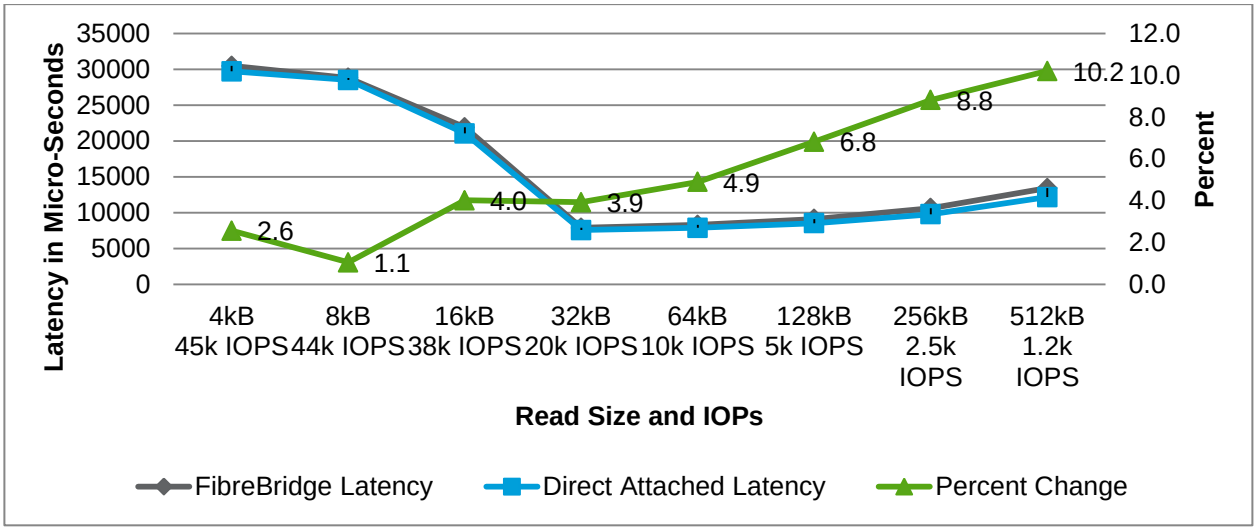
 © 2014 NetApp, Inc. All Rights Reserved.

Latency with SAS disks

In a test with 9 DS2246 shelves with 216 SAS disks (10k rpm), the FibreBridge introduced minimal latency compared to direct-attached storage.

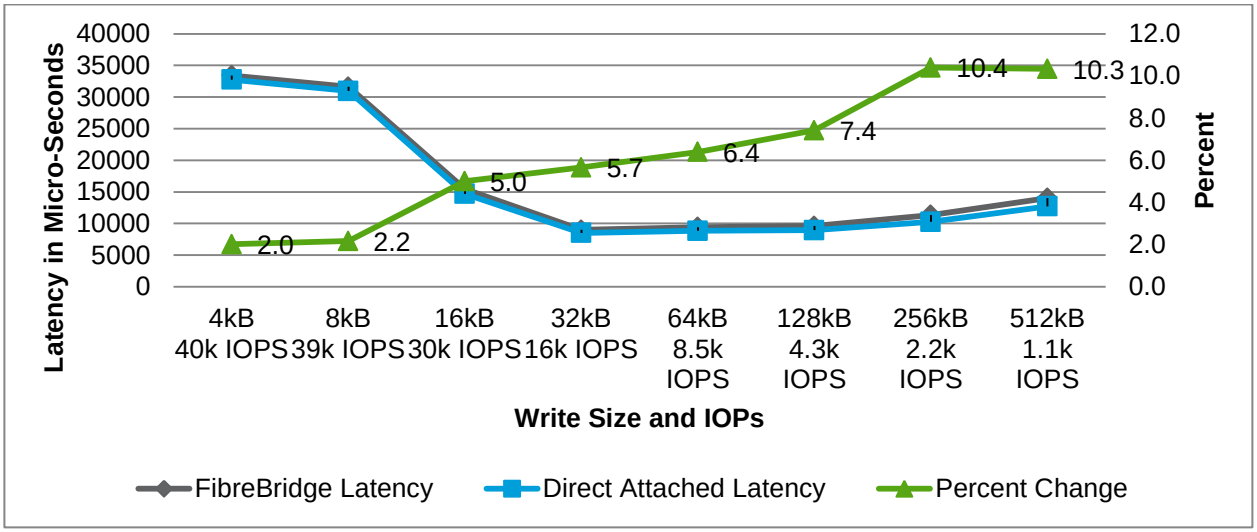
Read Latency

Figure 21) Read latency: FibreBridge compared to direct attached.



Write Latency

Figure 22) Write latency: FibreBridge compared to direct attached.



Conclusion: FibreBridge does not introduce significant latency compared to direct-attached SAS disks. The same conclusion can be extrapolated for SATA disks.

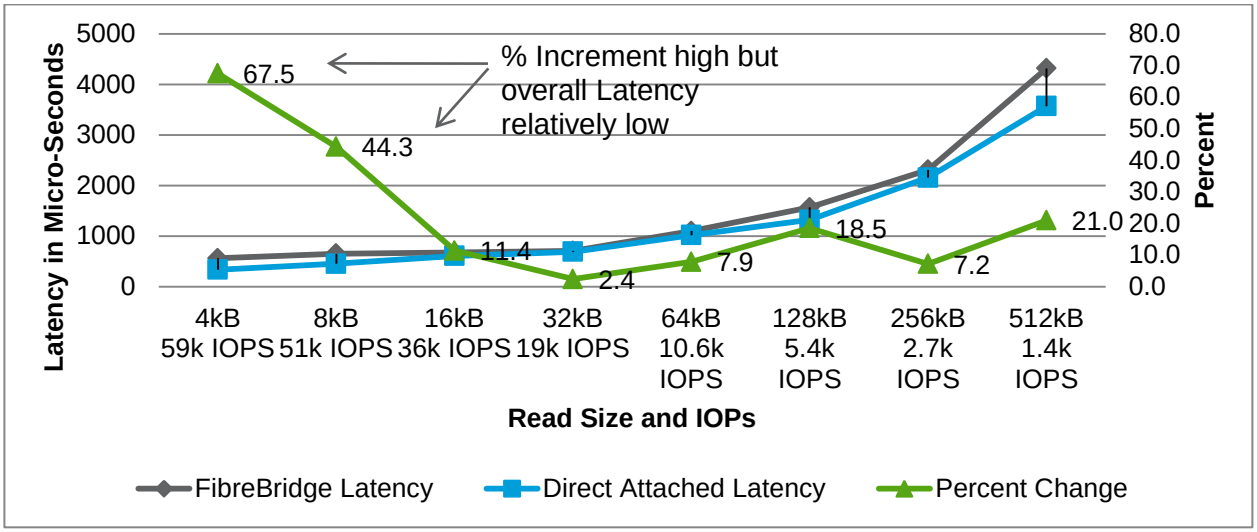
 © 2014 NetApp, Inc. All Rights Reserved.

Latency with SSD

To push FibreBridge limits, tests were run with 10 solid-state drives (SSDs). Again, the FibreBridge introduced minimal latency when compared to direct-attached storage.

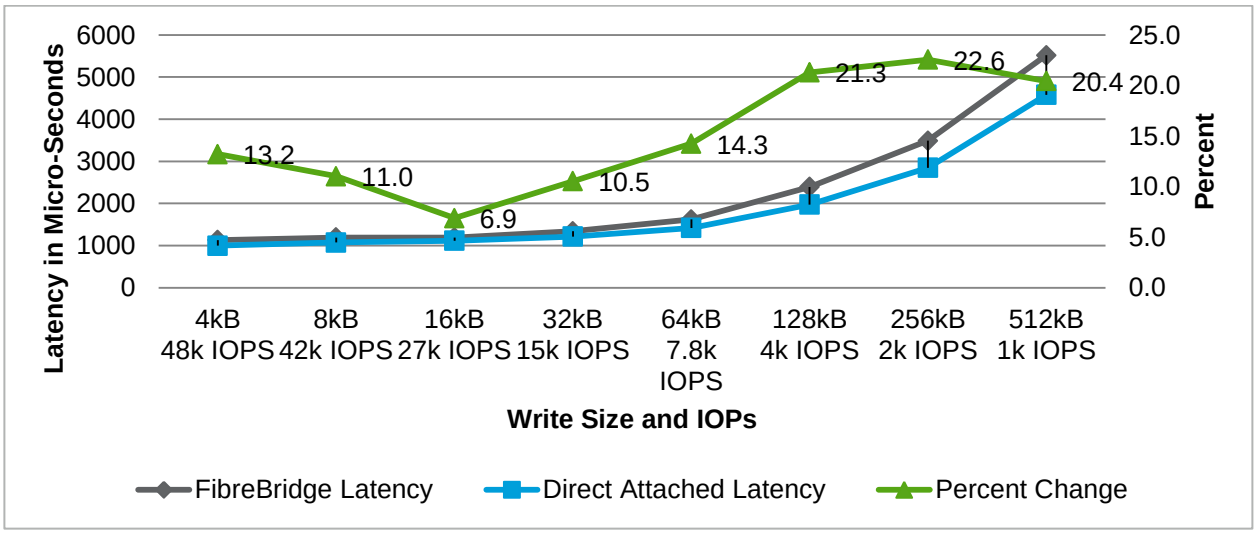
Read Latency

Figure 23) Read latency: FibreBridge compared to direct attached.



Write Latency

Figure 24) Write latency: FibreBridge compared to direct attached.



Conclusion: Though % latency increments are high at certain points, actual latency numbers are low. There are considerable benefits to using SSDs (over SAS disks). Flash Pool support for SAS-based MetroCluster configurations starts in Data ONTAP 8.1.1.

10.5 Flash Pool Performance

Starting in Data ONTAP 8.1.1, MetroCluster supports Flash Pool. Flash Pool allows flash technology in the form of SSDs to be combined with traditional hard disk drives (HDDs) in a single Data ONTAP

 © 2014 NetApp, Inc. All Rights Reserved.

aggregate. When SSD and HDD technologies are combined in a Data ONTAP aggregate, the NetApp storage system takes advantage of the latency and throughput benefits of the SSD while maintaining the mass storage capacity of HDD.

In general, Flash Pool behaves no differently when used in a MetroCluster configuration. However, because MetroCluster requires the FibreBridge to support SAS storage, we must be cognizant of FibreBridge limits (as stated in the previous section), which can potentially cap Flash Pool performance.

For example, with 32k random reads, a FibreBridge maxes out at approximately 700MB/sec throughput. Because FibreBridges per stack of shelves are used, there is a theoretical maximum throughput of 1400MB/sec. Assume that each SSD data disk increases throughput by 100MB/sec; with a 32k random read workload we gain no benefit from more than 14 SSD data disks.

Follow the storage recommended best practices when sizing and creating Flash Pool. For more information on Flash Pool, see TR-4070: NetApp Flash Pool Design and Implementation Guide.

10.6 Sizing MetroCluster Interswitch Links (ISLs)

During normal operation, write traffic traverses the ISLs to mirror to the remote plex and for NVRAM mirroring. Thus, sizing ISLs is a function of write load. Differences exist depending on whether there are one or two ISLs per fabric. The following table presents a methodology for theoretical calculations. When using this method, be conservative and maintain a bandwidth cushion when sizing. Estimate the peak write load and use this value in the calculations that follow.

Table 4) ISL sizing guidelines.

One ISL per Fabric (Two ISLs Total)	
Sizing Guidelines for *X*MB/Sec Write Load (Use Peak Load)	
Note: Using one ISL per fabric, theoretical max write throughput is 2/3 ISL bandwidth.	
Storage traffic	*X*MB/sec
Cluster interconnect traffic	*X*MB/sec
Total traffic traversing ISL	$X + X = 2X$MB/sec
Theoretical ISL bandwidth	• $(X + X)/(2/3) = 3X$ • Divide by 2/3 based on theoretical max write throughput of 2/3 of ISL bandwidth • Note that this is theoretical; actual will be less • Convert to Gbps 3X * 8/1000
ISL bandwidth choices	Conservatively choose ISL bandwidth rounded up from the previous step; choices are 1, 2, 4, 8Gbps

 © 2014 NetApp, Inc. All Rights Reserved.

One ISL per Fabric (Two ISLs Total)
Example: With a 100MB/sec peak write load: • Storage traffic = 100MB/sec • Cluster interconnect traffic = 100MB/sec • Theoretical ISL bandwidth = (100 plus 100)/(2/3), giving us 300*8/1000 = 2.4Gbps • Choices are 1, 2, 4, 8, 16Gbps • Choose 4Gbps ISL bandwidth

Two ISLs per Fabric (Four ISLs Total)	
Sizing Guidelines for *X*MB/Sec Write Load (Use Peak Load)	
Note: Using two ISLs per fabric, theoretical max write throughput is 100% ISL bandwidth.	
Storage traffic	*X*MB/sec
Cluster interconnect traffic	*X*MB/sec
Total traffic traversing storage ISL	*X*MB/sec
Total traffic traversing cluster interconnect ISL	*X*MB/sec
Theoretical storage ISL bandwidth	*X* * 8Mbps
Theoretical cluster interconnect ISL bandwidth	*X* * 8Mbps
ISL bandwidth choices	Conservatively choose ISL bandwidth rounded up from the previous step; choices are 1, 2, 4, 8Gbps
Example: 100MB/sec peak write load: • Storage traffic = 100MB/sec • Cluster interconnect traffic = 100MB/sec • Theoretical storage ISL bandwidth = 100*8 = 800Mbps • Theoretical cluster interconnect bandwidth = 100*8 = 800Mbps • Choices are 1, 2, 4, 8, 16Gbps • Choose 2Gbps for storage ISL (instead of 1Gbps for bandwidth cushion) • Choose 2Gbps for cluster interconnect ISL bandwidth (instead of 1Gbps for bandwidth cushion)	

 © 2014 NetApp, Inc. All Rights Reserved.

10.7 FC-VI Speed

Data ONTAP 8.2.1 introduces support for the 16Gb FC-VI card for FAS62xx, FAS3220, FAS3250, and FAS80xx platforms. This offers substantially higher throughput for improved performance. Testing on FAS80xx platforms demonstrated from 20% to 50% higher throughput and 30% to 50% lower latency on a variety of workloads, in comparing performance of the 8Gb and 16Gb FC-VI cards.

Refer to the IMT for supported platforms for the 16Gb FC-VI card.

Supported controllers running Data ONTAP 8.2.1 can upgrade from the 8Gb to the 16Gb FC-VI card; however the process is disruptive – both controllers in the MetroCluster must be shut down. See KB article 1014505 for the detailed procedure.

10.8 Unidirectional Port Mirroring

Data ONTAP 8.2.1 introduces support for Unidirectional Port Mirroring (UDPM). Before Data ONTAP 8.2.1, NVRAM mirroring uses only one FC-VI port for traffic in both directions (incoming and outgoing). If the port or link for NVRAM mirroring being used goes down, the traffic will fail over to the other port. With UDPM enabled, both FC-VI ports are used: one for incoming and one for outgoing traffic on each controller when client or host traffic is active on both MetroCluster controllers. Selection of the interconnect link to be used by incoming and outgoing traffic is based on the system serial number of the nodes. The node with the higher serial number uses FC-VI link 0 for outgoing data traffic and FC-VI link 1 for incoming data traffic. The partner node will therefore use FC-VI link 0 for incoming data traffic and FC-VI link 1 for outgoing data traffic.

Testing on FAS80xx platforms demonstrated from 7% to 30% higher throughput and 13% to 28% lower latency on a variety of workloads, in comparing performance with UDPM enabled, compared to UDPM disabled.

Unidirectional Port Mirroring is automatically enabled when both controllers in the MetroCluster configuration are running Data ONTAP 8.2.1, provided that the preferred primary port is disabled on both the nodes. UDPM can be enabled or disabled by CLI. Enabling UDPM similarly requires disabling the preferred primary port. Both nodes of the MetroCluster configuration should have the same setting for UDPM, either enabled or disabled. It is recommended to maintain the default enablement of UDPM after upgrading to Data ONTAP 8.2.1. Refer to the High Availability and MetroCluster Configuration Guide for more information on querying the current status and enabling UDPM.

10.9 ISL Encryption

Data ONTAP 8.2.1 introduces support for ISL encryption with the Brocade 6510 switch only. Encryption does impose an overhead to performance. Testing on FAS80xx platforms demonstrated between 10 and 30% decrease in throughput for sequential workloads with up to double the latency (depending on controller model). For random workloads, the effect was far less (less than 5% drop in throughput and less than 8% increased latency).

11 Configuring MetroCluster

Refer to the product documentation on the NetApp Support site for detailed configuration steps. The relevant publications are:

- High-Availability and MetroCluster Configuration Guide
- Fabric MetroCluster Brocade and Cisco Switch Configuration Guide (for fabric MetroCluster switches)
- Atto FibreBridge Configuration Guide (for SAS-based MetroCluster FibreBridges)

This section presents checklists and high-level steps to provide guidance on how best to leverage the product documentation. It is not meant to replace or supersede the product documentation in any way.

 © 2014 NetApp, Inc. All Rights Reserved.

11.1 Stretch MetroCluster Configuration

Make sure that you have the correct hardware and software as described in section 2. The following steps, which should be performed in order, summarize the configuration process and must be performed on the controllers in both sites. MetroCluster works as an active-active configuration with both controllers actively serving data. More information on these steps is available in the High-Availability and MetroCluster Configuration Guide.

1. Hardware setup:

   Location and distance: When installing stretch MetroCluster hardware:

   - Controllers and shelves are in geographically separated locations:
     - SiteA: ControllerA, ControllerA_Local_Storage, ControllerB_Mirrored_Storage
     - SiteB: ControllerB, ControllerB_Local_Storage, ControllerA_Mirrored_Storage
   - Distance including patch panels must not exceed stretch MetroCluster limits.

   Cabling: All storage must attach to both controllers such that each controller sees all storage. Use cables of the required length for cross-site connections. Note that NetApp does not source long fiber optic cables.

2. License installation:

   In Data ONTAP 8.2 and later, set the following options. Features specific to MetroCluster are enabled during configuration as described in the High-Availability and MetroCluster Configuration Guide.

   The following Data ONTAP options must be enabled on both nodes:

   - *cf.mode:* You must set this option to ha.
   - *cf.remote_syncmirror.enable:* Set the option to on.

   In releases prior to 8.2, install the MetroCluster licenses on both controllers:

   - syncmirror_local license
   - cluster license
   - cluster_remote license

3. Pools, plexes, and aggregates:

   a. Make sure the pool assignments are correct. Each controller should have pool0 (local plex) on local storage and pool1 (remote plex) on geographically separated remote storage.

      Disk assignment depends on whether dedicated or mixed pools are used.

      Dedicated pool configuration (when disks in a stack/loop are assigned to only one pool): All disks on the same stack/loop are assigned to the same system and pool. All stacks/loops connected to the same adapter are assigned to the same pool.

      Mixed pool configuration (when disks in the same stack/loop are shared between two pools): Typical mixed pool configurations include a single stack with both the local controller's pool0 on one or more shelves and the remote controller's pool1 on another shelf/shelves. In this situation run the `options disk.auto_assign off` command on both controllers to disable disk autoassignment. This prevents the disk drives from being automatically assigned to the same controller. The `disk assign` command should be used to assign disks to the correct owners (local or remote controller).

   b. Verify connections: Confirm that the disks are visible and have dual paths.

   c. Configure network: Historically, network interface configuration for MetroCluster configurations is no different from that for standard HA pairs. However, effective with Data ONTAP 7.3.2, each controller of the MetroCluster configuration may reside on a different subnet. To configure this, do the following:

      - Set the option.cf.takeover.use_mcrc_file to on.
      - Configure the /etc/mcrc file on each controller.

 © 2014 NetApp, Inc. All Rights Reserved.

d. Create aggregates and mirrors: SyncMirror is used to create aggregate mirrors. When planning mirrors, keep in mind:

- Aggregate mirrors should be on the remote site (geographically separated).
- In normal mode (no takeover), aggregate mirrors cannot be served.
- Aggregate mirrors can exist only between like or compatible drive types.

When the SyncMirror license is installed (or option enabled), disks are divided into pools (pool0 local, pool1 remote/mirror). When a mirror is created, Data ONTAP pulls disks from pool0 for the local aggregate and from pool1 for the mirrored aggregate.

Verify the correct number of disks are in each pool before creating the aggregates. Any of the following commands can be used:

- `sysconfig -r`
- `aggr status -r`
- `vol status -r`

Create a mirrored aggregate: After the pools have been verified, create new mirrored aggregates with the –m flag. For example, `aggr create aggrA -m 10` creates a mirrored aggregate called aggrA with ten drives (five in plex0, five in plex 1).

To mirror an existing aggregate, use the aggr mirror command.

Note: The root aggregate must be mirrored for the CFOD functionality to work after a site disaster.

4. Test failure scenarios:

Before putting the configuration into production, perform the following tests to verify operation:

a. Tests for hardware component redundancy:

- Disconnect cables:
  - Controller to shelf
  - Controller to controller
- Power down
  - Shelf
  - Controller

Note: Disconnecting or powering down two "like components" leads to data loss (example: powering down a local shelf and then powering down a remote shelf before bringing the local shelf up) because the redundancy is voided by the dual failure.

b. Test failover: Verify failover and giveback functionality:

- On each controller (one controller per test):
  - Enter the `cluster failover` command (`cf takeover`).
  - Enter the `cluster giveback` command (`cf giveback`) once the partner controller is waiting for giveback.

Test force takeover: Verify force takeover and site recovery functionality:

- On each controller (one controller per test):
  - Shut down one controller and all attached shelves on site.
  - Enter `cluster forced takeover` command (`cf forcetakeover -d`).
  - Follow the giveback process in section 12, "Site Failover and Recovery."

11.2 Fabric MetroCluster Configuration

Make sure that you have the correct hardware and software as described in section 2. The following steps, which should be performed in order, summarize the configuration process and must be performed on the controllers in both sites. MetroCluster works as an active-active configuration with both controllers actively serving data. More information on these steps is available in the High-Availability and MetroCluster Configuration Guide, “Fabric MetroCluster Brocade and Cisco Switch Configuration Guide,” and “Atto FibreBridge Configuration Guide” on the NetApp Support site.

1. Hardware setup:

 © 2014 NetApp, Inc. All Rights Reserved.

Location: When installing fabric MetroCluster hardware, each set of controllers and shelves are in geographically separated locations:

- SiteA: ControllerA, ControllerA_Local_Storage, ControllerB_Mirrored_Storage
- SiteB: ControllerB, ControllerB_Local_Storage, ControllerA_Mirrored_Storage

Distance: The distance between controllers must not exceed fabric MetroCluster limits based on the switches and switch speed selected; see the Interoperability Matrix.

Note: The maximum distance supported between sites is 200km if using SAS-based storage.

Cabling: Direct-attached storage is not supported in fabric MetroCluster configurations. Both controllers and all storage (or FibreBridge when using SAS storage) must connect to the fabric switches.

- Controller-to-switch cabling:
  - FC initiator ports: Four FC initiator ports are required on the controller. Two FC initiator ports connect to each switch for redundancy and sufficient bandwidth.
  - Cluster interconnect ports: There is one connection per switch from the FC-VI card.
- Shelf-to-switch cabling for FibreBridge-based SAS configurations:
  - Each stack can contain up to 10 shelves except as documented in section 9.7.
  - Two FibreBridges are required per stack.
  - Each FibreBridge connects to a single switch.
  - There are no spindle limits imposed by fabric MetroCluster, just the platform spindle limit.
- Shelf-to-switch cabling for DS14FC configurations:
  - Each shelf loop connects to both switches.
  - Maximum of two shelves per loop: We recommend starting with one shelf per loop and, when all switch ports have been expended, adding shelves to the existing shelf loops.
  - Be cognizant of imposed spindle limits (see the fabric MetroCluster considerations section).
- Shelf-to-switch cabling for DS14FC and FibreBridge-based SAS mixed configurations:

  When expanding an existing DS14 FC fabric MetroCluster configuration with FibreBridge-based SAS storage, the FibreBridge-based SAS storage simply plugs into the existing environment. Be cognizant of DS14 FC spindle limits for the DS14FC portion on the environment.
- Switch-to-switch cabling: interswitch links (ISLs):

  If customers already own or buy dark fiber: Switches will connect to each other over the dark fiber links (using long-wave SFPs on the switch ports sourced either by NetApp or externally [NetApp does not source long-wave SFPs for distances >30km]).

  If customers lease metrowide transport services from a service provider: Switches connect into the DWDM/TDM devices (long-wave SFPs are not required if the DWDM/TDM devices come supplied with long-wave SFPs).

  Refer to section 9.3 for more information.

2. License installation:

   In Data ONTAP 8.2, set the following options. Features specific to MetroCluster are enabled during configuration as described in the High-Availability and MetroCluster Configuration Guide.

   The following Data ONTAP options must be enabled on both nodes:

   - *cf.mode:* You must set this option to ha.
   - *cf.remote_syncmirror.enable:* Set the option to on.

   In releases prior to 8.2, install the MetroCluster licenses on both controllers:

   - syncmirror_local license
   - cluster license
   - cluster_remote license

3. FibreBridge firmware:

 © 2014 NetApp, Inc. All Rights Reserved.

Download and install the correct firmware from the NetApp Support site when using the FibreBridge for SAS-based MetroCluster configurations.

4. Switch configuration:

   Refer to the Interoperability Matrix on the NetApp Support site to verify correct switch vendor firmware.

   Refer to “Fabric MetroCluster Brocade and Cisco Switch Configuration Guide” on the NetApp Support site for step-by-step information on how to configure switches for DS14FC-based fabric MetroCluster.

   Refer to “Fabric MetroCluster FibreBridge 6500N Setup Guide” on the NetApp Support site for step-by-step information on how to configure switches for SAS-based fabric MetroCluster.

5. Pools, plexes, and aggregates:

   a. Make sure the pool assignments are correct. Each controller should have pool0 (local/active plex) on local storage and pool1 (remote plex) on geographically separated remote storage.

      Disk assignment depends on whether dedicated or mixed pools are used.

      Dedicated pool configuration (when disks in a stack/loop are assigned to only one pool): All disks on the same stack/loop are assigned to the same system and pool. All stacks/loops connected to the same adapter are assigned to the same pool.

      Mixed pool configuration (when disks in a stack/loop are shared between two pools): A mixed pool configuration is where a single stack has both the local controller’s pool0 on one or more shelves and the remote controller’s pool1 on another shelf/shelves. In this situation run the `options disk.auto_assign off` command on both controllers to disable disk autoassignment. This prevents the disk drives from being automatically assigned to the same controller. The `disk assign` command should be used to assign disks to the correct owners (local or remote controller).

   b. Verify connections: Confirm that the disks are visible and have dual paths.

   c. Configure network: Historically, network interface configuration for MetroCluster configurations is no different from that for standard HA pairs. However, effective with Data ONTAP 7.3.2, each controller of the MetroCluster configuration may reside on a different subnet. To configure this, do the following:

      - Set the option.cf.takeover.use_mcrc_file to on.
      - Configure the /etc/mcrc file on each controller.

   d. Create aggregates and mirrors: SyncMirror is used to create aggregate mirrors. When planning mirrors, keep in mind:

      - Aggregate mirrors should be on the remote site (geographically separated).
      - In normal mode (no takeover), aggregate mirrors cannot be served.
      - Aggregate mirrors can exist only between like or compatible drive types.

      When the SyncMirror license is installed (or option enabled), disks are divided into pools (pool0 local, pool1 remote/mirror). When a mirror is created, Data ONTAP pulls disks from pool0 for the local aggregate and from pool1 for the mirrored aggregate.

      Verify the correct number of disks are in each pool before creating the aggregates. Any of the following commands can be used:

      - `sysconfig –r`
      - `aggr status –r`
      - `vol status –r`

      Creating a mirrored aggregate: After the pools have been verified, create new mirrored aggregates with the –m flag. For example, `aggr create aggrA –m 6` creates a mirrored aggregate called aggrA with six drives (three in plex0, three in plex 1).

      To mirror an existing aggregate, use the aggr mirror command.

      The root aggregate must be mirrored for the CFOD functionality to work on a site disaster.

 © 2014 NetApp, Inc. All Rights Reserved.

6. Test failure scenarios:

   Before putting the configuration into production, perform the following tests to verify operation:

   a. Tests for hardware component redundancy:
      - Disconnect cables or simply disable the appropriate switch ports:
        - Controller to switch
        - Shelf to switch
        - If using SAS storage, FibreBridge to switch
        - If using SAS storage, FibreBridge to shelf
      - Power down shelf, controller, switch, bridge

   Note: Disconnecting or powering down two "like components" leads to data loss (example: powering down a local shelf and then powering down a remote shelf before bringing the local shelf up) because the redundancy is voided by the dual failure.

   b. Test failover: Verify failover and giveback functionality:
      - On each controller (one controller per test):
        - Enter the `cluster failover` command (`cf takeover`).
        - Enter the `cluster giveback` command (`cf giveback`) once the partner controller is waiting for giveback.

      Test force takeover: Verify force takeover and site recovery functionality:
      - On each controller (one controller per test):
        - Shut down one controller and all attached shelves on site.
        - Enter `cluster forced takeover` command (`cf forcetakeover -d`).
        - Follow the giveback process in section 12, "Site Failover and Recovery."

12 Site Failover and Recovery

12.1 Site Failover

A site failover might be necessary for many reasons, including

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

Upon determining that one of the sites failed, the administrator must execute a specific command on the surviving node to initiate a site takeover: `cf forcetakeover -d`.

This process is manual to prevent a split-brain scenario, which occurs when connectivity is lost between sites but each site continues data-serving operations. In this case a forced takeover might not be desirable.
Remember, the process is manual only when dealing with a site disaster. All other failures are handled automatically (with normal HA takeover and local SyncMirror).

Note the difference between normal HA takeover and forced takeover initiated by MetroCluster on disaster:

- In an HA pair configuration other than MetroCluster, if a storage controller fails, the partner detects the failure and automatically 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 is not performed.
- In a MetroCluster configuration, manually executing a single command allows a takeover to occur in spite of the lack of a quorum of disks.

 © 2014 NetApp, Inc. All Rights Reserved.

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

- Volumes have a new file system ID (FSID) in order to avoid conflict with the original volumes (the original FSID can be preserved after Data ONTAP 7.2.4).
- 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 offline so that only the desired LUNs are brought online after the site failure. From Data ONTAP 7.2.4 onwards, 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 online automatically. This option is called `cf.takeover.change_fsid`. If set to off (0), the original FSID is preserved.

12.2 Split-Brain Scenario

The `cf forcetakeover -d` 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 without MetroCluster). After the problem at the failed site is resolved, the administrator must prevent the failed controller from trying to rejoin the HA pair. If the failed node does resume operation, a split-brain scenario might occur because the recovered controller does not know that the other site has taken over. If both controllers are trying to serve data, this could lead to data corruption.

To restrict access to the previously failed controller, do one or more of the following:

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

12.3 Recovery Process

Although a complete site failover can be performed with a single command, there are cases in which another step or two might 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 on:`

- **NFS volumes** must be remounted. For more information about mounting volumes, see the "File Access and Protocols Management Guide" on the NetApp Support site.
- **iSCSI and Fibre Channel LUNs** might need to be rescanned by the application if the application (for example, VMware®) relies on the LUN serial number. When a new FSID is assigned, the LUN serial number changes.

After a forced takeover, all LUNs that were served from the failed site are now served by the surviving site. However, each of these LUNs must be brought online.

For example: `lun online /vol/vol1/lun0 and lun online /vol/vol1/lun1`

 © 2014 NetApp, Inc. All Rights Reserved.

The reason they are offline 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 offline before the disaster. If the LUN that was previously offline came online first, the second LUN would not be accessible because two LUNs with the same ID mapped to the same host cannot be brought online.

These steps can be automated using scripting. The main challenge for the script is to understand the state of the LUNs before the site failure so that the script brings online only LUNs that were online prior to the site failure.

However, as mentioned, 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 online automatically. This option is called `cf.takeover.change_fsid`. If set to off (0), the original FSID is preserved.

12.4 Giveback Process

After all the conditions that caused the site failure have been resolved and it has been verified that the controller at the failed site is offline, it is time to prepare for the giveback so that the sites can return to their normal operation. Use the following procedure to resynchronize the mirrors and perform the giveback.

Command	Result
Power on the disk shelves and FC switches at the disaster site.	
`aggr status -r`	Validate that you can access the remote storage from the surviving node. If the remote shelves do not 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 command at the left. 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 online.
If aggregates at the disaster site are online, take them offline 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 offline.
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.
Boot the node at the disaster site to wait for giveback.	Confirm that the console of the node at the disaster site shows it is waiting for giveback state.
Enter the following command at the partner node: `cf giveback`	The node at the disaster site reboots.

 © 2014 NetApp, Inc. All Rights Reserved.

13 Monitoring MetroCluster and MetroCluster Tools

Currently MetroCluster has no direct integration into manageability tools such as OnCommand® Protection Manager and Operations Manager. However, MetroCluster does integrate with several other tools offering monitoring and management capability.

13.1 Fabric MetroCluster Data Collector (FMC_DC)

Fabric MetroCluster Data Collector is a Java® program with the capability to do thorough data collection and analysis for a MetroCluster configuration.

Overview:

- FMC_DC monitors an entry called a node: A node can be a complete fabric MetroCluster configuration or a component within a fabric MetroCluster configuration such as a controller or switch.
- Data collection is run on the nodes: hourly data collection for concise data and full data collection for comprehensive data.
- Collection schedules can be altered, and manual collections are possible.
- Each node has an independent collection schedule.
- Data, after being collected, is analyzed and checked for errors.
- Data types collected: switch, controller, the AutoSupport™ tool, and storage information.

Workflow:

1. Nodes are identified to be monitored by FMC_DC.
2. Data collection is performed based on the schedules set/altered.
3. Data collected is checked for errors (the program FMC_Check is used).
4. An analysis report is created and e-mailed if necessary.

MetroCluster monitor helps identify issues early on, and FMC_DC helps identify the cause.

Download: NetApp Support site toolchest
Support: Best-effort support from NetApp Support

13.2 MetroCluster Tie-Breaker

MetroCluster tie-breaker (MCTB) software helps differentiate between two scenarios:

- Connectivity between sites or between the two halves of a MetroCluster configuration is lost.
- Site disaster: The site housing one-half of the MetroCluster configuration goes down.

Without the tie-breaker, a MetroCluster configuration has no mechanism to differentiate between the preceding two scenarios. The tie-breaker behaves like a witness and sits at a third site monitoring the two halves of the MetroCluster configuration. The tie-breaker employs the following logic.

- If loss of connectivity occurs between sites, simply notify of the loss of the link between sites but take no action. The rationale is that after connectivity is restored, data changes are resynchronized between the two halves of the MetroCluster configuration.
- Site disaster: If a site disaster occurs, the tie-breaker initiates a forced takeover (CFOD) on the surviving site. The surviving site then serves the disaster-stricken site's data as well.

The end result is a third point of view to determine when a forced takeover is necessary.

The MetroCluster tie-breaker can be acquired in one of two ways:

 © 2014 NetApp, Inc. All Rights Reserved.

- PVR with the Rapid Response Engineering (RRE) team
- A NetApp representative

13.3 OnCommand Site Recovery

OnCommand Site Recovery (OCSR) software extends Windows Server® failover cluster services (WSFC) to orchestrate MetroCluster storage failover with application failover. OCSR works in conjunction with WSFC, relying on a Windows® cluster quorum to prevent split-brain scenarios.

OCSR overview:

- Extends existing WSFC
- Provides native unattended failover in the event of a failure
- Eliminates the need for complex scripting
- Provides maximum availability for business-critical services

OCSR works in conjunction with WSFC, relying on a Windows cluster quorum to prevent split-brain scenarios. After a quorum is established, OCSR checks the storage. If a disaster is detected, OCSR stops the service and then triggers a forced takeover (CFOD) to bring the application back online after failover completes.

For additional information and documentation, go to the Software section of the NetApp Support site.

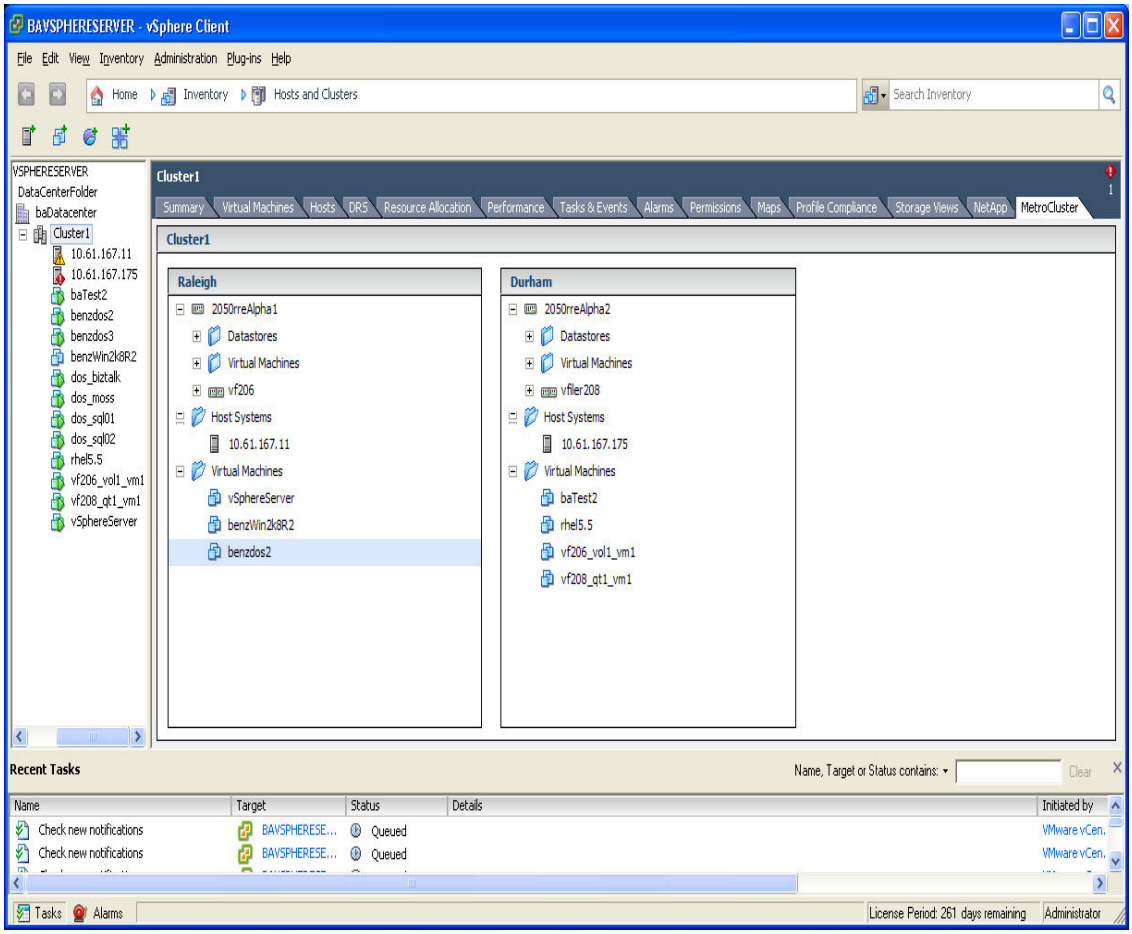
13.4 MetroCluster vCenter Easy Button

The MetroCluster vCenter™ easy button is a VMware vSphere® client plug-in that provides a MetroCluster tab in the vCenter client. It provides the following functionality:

- Gives a "site-centric" view of vSphere cluster inventory, including controllers, datastores, VMware ESX® and ESXi™ hosts, and virtual machines
- Can perform a controller takeover:
  - CFO (normal HA cluster failover)
  - CFOD (cluster failover on disaster: `cf forcetakeover -d`)
- Can perform a controller giveback after an HA takeover occurred
- Can evacuate virtual machines from one site to the other by placing the ESX or ESXi hosts in maintenance mode (requires VMware DRS)

The plug-in cannot automate the recovery process (rejoining of aggregates) after a site disaster; user intervention is required.

 © 2014 NetApp, Inc. All Rights Reserved.



The plug-in is available as a PVR with the Rapid Response Engineering team or through a NetApp representative.

14 Nondisruptive Operation

A number of maintenance operations can be performed nondisruptively in MetroCluster.

14.1 Nondisruptive Shelf Replacement (NDSR) for Fabric MetroCluster

Physical replacement of a disk shelf might become necessary. MetroCluster allows this task to be achieved nondisruptively. A prerequisite is that all disks on loops affected by a disk shelf must be mirrored.

NDSR for SAS shelves: NDSR for SAS shelves follows the same procedure as for DS14 shelves.

NDSR for DS14FC shelves: The following table summarizes the steps for NDSR for DS14FC; however, also refer to the “High-Availability and MetroCluster Configuration Guide” on the NetApp Support site. **Note**: For the purposes of this procedure, the two MetroCluster nodes are mc-nodeA and mc-nodeB. Aggregate and plex names are examples only.

 © 2014 NetApp, Inc. All Rights Reserved.

	Description	Commands
Note: For the purposes of this procedure, NetApp assumes that the disk shelf needing replacement is part of the loop connected to port 9 of the FC switches on mc-nodeB.		

1	Verify that all aggregates and volumes that own disks on the affected loop are mirrored and that the mirror is operational. Run the following commands to confirm mirror and operational state.	`mc-nodeB> aggr status`
		`mc-nodeB> sysconfig -r`
		`mc-nodeA> aggr status`
		`mc-nodeA> sysconfig -r`

2	Trigger an AutoSupport message from mc-nodeB and mc-nodeA indicating the start of the disk shelf replacement process.	`mc-nodeB> options autosupport.doit "SHELF REPLACE: START"` `mc-nodeA> options autosupport.doit "SHELF REPLACE: START"`
3	Run sysconfig on both mc-nodeB and mc-nodeA and save the output locally. Review the total number of disks seen on both NetApp controllers. This output will be used for comparison purposes at step 21.	`mc-nodeB> sysconfig` `mc-nodeA> sysconfig`
4	Inspect aggregate free space. If the total aggregate free space is less than the rate of change during shelf replacement, Snapshot™ autodelete might remove common SyncMirror aggregate Snapshot copies. If this occurs, rebaselining of mirrored data will be required.	
5	Offline aggregate plex "aggr1/plex0" on mc-nodeB.	`mc-nodeB> aggr offline aggr1/plex0`
6	On mc-nodeB_sw0 and mc-nodeB_sw2, log in as admin and disable switch port 9.	`mc-nodeB_sw0> portdisable 9` `mc-nodeB_sw2> portdisable 9`

7	Wait until all disks missing notifications on mc-nodeB and mc-nodeA complete. Verify this by running "`sysconfig`" and `sysconfig -a` to confirm that the disks and shelves are no longer visible.	`mc-nodeB> sysconfig`
		`mc-nodeB> sysconfig -a`
		`mc-nodeA> sysconfig`
		`mc-nodeA> sysconfig -a`

8	Power off the disk shelf connected to port 9 of mc-nodeB_sw0 and mc-nodeB_sw2.	
9	Remove disks from the shelf and keep them in a safe place where they will not accidentally drop off or experience any other mechanical impact due to replacement activity.	
10	Disconnect all FC, SFP, or SAS cables from the disk shelf connected to loop/stack 9.	
11	Remove the disk shelf connected to loop 9 from the rack cabinet.	

 © 2014 NetApp, Inc. All Rights Reserved.

	Description	Commands
12	Remove module A (top slot) from the disk shelf and insert into slot A (top slot) on the replacement shelf. (Replace any module hardware as needed.)	
13	Remove module B (bottom slot) from the disk shelf and insert into slot B (bottom slot) on the replacement shelf. (Replace any module hardware as needed.)	
14	Insert the replacement shelf into the rack cabinet. Set the module speed properly and make sure that the shelf ID is the same as the one that was replaced.	
15	Recable all connections. (Replace cables if needed.)	
16	Reconnect all SFP cables between the new disk shelf and the disk shelf of the same loop 9. (This step does not apply if switch port 9 has a single disk shelf loop.)	
17	Insert disks removed at step 9 into the replacement shelf **only** after the shelf is replaced (completely installed with requisite install kit) into rack cabinet.	
18	Power on the disk shelf on loop 9. Verify that the power comes up and no alarms are reported.	
19	On mc-nodeB_sw0 and mc-nodeB_sw2, enable switch port 9.	`mc-nodeB_sw0> portenable 9` `mc-nodeB_sw2> portenable 9`

20	Wait for Data ONTAP to finish recognizing all disks in the configuration. Check each aggregate and verify that all disks appear correctly in each offline plex. Also verify that the system recognizes all disks and shelves in the configuration.	`mc-nodeB> aggr status -r aggr1`
		`mc-nodeB> sysconfig -a`
		`mc-nodeB> sysconfig -r`

21	Run `sysconfig` on both mc-nodeB and mc-nodeA and save the output locally.	`mc-nodeB> sysconfig` `mc-nodeA> sysconfig`
22	Compare results to the sysconfig output from step 3 and confirm that the same number of disks appears under each FC host adapter listed as active and online before and after the shelf replacement. Correct any disk path issues that exist before continuing to step 23.	
23	Online aggregate plex “aggr1/plex0” on mc-nodeB.	`mc-nodeB> aggr online aggr1/plex0`
24	Wait for all aggr1 volumes on mc-nodeB to finish resyncing and return to a mirrored state. Depending on the change rate with the aggregate, this step can take minutes or hours.	
25	On mc-nodeB, turn Snapshot autodelete back on for each aggregate.	`mc-nodeB> aggr options aggr1 snapshot_autodelete on`

 © 2014 NetApp, Inc. All Rights Reserved.

	Description	Commands
26	Trigger AutoSupport from mc-nodeB and mc-nodeA, indicating the finish of the disk shelf replacement process.	`mc-nodeB> options autosupport.doit "SHELF REPLACE: FINISH"` `mc-nodeA> options autosupport.doit "SHELF REPLACE: FINISH"`

14.2 Nondisruptive Hardware Changes

For hardware upgrade procedures, refer to the nondisruptive hardware changes section of the "High-Availability and MetroCluster Configuration Guide" on the NetApp Support site or the platform-specific FRU replacement procedures in the hardware information library on the NetApp Support site.

You can use the nondisruptive upgrade method to perform the following hardware upgrade procedures:

- Replace the controller (with a controller of the same type and with the same adapters).
- Replace the motherboard.
- Replace or add an adapter: You can replace NVRAM, disk, or NIC components, either with the same component or with an improved component (for example, you can upgrade from 2-port to 4-port Gigabit Ethernet or 2-port to 4-port Fibre Channel).
- Replace the cluster interconnect adapter.
- Install on-board firmware on various platforms.

Note: Upgrading the FC-VI card from 8Gb to 16Gb is a disruptive procedure, as described in KB article 1014505.

14.3 Nondisruptive Operation for SAS Shelves

Refer to the Atto 6500N documentation on the NetApp Support site for FibreBridge-related nondisruptive hardware upgrade procedures.

The procedures that can be performed include:

- Hot-add a stack of SAS shelves to an existing MetroCluster configuration
- Hot-swap a FibreBridge
- Hot-add a SAS shelf to a stack of SAS shelves
- Hot-swap individual cables

15 MetroCluster Conversions

15.1 HA Pair to Stretch MetroCluster

An HA pair can be converted to a stretch MetroCluster configuration. The procedure requires downtime but is performed in place; no data migration is required. The following additional parts are needed.

HA Pair to Stretch MetroCluster	
Disk shelves	Additional shelves for mirroring.

 © 2014 NetApp, Inc. All Rights Reserved.

HA Pair to Stretch MetroCluster	
Cables for disk shelves	MPHA required for all new stretch MetroCluster configs. SAS copper or optical cables depending on implementation.
FibreBridges (when NOT using SAS copper or optical cables)	2 FibreBridges per stack of shelves. Consider the additional cables: shelf to FibreBridge (SAS cables) and FibreBridge to controller (FC cables).
Storage initiator HBAs	Depending on the exact configuration you might need to add HBAs. All currently shipping platforms have on-board initiator ports. Consult the "System Configuration Guide" on the NetApp Support site for specific models and support.
Chassis	If the HA pair consists of two controllers in the same chassis, a separate chassis is needed for one of the controllers to provide the physical separation.
FC-VI card	If the HA pair consists of two controllers in the same chassis, then the internal cluster interconnect can no longer be used. A separate FC-VI card must be installed in each controller as part of the upgrade process. This card, when installed, disables the internal backplane cluster interconnect.
Controller to disk shelf cables	Make sure you have the proper length, number, and type of Fibre Channel cables for interconnecting the disk paths and the two cluster interconnect ports between controllers. Also remember the connector types, especially if the connections between MetroCluster nodes will traverse through patch panels of any kind.
Copper to fiber adapters	For a stretch MetroCluster configuration using a non-31xx/32xx/62xx/80x0 platform, the existing cluster interconnect on the NVRAM card may be used. In order to accommodate the increased distances, the copper interface needs to be converted to fiber so that FC cabling may be used.
Comments: • **Disk and shelf firmware compatibility**: Inventory the current disk shelves to verify that firmware levels are compatible with the Interoperability Matrix on the NetApp Support site. • **Maximum spindle count = platform maximum**: The total maximum number of disk spindles for a FAS stretch MetroCluster configuration is equal to whatever the limit is for that platform. Refer to the "System Configuration Guide" on the NetApp Support site for the platform being used.	
Licenses	Assuming that this was an active-active or an HA configuration, the cluster license should already be installed on both systems. In that case only the syncmirror_local and cluster_remote licenses need to be ordered and installed. All three are required for a MetroCluster configuration. In Data ONTAP 7.3 and higher, the cluster license, syncmirror_local license, and cluster_remote license are part of the base software bundle.

 © 2014 NetApp, Inc. All Rights Reserved.

Procedure to convert from HA pair to stretch MetroCluster: Plan for downtime and follow the normal stretch MetroCluster setup steps in the “High-Availability and MetroCluster Configuration Guide” on the NetApp Support site.

15.2 HA Pair to Fabric MetroCluster

An HA pair can be converted to a fabric MetroCluster configuration. The procedure requires downtime but is performed in place; no data migration is required. The following additional parts are needed.

HA Pair to Fabric MetroCluster	
Disk shelves	Because DS14ATA disk shelves are not supported in fabric MetroCluster, additional DS14FC or SAS shelves might be necessary in order to migrate the data from the DS14ATA disk shelves. Also, additional shelves are needed for the aggregate mirrors.
FibreBridges	2 FibreBridges per stack of shelves. Consider the cabling required: disk shelf to FibreBridge (SAS cables), FibreBridge to switch (FC cables).
Fibre Channel switches	4 dedicated Fibre Channel (FC) switches are required for the fabric MetroCluster configuration, 2 per site. The switches must be supplied by NetApp. Verify that the switches and switch firmware chosen are compatible with the controllers and Data ONTAP version (see the Interoperability Matrix on the NetApp Support site).
Fibre Channel HBAs	Depending on the exact configuration, you might need to add FC HBAs. All of the currently shipping platforms have on-board FC ports. However, if these controllers also connect to a front-end SAN, then more ports might be necessary. Consult the “System Configuration Guide” on the NetApp Support site for specific models and support.
FC-VI card	A separate FC-VI card must be installed in each controller as part of the upgrade process. This card, when installed, disables the internal backplane cluster interconnect.
Fibre Channel interswitch links (ISLs)	These FC connections carry the cluster interconnect and disk traffic between nodes. Either 1 ISL per fabric (thus, a total of 2) or 2 ISLs per fabric (thus, a total of 4) are required. If 2 ISLs per fabric are used, then traffic isolation (TI) must be implemented with Brocade switches. The distance between the two locations will determine the type of cabling, the Small Form-Factor Pluggables (SFPs) needed for the ISLs, and possibly the switch model itself. Refer to section 9.3 and to the Interoperability Matrix on the NetApp Support site for more information.
Wave division multiplexers (customer supplied)	Because fiber cabling is expensive, customers often deploy multiplexing devices in order to use the fiber for multiple purposes. NetApp MetroCluster supports whatever device is used as long as it is listed on the switch vendor’s (Brocade’s or Cisco’s) compatibility guide for the model switch being used. This guide may be found at the switch vendor’s website.

 © 2014 NetApp, Inc. All Rights Reserved.

HA Pair to Fabric MetroCluster	
Cables	In a standard HA pair, many shelves can be connected in a physical disk loop or stack. In a DS14 FC fabric MetroCluster configuration there is a limit of 2 physical DS14FC disk shelves per loop, so accommodation might have to be made for additional loops and therefore additional Fibre Channel switch ports (1 loop [2 shelves] per port) and cables. SAS-based MetroCluster configurations have a limit of 10 physical SAS disk shelves per stack. Consider the cabling required for SAS-based fabric MetroCluster configurations: disk shelf to FibreBridge (SAS cables), FibreBridge to switch (FC cables).
Comments: • **Disk and shelf firmware compatibility:** Inventory the current disk shelves to verify that firmware levels are compatible with the Interoperability Matrix located on the NetApp Support site. If they are not, then the firmware must be upgraded as part of the process. • **Maximum spindle limits**: Be aware of the DS14FC spindle limits of a fabric MetroCluster configuration. See the Interoperability Matrix on the NetApp Support site for the latest information.	
Licenses	Verify the correct fabric MetroCluster licenses are installed as listed in “Data ONTAP Licenses” in section 2.1.
Disk ownership	If the current stretch MetroCluster configuration is based on the older FAS3020 or 3050, then the disk ownership model will have to be changed in order to upgrade to a fabric MetroCluster configuration. Make sure that these controllers are compatible with the Data ONTAP version and switches being used.
Fabric switch licenses	It is extremely important that the proper licenses for the switches be ordered and obtained prior to the start of the upgrade. See the section on fabric MetroCluster components for the proper licenses for the fabric switches.

Procedure to convert from HA pair to fabric MetroCluster configuration: Plan for downtime and follow the HA pair to fabric MetroCluster upgrade steps in the “High-Availability and MetroCluster Configuration Guide” on the NetApp Support site.

15.3 Stretch MetroCluster to Fabric MetroCluster

A stretch MetroCluster configuration can be converted to a fabric MetroCluster configuration. The procedure requires downtime but is performed in place; no data migration is required. The following additional parts are needed.

Stretch MetroCluster to Fabric MetroCluster	
Disk shelves	Because DS14ATA disk shelves are not supported in fabric MetroCluster, additional DS14FC or SAS shelves might be necessary in order to migrate the data from the DS14ATA disk shelves.
FibreBridges	2 FibreBridges per stack of shelves.

 © 2014 NetApp, Inc. All Rights Reserved.

Stretch MetroCluster to Fabric MetroCluster	
Fibre Channel switches	4 dedicated Fibre Channel (FC) switches are required for the fabric MetroCluster configuration, 2 per site. The switches must be supplied by NetApp. Verify that the switches and switch firmware chosen are compatible with the controllers and Data ONTAP version (see the Interoperability Matrix on the NetApp Support site).
Fibre Channel HBAs	Depending on the exact configuration, you might need to add FC HBAs. All of the currently shipping platforms have on-board FC ports. However, if these controllers also connect to a front-end SAN, then more ports might be necessary. Consult the "System Configuration Guide" on the NetApp Support site for specific models and support.
FC-VI card	If the current stretch MetroCluster configuration is not 31xx/32xx/62xx/80x0 series, then a separate FC-VI card must be installed in each controller as part of the upgrade process. (FAS31xx/32xx/62xx/80x0 stretch MetroCluster configurations already use the FC-VI card.)
Fibre Channel interswitch links (ISLs)	These FC connections carry the cluster interconnect and disk traffic between nodes. 1 ISL per fabric (thus, a total of 2) or 2 ISLs per fabric (thus, a total of 4) are required. If 2 ISLs per fabric are used, then traffic isolation (TI) must be implemented for Brocade switches. The distance between the two locations will determine the type of cabling, the Small Form-Factor Pluggables (SFPs) needed for the ISLs, and possibly the switch model itself. Refer to section 9.3 and to the Interoperability Matrix on the NetApp Support site for more information.
Wave division multiplexers (customer supplied)	Because fiber cabling is expensive, customers often deploy multiplexing devices in order to use the fiber for multiple purposes. NetApp MetroCluster supports whatever device is used as long as it is listed on the switch vendor's (Brocade's or Cisco's) compatibility guide for the model switch being used. This guide may be found at the switch vendor's website.
Cables	In a standard HA pair, many shelves can be connected in a physical disk loop. In a DS14FC fabric MetroCluster configuration there is a limit of 2 DS14FC physical disk shelves per loop, so accommodation might have to be made for additional loops and therefore additional Fibre Channel switch ports (1 loop [2 shelves] per port) and cabling. In SAS-based MetroCluster configurations, there is a limit of 10 SAS shelves per stack. Consider the cabling required for SAS-based fabric MetroCluster configurations: disk shelf to FibreBridge (SAS cables), FibreBridge to switch (FC cables).
Comments: • **Disk and shelf firmware compatibility**: Inventory the current disk shelves to verify that firmware levels are compatible with the Interoperability Matrix located on the NetApp Support site. If they are not, then the firmware must be upgraded as part of the process. • **Maximum spindle limits**: Be aware of the DS14FC spindle limits of a fabric MetroCluster configuration. See the Interoperability Matrix on the NetApp Support site for the latest information.	
Data ONTAP version	If a Data ONTAP operating system upgrade is planned, make sure it is compatible based on the Interoperability Matrix on the NetApp Support site.

 © 2014 NetApp, Inc. All Rights Reserved.

Stretch MetroCluster to Fabric MetroCluster	
Disk ownership	If the current stretch MetroCluster configuration is based on the older FAS3020 or 3050, then the disk ownership model will have to be changed in order to upgrade to a fabric MetroCluster configuration. Make sure that these controllers are compatible with the Data ONTAP version and switches being used.
Fabric switch licenses	It is extremely important that the proper licenses for the switches be ordered and obtained prior to the start of the upgrade. See the section on fabric MetroCluster components for the proper licenses for the fabric switches.

Procedure to convert from a stretch MetroCluster configuration to a fabric MetroCluster configuration: Plan for downtime and follow the fabric MetroCluster configuration steps in the "High-Availability and MetroCluster Configuration Guide" on the NetApp Support site.

15.4 Controller-Only Upgrades

Controller-Only Upgrades	
Chassis	If the HA pair consists of two controllers in the same chassis, a separate chassis is needed for one of the controllers to provide the physical separation.
FC-VI card	Depending on the specific model of controllers being replaced and the version of Data ONTAP to be used, a new FC-VI card might be necessary for each of the controllers. For example, the FC-VI card used on a 9xx or 3020/3050 is a 2Gbps card. Newer controllers and Data ONTAP 8.x and higher do not support this card. A 4Gbps or 8Gbps card will have to be ordered and installed as part of the upgrade process.
Comments: • **Disk and shelf firmware compatibility**: Inventory the current disk shelves to verify that firmware levels are compatible with the Interoperability Matrix located on the NetApp Support site. If they are not, then the firmware must be upgraded as part of the process. • **Maximum spindle count**: Depending on the specific model of controllers being replaced, the type of MetroCluster configuration (stretch or fabric), and the version of Data ONTAP to be used, if DS14 FC disks are used, the maximum number of DS14 FC disks might change. Refer to the "System Configuration Guide" or the Interoperability Matrix on the NetApp Support site for the platform being used.	
Data ONTAP version	If a Data ONTAP operating system upgrade is planned, make sure the upgrade is compatible based on the Interoperability Matrix.
Disk ownership	If the current active-active or HA pair is based on the older 3020 or 3050, then the disk ownership model must be changed in order to upgrade to a fabric MetroCluster configuration. Make sure that these controllers are compatible with the Data ONTAP version and switches being used.

Procedure for controller-only upgrades: Refer to the steps in the "High-Availability and MetroCluster Configuration Guide" on the NetApp Support site.

 © 2014 NetApp, Inc. All Rights Reserved.

15.5 MetroCluster to HA Pair

The high-level steps to convert a MetroCluster (both stretch and fabric) configuration to an HA pair follow. Plan for downtime because this is a disruptive procedure. Assume the MetroCluster configuration is composed of NodeA and NodeB.

1. Perform `cf takeover` on NodeB.
2. Offline plexes formed by NodeA_pool0 and NodeB_pool1 (plexes on the NodeA site).
3. If fabric MetroCluster: Disable the ISLs between sites. If stretch MetroCluster: Disconnect cross-site cabling. Data is then served by NodeB from NodeA_pool1 and NodeB_pool0.
4. Power off the equipment at the NodeA site.
5. Move NodeA and shelves over to the NodeB site.
6. Halt NodeB.
7. Recable NodeA and NodeB as an HA pair.
8. Power up the equipment.
9. Aggregates should resync; then you have local HA and SyncMirror.
10. NodeA should be waiting for giveback.
11. Perform a `cf giveback` from NodeB to NodeA.
12. Break mirrors and reclaim space if SyncMirror is not required.

15.6 Relocating a MetroCluster Configuration

The high-level steps to relocate a MetroCluster (both stretch and fabric) configuration follow. This is a nondisruptive procedure. The MetroCluster configuration is composed of NodeA and NodeB. We are relocating NodeA.

1. Perform `cf takeover` on NodeB.
2. Offline plexes formed by A_pool0 and B_pool1 (plexes on the NodeA site).
3. If fabric MetroCluster: Disable the ISLs between sites. If stretch MetroCluster: Disconnect cross-site cabling. Data is then served by NodeB from NodeA_pool1 and NodeB_pool0.
4. Power off the equipment at the NodeA site.
5. Relocate the equipment to the new location.
6. If fabric MetroCluster: Power on switches and shelves. If stretch MetroCluster: Power on shelves.
7. If fabric MetroCluster: Enable ISLs. If stretch MetroCluster: Reconnect cross-site connections.
8. Online (previously offlined) plexes.
9. Wait for resync to complete.
10. Power on NodeA.
11. NodeA should be waiting for giveback.
12. Perform `cf giveback` on NodeB.

16 MetroCluster Interoperability

This section deals with guidelines for topics that are generally treated as external to the MetroCluster configuration.

 © 2014 NetApp, Inc. All Rights Reserved.

16.1 Fabric MetroCluster and the Front End

As a best practice, the front-end (fabric or Ethernet) network should span both sites. There should be redundancy in the front end—redundant networks or fabrics—for an end-to-end redundant solution. If the front-end fabric or Ethernet network does not span both sites, normal takeover/giveback operations will not work.

16.2 Fabric MetroCluster and FCP Clients

Keep in mind the following when using FCP clients and fabric MetroCluster connectivity between sites:

- Front-end SAN should span both sites (if it doesn't, normal takeover/giveback won't work).
- If MetroCluster back-end fabric connectivity between sites is lost, certain SAN host multipathing stacks might not respond correctly to the situation, causing I/O failure to the host.
- MetroCluster supports all host utilities/multipathing stacks supported by NetApp. Caveats might exist, and additional configuration might be necessary.

16.3 V-Series/FlexArray Virtualization MetroCluster

V-Series/FlexArray virtualization MetroCluster follows the same support structure as FAS MetroCluster. Key points to note are:

- Supports the same fabric switches as FAS MetroCluster (Brocade and Cisco)
- No support for open SAN or shared SAN
- Support for native NetApp storage with Data ONTAP 8.2
- Support for SAS optical with Data ONTAP 8.2.1

See the Interoperability Matrix on the NetApp Support site for the latest support information.

For information about V-Series/FlexArray virtualization MetroCluster configurations with back-end storage, see the following resources on the NetApp Support site:

For Data ONTAP 8.0.3 and later, see the High-Availability and MetroCluster Configuration Guide.

For earlier Data ONTAP releases, see the "V-Series MetroCluster Guide."

16.4 MetroCluster and SnapMirror

Using MetroCluster in combination with NetApp SnapMirror® (async) technology creates a compelling DR (disaster recovery) solution spanning metro as well as regional distances. A fabric MetroCluster configuration provides metrowide protection, and SnapMirror manages disaster recovery over regional distances.

16.5 MetroCluster and Storage Efficiency

Deduplication functionality is supported on Data ONTAP 7.2.5.1 or later and 7.3.1 or later.

Compression supports both fabric and stretch MetroCluster beginning with Data ONTAP 8.0.2. When using MetroCluster with compression and/or deduplication, consider the following:

- Deduplication has an impact on CPU resources as a result of extra disk write operations. The increase is due to writing to two plexes. On most platforms the impact is less than 10%. This impact is more pronounced on low-end systems than on high-end systems.
- Compression continues in takeover mode.
- A node in takeover mode takes over the servicing of I/Os targeted at the partner volumes as well as its compression and change logging. As a result, additional system resources are consumed, which might require that the system workload be adjusted.
- If the compression scanner is running before takeover or giveback, it needs to be restarted after takeover or giveback completes.

 © 2014 NetApp, Inc. All Rights Reserved.

- Only a subset of deduplication commands for the partner volumes is available in takeover mode. These commands are `sis status, sis stat, sis on, sis off.`
- Deduplication and compression (if applicable) must be licensed on both nodes.
- Prior to 8.1, in takeover mode, writes to partner flexible volumes are change logged. The deduplication process does not run on the partner flexible volumes while in takeover mode. Upon giveback, data in the change logs is processed and data gets deduplicated.
- Prior to 8.1, in takeover mode, change logging continues until the change log is full. This can occur if the node remains in takeover mode for a long period of time, such as during a disaster. All data continues to be accessible regardless of change log availability.

In Data ONTAP 8.1 and later:

- In takeover mode, compression and deduplication continue to run normally as per the schedule.
- A node in takeover mode takes over the servicing of I/Os targeted at the partner volumes. As a result, additional system resources are consumed, which might require that the system workload be adjusted.

17 Solutions on MetroCluster

Following are the technical reports on solutions leveraging MetroCluster. Note that the reports are owned by the solution business unit.

Solution Area	TR	Last Update
Oracle® RAC, Oracle ASM on NetApp MetroCluster	TR 3816	February 2010
DB2 9.7 HA on NetApp MetroCluster	TR 3807	November 2009
VMware HA, FT on MetroCluster	TR 3854	May 2010
Microsoft® Exchange, SQL Server®, SharePoint® on Microsoft Hyper-V®, and MetroCluster	TR 3804	January 2010
VMware on NetApp	TR 3788	June 2010
HA and DR using VMware, MetroCluster, and SnapMirror	TR 3606	November 2009
Sun Cluster with MetroCluster	TR 3639	October 2009
(POC) Xen on MetroCluster	TR 3755	March 2009
(POC) Microsoft Exchange on FCP on MetroCluster	TR 3412	January 2008

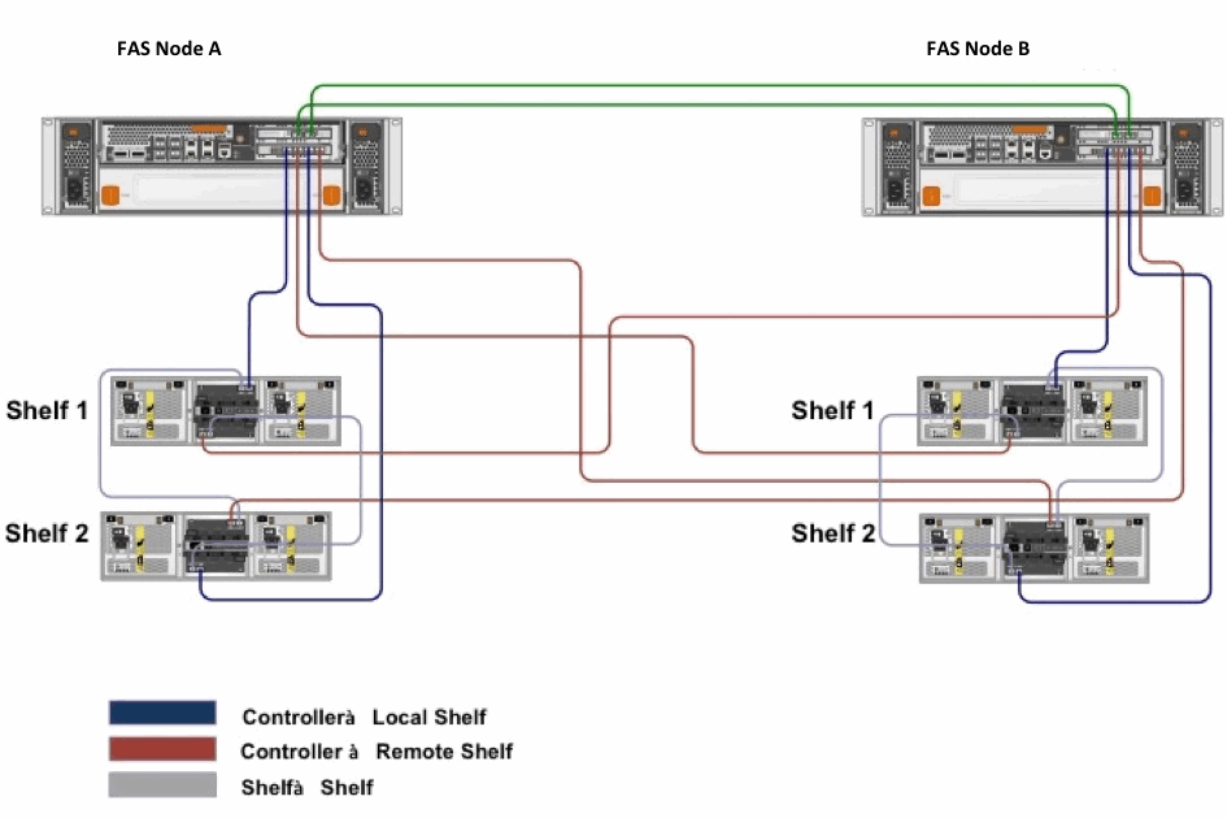
 © 2014 NetApp, Inc. All Rights Reserved.

Appendix

Cabling Diagrams

FAS32xx Stretch MetroCluster Using DS14 Shelves

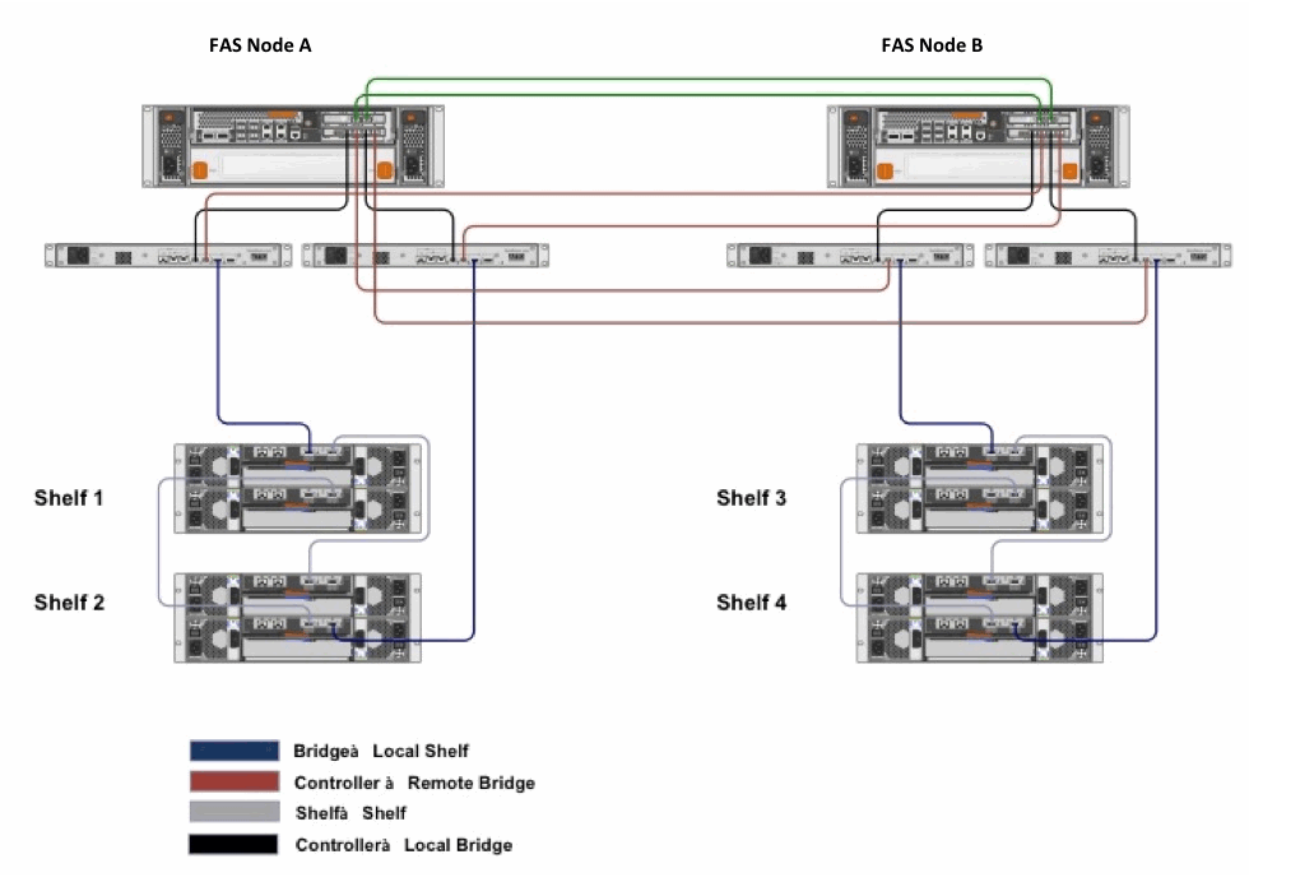
Figure 25) FAS32xx stretch MetroCluster using DS14 shelves.



 © 2014 NetApp, Inc. All Rights Reserved.

FAS32xx Stretch MetroCluster Using SAS Shelves with FibreBridge

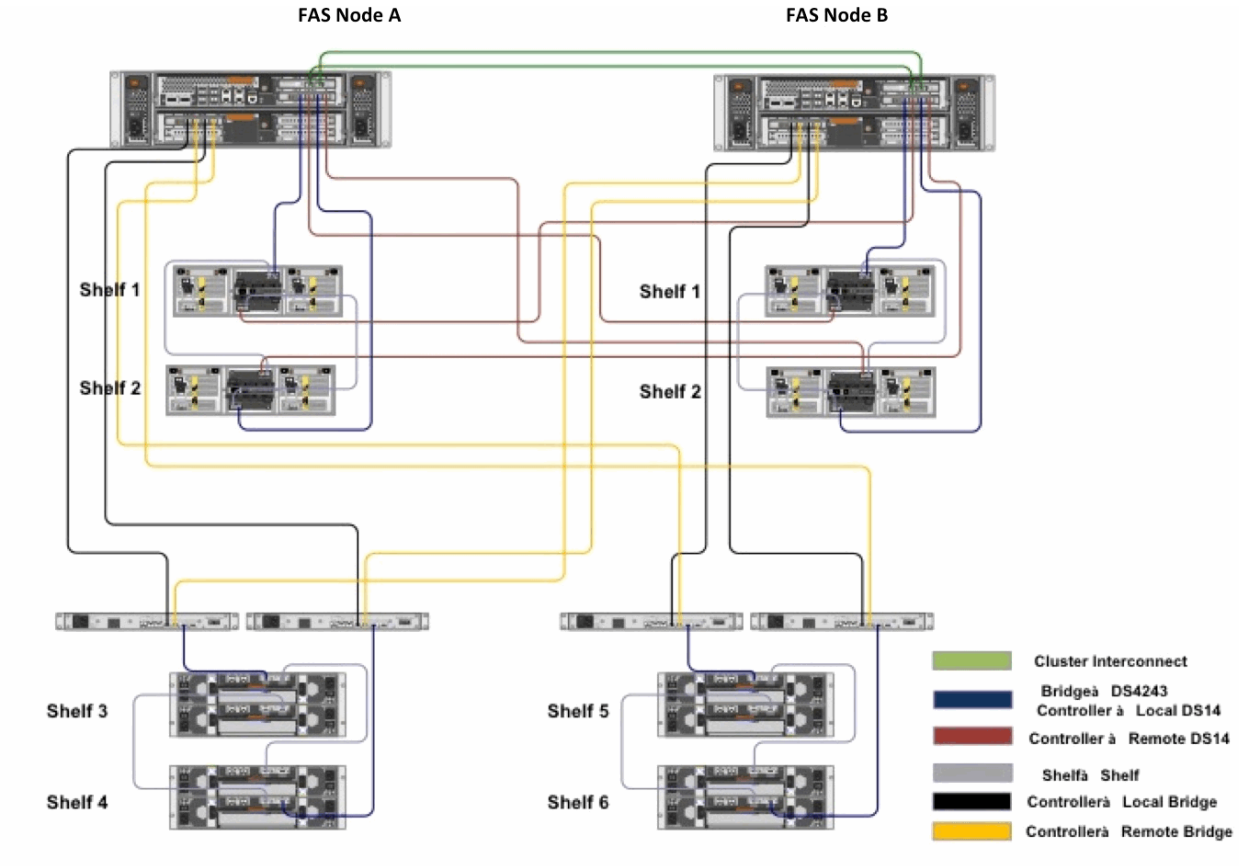
Figure 26) FAS32xx stretch MetroCluster using SAS shelves with ~~the~~ FibreBridge.



 © 2014 NetApp, Inc. All Rights Reserved.

FAS32xx Stretch MetroCluster Using DS14 and SAS Shelves (with FibreBridge)

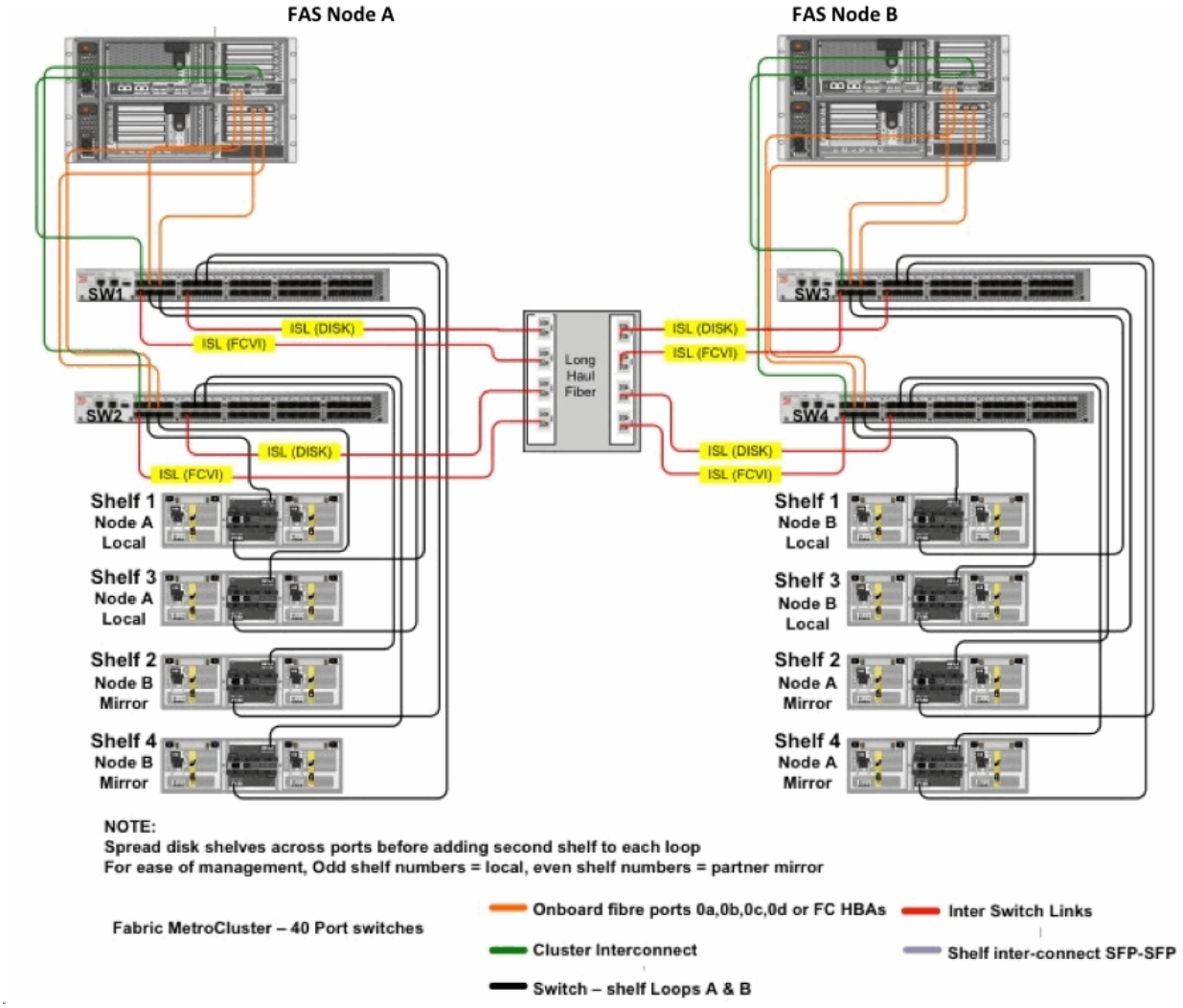
Figure 27) FAS32xx stretch MetroCluster using DS14 and SAS shelves (with FibreBridge).



 © 2014 NetApp, Inc. All Rights Reserved.

FAS62xx Fabric MetroCluster Using DS14FC Shelves

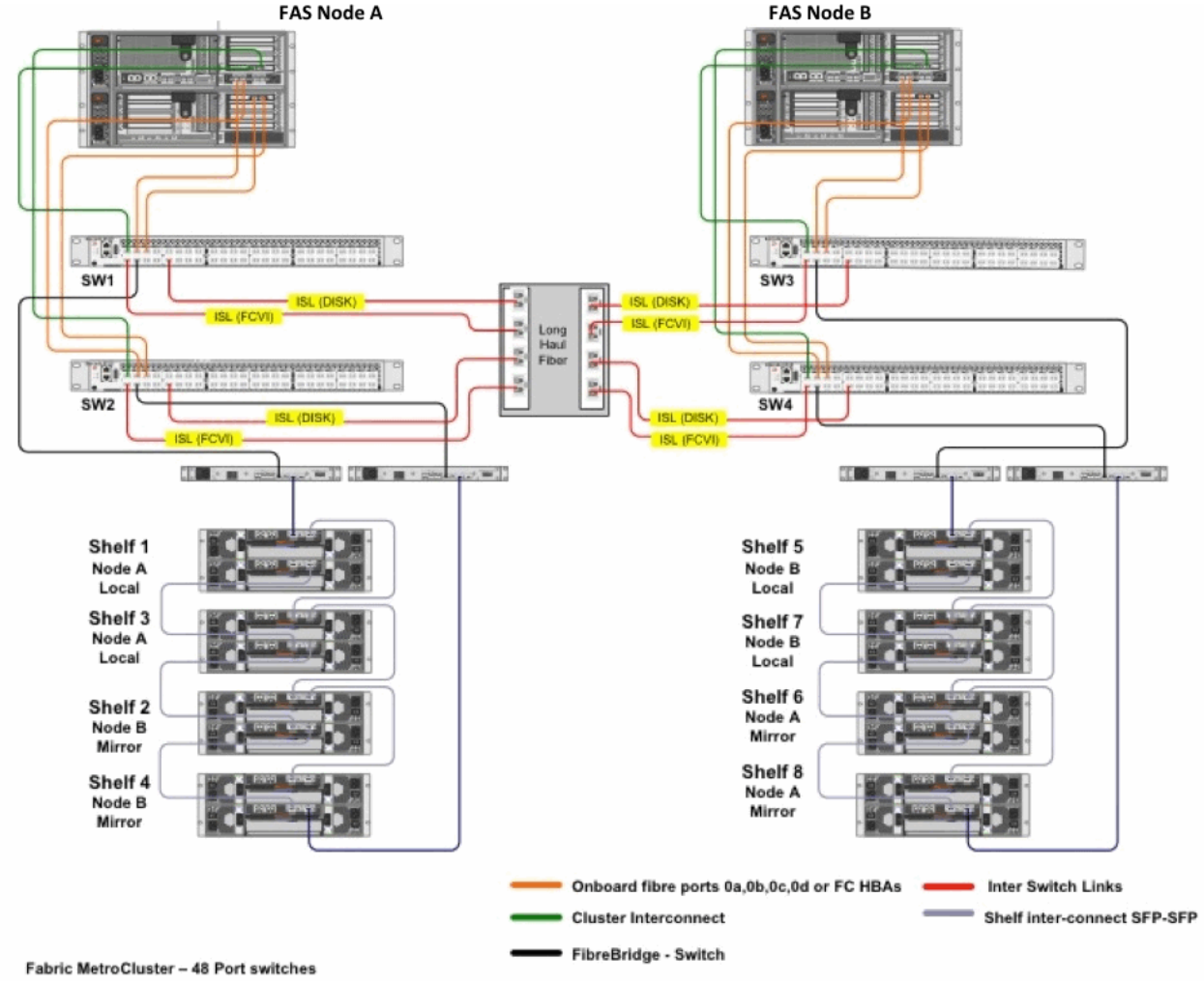
Figure 28) FAS62xx fabric MetroCluster using DS14FC shelves.



 © 2014 NetApp, Inc. All Rights Reserved.

FAS62xx Fabric MetroCluster Using SAS Shelves with FibreBridge

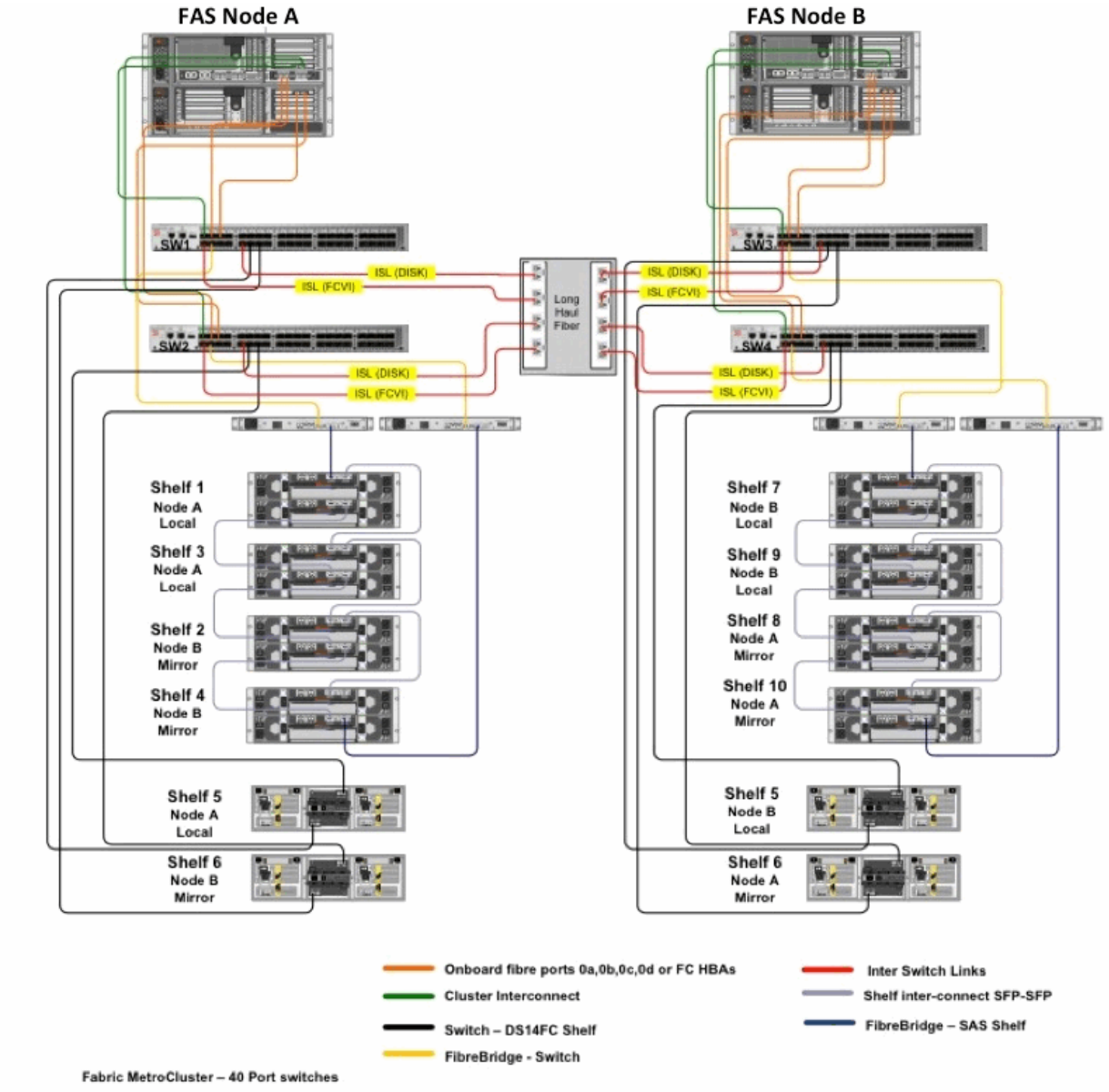
Figure 29) FAS62xx fabric MetroCluster using SAS shelves with FibreBridge.



 © 2014 NetApp, Inc. All Rights Reserved.

FAS62xx Fabric MetroCluster Using DS14FC and SAS Shelves (with FibreBridge)

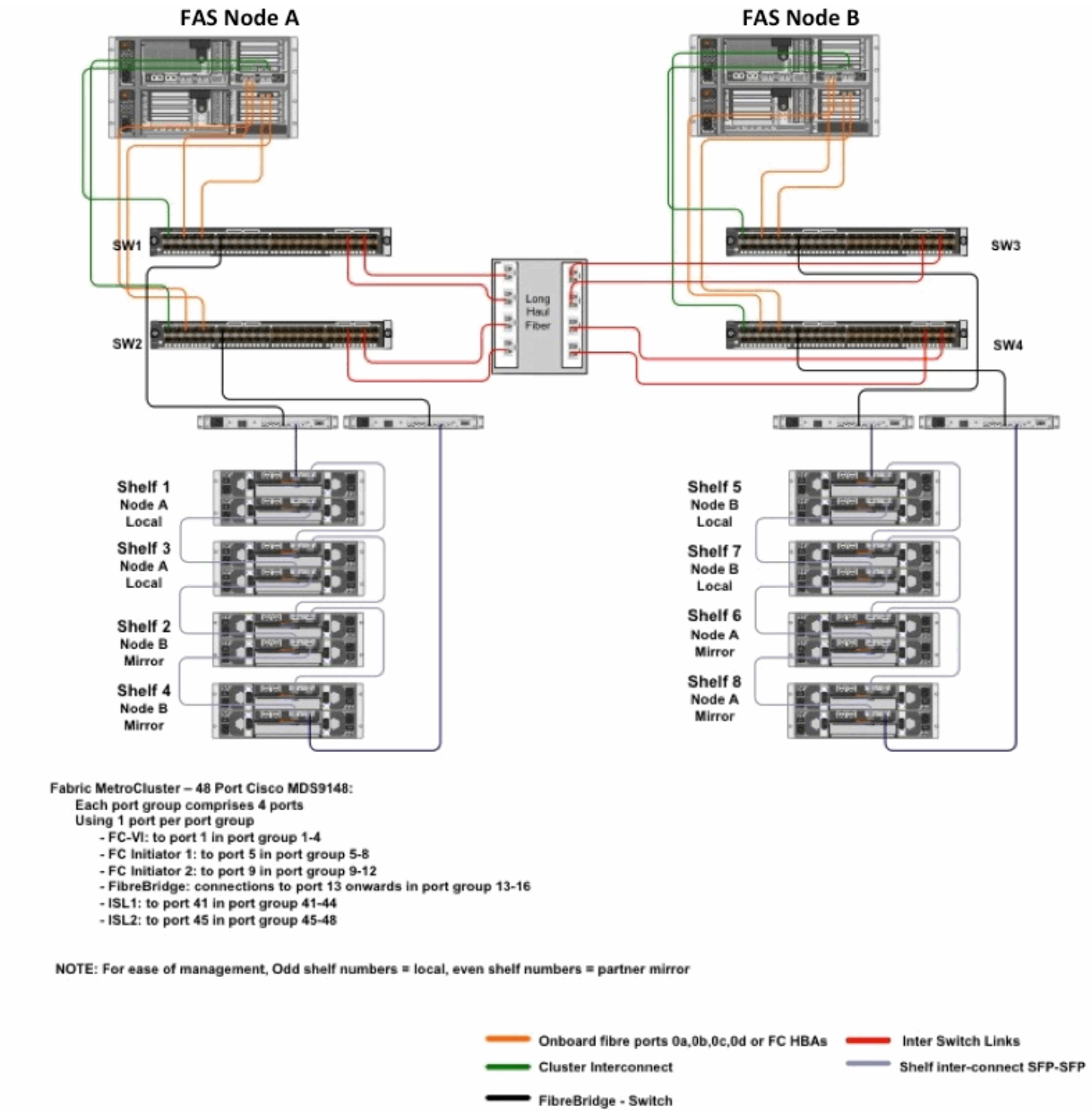
Figure 30) FAS62xx fabric MetroCluster using DS14FC and SAS shelves (with FibreBridge).



 © 2014 NetApp, Inc. All Rights Reserved.

FAS62xx Fabric MetroCluster with Cisco 9148

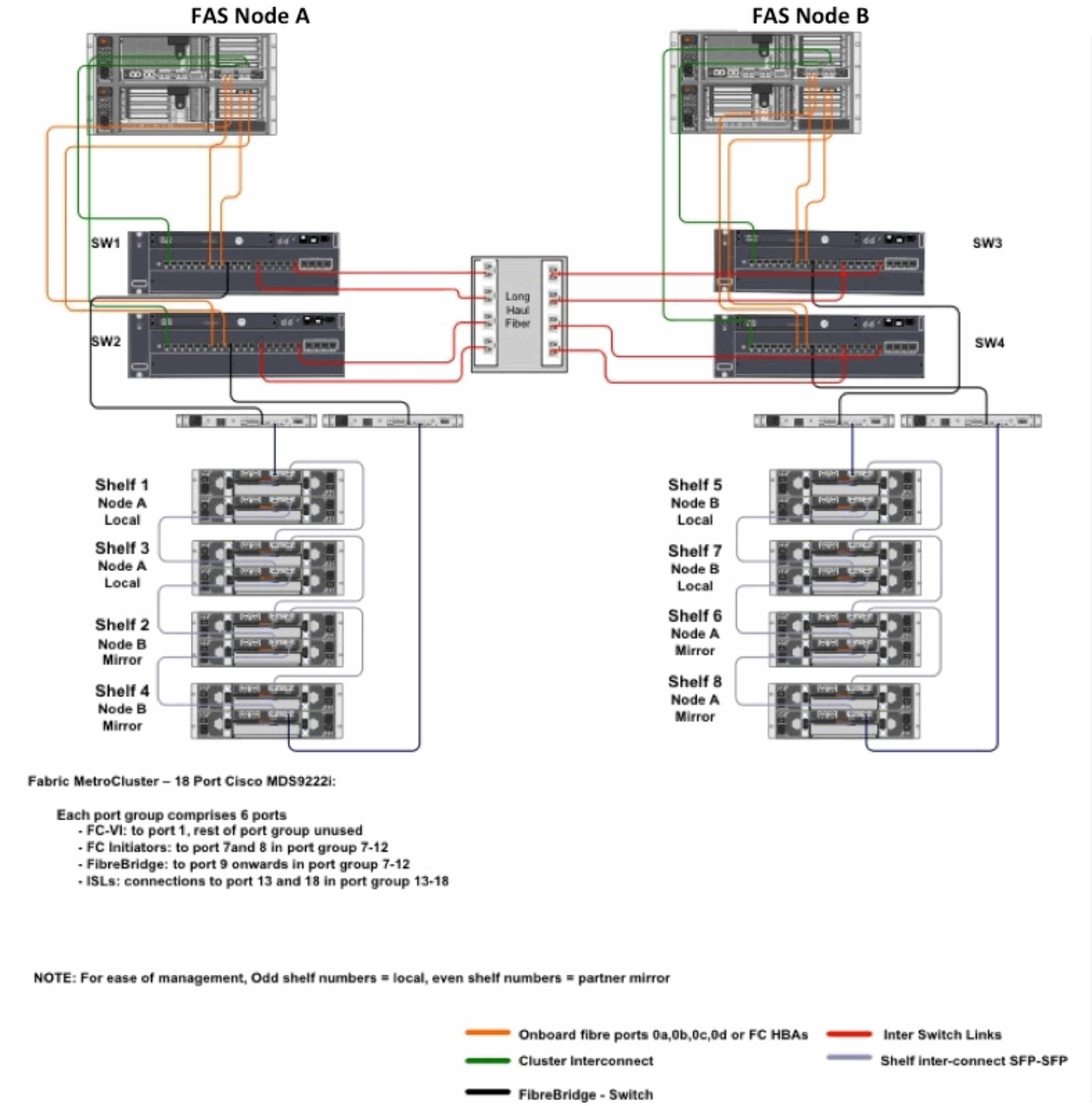
Figure 31) FAS62XX fabric MetroCluster with Cisco 9148.



 © 2014 NetApp, Inc. All Rights Reserved.

FAS62xx Fabric MetroCluster with Cisco 9222i

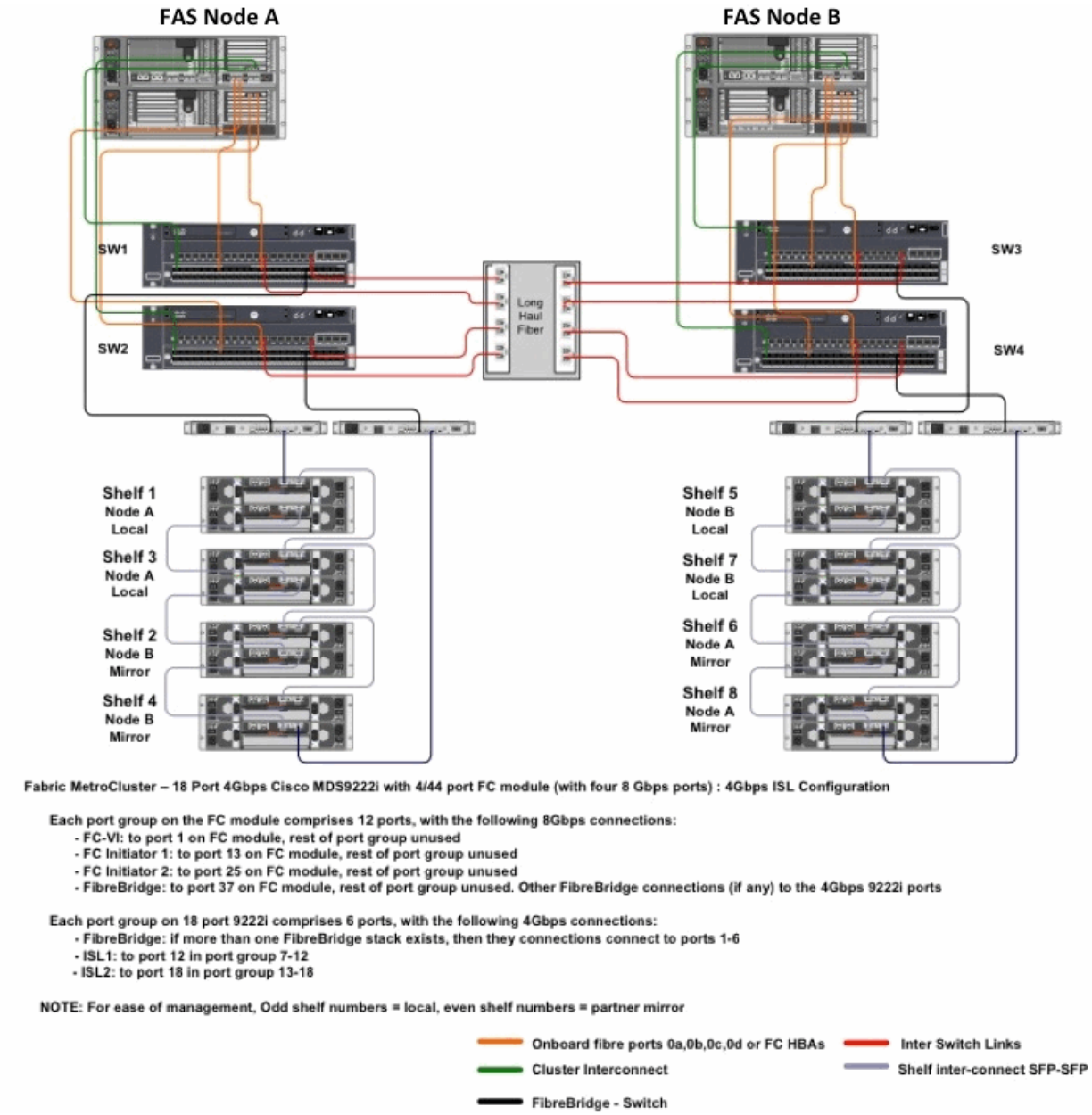
Figure 32) FAS62XX fabric MetroCluster with Cisco 9222i.



 © 2014 NetApp, Inc. All Rights Reserved.

FAS62xx Fabric MetroCluster with Cisco 9222i Plus 4/44 FC Module with 4Gbps ISLs

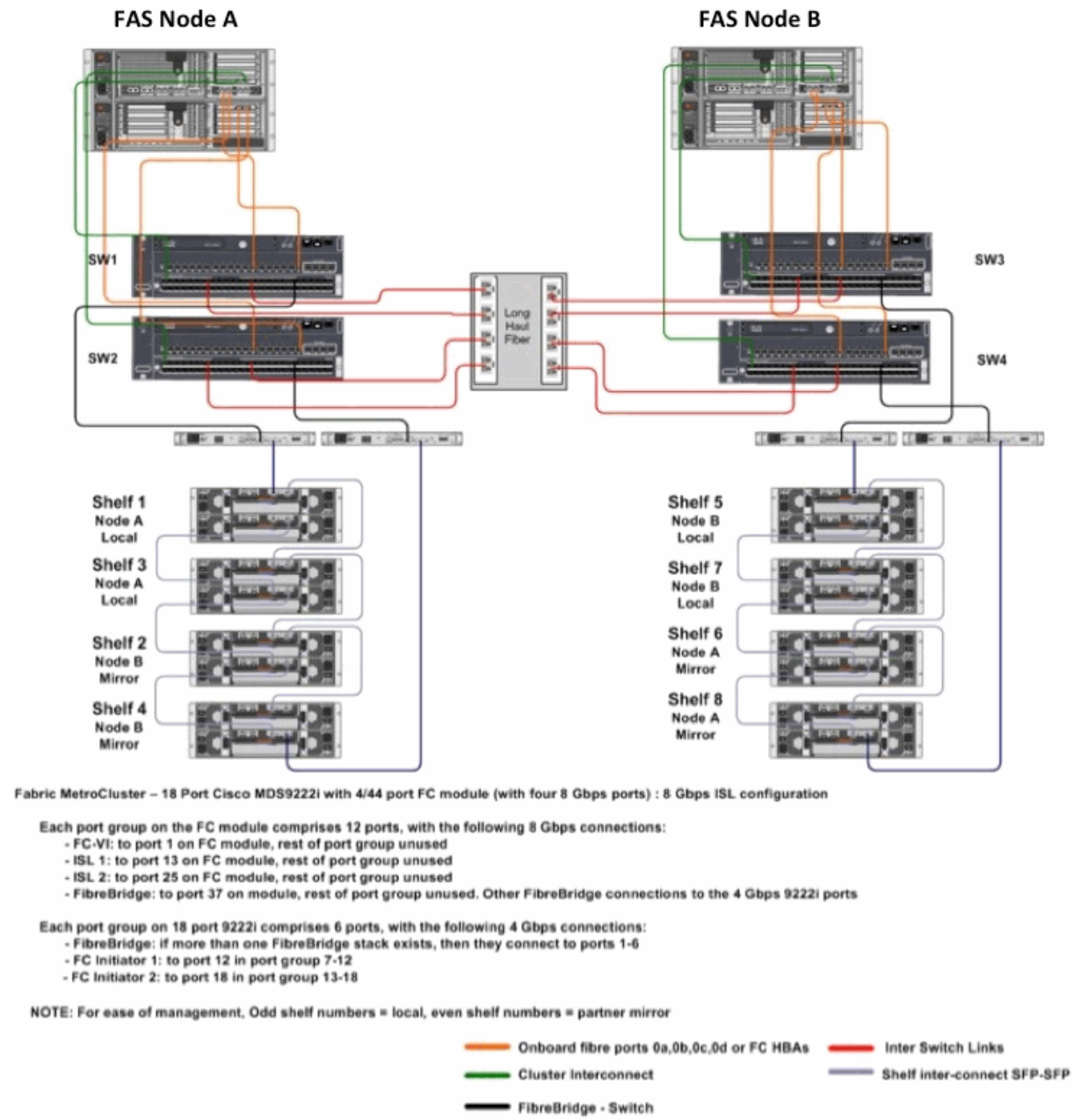
Figure 33) FAS62XX fabric MetroCluster with Cisco 9222i plus 4/44 FC module with 4Gbps ISLs.



 © 2014 NetApp, Inc. All Rights Reserved.

FAS62xx Fabric MetroCluster with Cisco 9222i Plus 4/44 FC Module with 8Gbps ISLs

Figure 34) FAS62XX fabric MetroCluster with Cisco 9222i plus 4/44 FC module with 8Gbps ISLs.



 © 2014 NetApp, Inc. All Rights Reserved.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NetApp provides no representations or warranties regarding the accuracy, reliability, or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

Go further, faster®

© 2014 NetApp, Inc. All rights reserved. No portions of this document may be reproduced without prior written consent of NetApp, Inc. Specifications are subject to change without notice. NetApp, the NetApp logo, Go further, faster, AutoSupport, ASUP, Data ONTAP, Flash Pool, MetroCluster, OnCommand, RAID-DP, SnapMirror, Snapshot, and SyncMirror are trademarks or registered trademarks of NetApp, Inc. in the United States and/or other countries. Hyper-V, Microsoft, SharePoint, SQL Server, Windows, and Windows Server are registered trademarks of Microsoft Corporation. Oracle and Java are registered trademarks of Oracle Corporation. Cisco is a registered trademark of Cisco Systems, Inc. ESX, VMware, and VMware vSphere are registered trademarks and ESXi and vCenter are trademarks of VMware, Inc. All other brands or products are trademarks or registered trademarks of their respective holders and should be treated as such. TR-3548-0514

NetApp®

www.netapp.com