Technical Report

High-Availability (HA) Pair Controller Configuration Overview and Best Practices

Jay Bounds, NetApp
February 2016 | TR-3450

Abstract

The NetApp® HA pair controller configuration delivers a robust and high-availability data service for business-critical environments. Each of the two identical storage controllers within the HA pair configuration serves data independently during normal operation. In the event of individual storage controller failure, the data service process transfers from the failed storage controller to the surviving partner. The HA pair controller configuration can also protect against other hardware failures, including network interface cards, FC-AL loops, and shelf I/O modules

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

In today's environment, business needs require 24/7 data availability. The storage industry delivers the base building block for IT infrastructure for data storage for all business applications and objectives. Therefore, constant data availability begins with architecting storage systems that facilitate nondisruptive operations (NDO). Nondisruptive operations have three main objectives: hardware resiliency, hardware and software lifecycle operations, and hardware and software maintenance operations. This technical report focuses on hardware resiliency and hardware and software maintenance operations in which continuous data availability during NDO is a function of the following components.

- **Performance.** Performance can be broken into two key perspectives from a data-availability point of view. The first is that customers have specific performance requirements that they must meet in order to satisfy applications that depend on storage system data being readily available. A data-availability outage means that the storage system may still respond to foreground I/O but fall below the requirements of the dependent applications' ability to function. The second perspective is that if a system's performance suffers to the extent that the system stops responding to foreground I/O, then a data-availability outage situation has been encountered.
- **Resiliency.** From the point of view of data availability, resiliency is the system's capability to suffer a single failure or several failures while continuing to respond to foreground I/O in the degraded state. Numerous options and features contribute to a system's capability to withstand failures; they are discussed throughout this document.
- **Recoverability.** Recoverability defines the system's capability to both automatically recover from failures and continue to respond to foreground I/O while conducting recovery operations on the storage system.

These three factors are further applied to the three layers of data availability.

- **Storage subsystem.** The storage subsystem layer addresses all hardware components and software features that relate to the storage system's internals. This layer can be considered to be from the HBA down through the attached storage arrays from a physical perspective, or around the storage and RAID software layers that are part of the NetApp Data ONTAP® operating system. Basically, this layer addresses the system's ability to communicate internally from the controller to the attached storage arrays.
- **System.** The system layer addresses the capability of a storage system to suffer failures. This layer focuses primarily on controller-level failures that affect the capability of a system to continue external communication. This applies to single-controller and high-availability (HA) pair configurations and the components that contribute to external controller communication such as network interfaces.
- **Site.** The site layer addresses the capability of a group of colocated storage systems to suffer failures. This layer focuses primarily on the features related to distributed storage system architecture that allow an entire storage system failure. Such a failure would probably be related to a site-level incident, such as a natural disaster or an act of terrorism**.**

The core foundation of NDO is the HA pair controller configuration, which provides high-availability solutions during planned and unplanned downtime events. The rest of this report is an overview and description of the technical concepts of the HA pair configurations with recommendations for best practices and solutions for different business requirements.

1.1 Scope

At the system level, NetApp delivers a robust and highly available data solution for mission-critical environments referred to as the HA pair controller configuration. Each of the two identical storage controllers within the HA pair configuration serves data independently during normal operation. In the event of individual storage controller failure, the data service process transfers from the failed storage controller to the surviving partner. The HA pair configuration can also protect against other hardware failures, including network interface cards, FC-AL loops, and shelf I/O modules.

This document covers:

- Overview of the hardware and software components of the HA pair configuration
- Best practices for evaluating HA pair solutions that meet the needs of customer environments
- Client interaction during failover and giveback operations
- Best practices to minimize client disruption
- Nondisruptive upgrade (NDU)
- Command line interface (CLI) parity between Data ONTAP operating in 7-Mode and clustered Data ONTAP

For information on the resiliency and recoverability of the storage subsystem, refer to TR-3437: Storage Subsystem Resiliency Guide.

1.2 Terminology Used in this Document

Table 1) Terminology used within the paper

Terminology	Description / Function
Storage controller, FAS system, node, partner node	The physical entity of the controller
Takeover, failover	The functional capability for a node to take over its partner's disks when the partner is downed for a planned or unplanned event
Giveback, failback	The functional capability of a node to return its partner's disk to it once the partner node is booted after a planned or unplanned event
Node, partner node	Controller within an HA pair controller configuration
Controller failover, CFO	The mechanism to fail over volumes in 7-Mode or CFO policy volumes (node root volume) in clustered Data ONTAP
Storage failover, SFO	The mechanism to fail over volumes in clustered Data ONTAP
Nondisruptive upgrade, NDU	The mechanism to update Data ONTAP software and firmware on a system and associated storage
Automated NDU, ANDU	An automated process for updating clustered Data ONTAP software and firmware on a system and associated storage
Rolling upgrade, rolling batch upgrade	The process of executing a Data ONTAP upgrade on several HA pair controllers in parallel within a cluster

2 Overview of HA Pair Controllers

The HA pair controller configuration consists of a pair of matching FAS storage controllers (local node and partner node); each of these nodes must be connected to the other's disk shelves. The Data ONTAP and firmware versions must be identical on both nodes. Similarly, the interconnect adapters on each node must be identical and configured with the same firmware version, and the interconnect adapters must be connected properly by appropriate interconnect cables. For cabling details, refer to the **High Availability Configuration Guide**.

In the HA pair controller environment, Data ONTAP on each node monitors the availability status of its partner by means of a heartbeat signal transmitted between the storage controllers through the interconnect cards and cables. It then stores this information on specialized mailbox disks. FAS storage controllers use battery-backed nonvolatile RAM (NVRAM) to prevent the loss of any data input/output requests that may have occurred after creation of the most recent consistency point. The NVRAM data of each controller node in the HA pair is always mirrored on the partner node. In the event of failover, the surviving node assumes control of the failed node's disks and maintains data consistency with the mirrored NVRAM. For additional details on NVRAM, see TR-3001.

The NetApp FAS2000 and versions of the FAS3100 series controllers do not use interconnect cards. The heartbeat signal and NVRAM data are transmitted between the nodes via integrated Ethernet ports.

2.1 How the Interconnect Works

The interconnect adapters are among the most critical components in HA pair controllers. Data ONTAP uses these adapters to transfer system data between the partner nodes, thereby maintaining data synchronization within the NVRAM on both controllers. Other critical information is also exchanged across the interconnect adapters, including the heartbeat signal, system time, and details concerning temporary disk unavailability due to pending disk firmware updates. The following section explains why NVRAM must be identical on both nodes.

Because NVRAM5 and NVRAM6 cards provide integrated interconnect hardware functionality, standalone interconnect cards are not used (or necessary) when NVRAM5 or NVRAM6 cards are present, except when using the fabric MetroCluster™ configuration, described later in this document.

2.2 How HA Pairs Handle NVRAM

Data ONTAP uses the WAFL® (Write Anywhere File Layout) file system to manage data processing and uses NVRAM to enable data consistency before committing writes to disks. Data in the NVRAM is copied to system memory through Direct Memory Access (DMA). If the storage controller encounters a power failure, the most current data is protected by the NVRAM, and file system integrity is maintained.

In the HA pair controller environment, each node reserves half of the total NVRAM size for the partner node's data so that exactly the same data exists in NVRAM on both storage controllers. Therefore, only half of the NVRAM in the HA pair controller is dedicated to the local node. Dividing the NVRAM in half to provide data consistency incurs approximately a 2% to 3% performance penalty. If failover occurs when the surviving node takes over the failed node, all WAFL checkpoints stored in NVRAM are flushed to disk. The surviving node then combines the split NVRAM and recovers the lost performance. Once the surviving node restores disk control and data processing to the recovered failed node, all NVRAM data belonging to the partner node is flushed to disk during the course of the giveback operation.

Single-Node Cluster Scalability and NVRAM

A single-node cluster provides the capability to have one node within its own cluster. A single-node cluster does not have any HA resiliency; therefore, a single-node cluster does not split the NVRAM of the node. A single-node cluster may not meet the demands of the business over time, requiring the single node to make the transition into a more resilient two-node cluster. In order to do that, the division of NVRAM on each storage controller would go into effect to properly mirror the partner node's data. For the

changes required to transition each node into an HA configuration and restructure the division, the NVRAM needs the controller to reboot. Once the single node makes the transition into a two-node cluster, all NDO capabilities facilitated by an HA controller configuration are in place.

The following diagrams describe the relationship between NVRAM and HA pair controller configurations.

Figure 1) HA pair controller configuration in normal operation

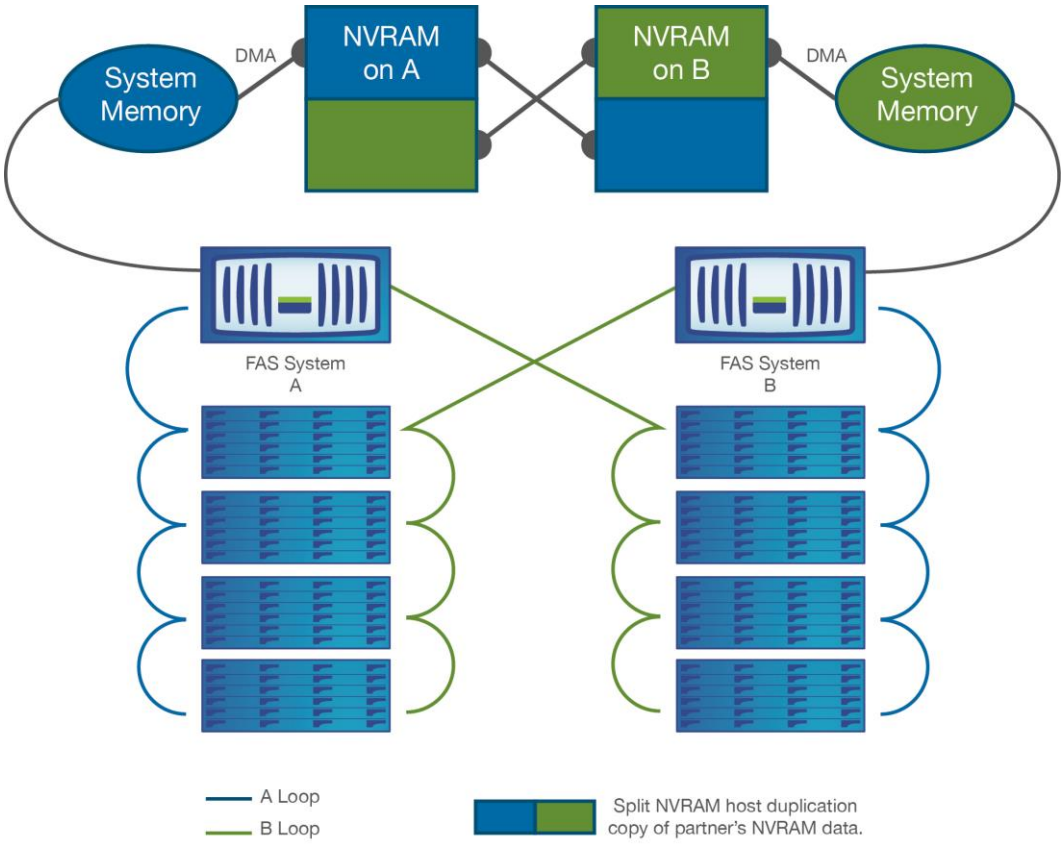
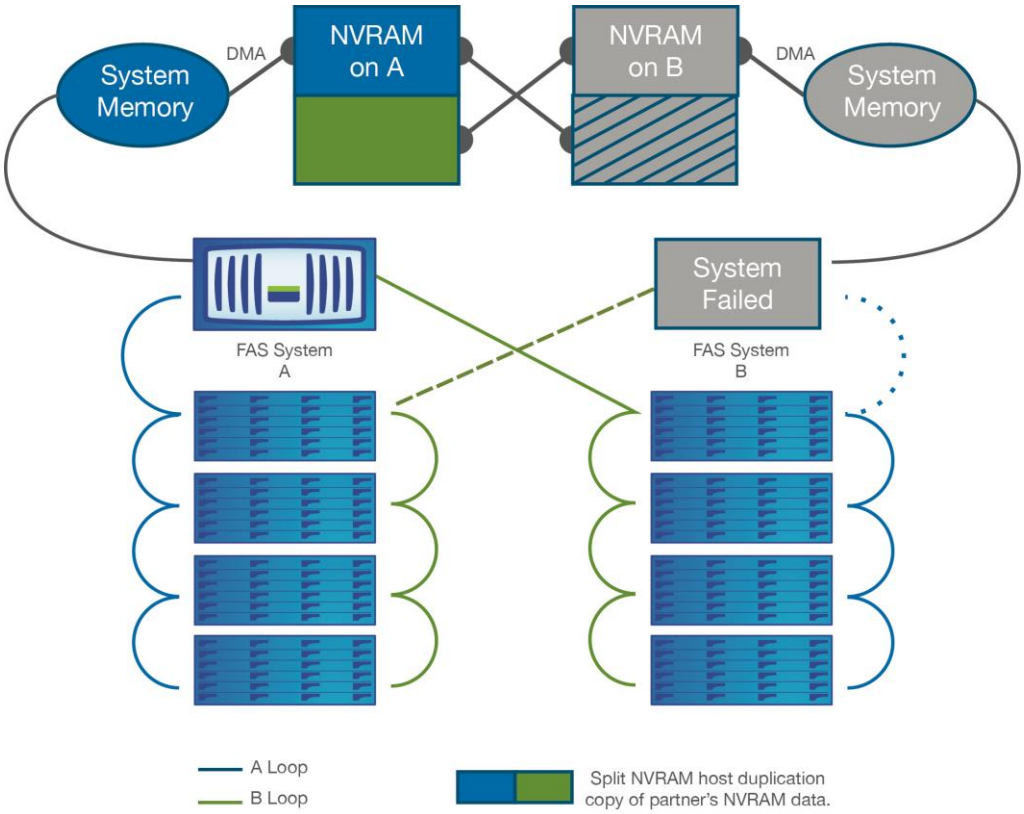


Figure 2) HA pair controller configuration in failover



2.3 Mailbox Disk Store for HA Pair Synchronization Information

So that both nodes within the HA pair controller configuration maintain the correct and current status of one another, the node's status and heartbeat information are stored on the mailbox disks of each node; a redundant set of disks is used in coordinating takeover or giveback operations. If one node stops functioning, the surviving partner node uses the information on the mailbox disks to perform takeover processing, which creates a virtual storage system. The mailbox heartbeat information prevents an unnecessary failover from occurring in the event of interconnect failure. Moreover, if HA information stored on the mailbox disks is out of sync at boot time, HA pair nodes automatically resolve the situation. The FAS system failover process is extremely robust, preventing split-brain issues from occurring.

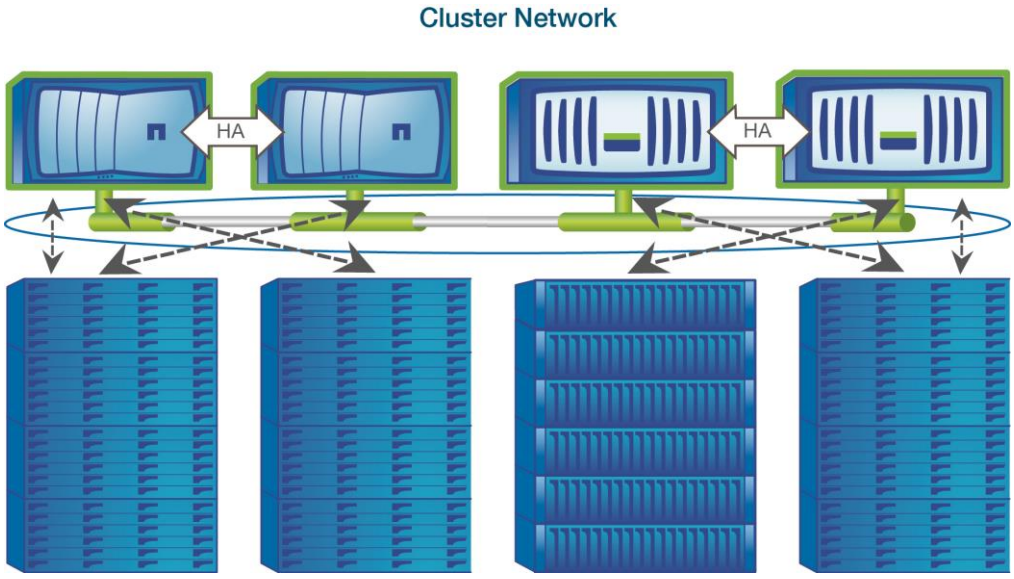
2.4 How HA Pairs Fit into Clustered Data ONTAP

Data ONTAP 8.0 introduced a new generation of storage systems, enabling the scale-out of numerous HA pairs into a single cluster. Previously, the term cluster was synonymous with an HA pair, but clustered Data ONTAP clearly delineates an HA pair from a cluster. An HA pair consists of the two partner nodes while a cluster is made up of numerous HA pairs. The HA pairs are joined together with a back-end network referred to as the cluster network. The HA pairs are the physical components that make up the larger logical entity that is referred to as the cluster. HA pairs provide storage resiliency at the system level for improved overall availability of the cluster.

The HA pair controller configuration maintains the same resiliency as a 7-Mode system. Each HA pair is directly attached (cabled) storage making use of multipath cabling for an additional layer of resiliency.

The following diagram shows a four-node cluster with Node 1, Node 2, Node 3, and Node 4. Node 1 and Node 2 are HA partners and Node 3 and Node 4 are HA partners. All four nodes are connected by the cluster network.

Figure 3) Cluster network connecting four nodes in a cluster



Although the physical architecture of the traditional 7-Mode HA pair controller configuration is the base building block of the cluster, there are several added benefits that the scale-out architecture provides beyond that of the traditional 7-Mode architecture.

Each node in an HA pair is linked to a cluster network that facilitates communication between each node within the cluster. The back-end cluster network is the communication backbone for both foreground and background I/O between the nodes. For example, clustered Data ONTAP delivers the additional benefit of allowing data to be moved around the cluster nondisruptively to enhance the capabilities of the HA pair solutions and enhance the NDO capabilities of NetApp storage systems. The movement of volumes and the replication of data can be done for any node within the cluster and is done over the cluster network. For more information on data mobility solutions, refer to TR-3975, DataMotion for Volumes Overview.

2.5 Cluster Failover (CFO) and Storage Failover (SFO)

There is also a layer of virtualization between the storage resources and the network resources, both of which are interfaced with the storage controllers in the cluster. The cluster network allows storage resources to be accessed through a network resource in the cluster. Therefore, a foreground I/O request can be received through a network interface on any node in the cluster and directed to the appropriate storage controller (and associate storage resource) to access the required data. Cluster failover refers to the policy associated with volumes on 7-Mode systems. Storage failover refers to the policy associated with volumes on clustered Data ONTAP systems. However, a node root volume on a clustered Data ONTAP system will retain CFO policy.

In a 7-Mode system, all aggregates are failed over and given back together. The failover or giveback is not complete until all aggregates are returned to the partner node. Alternatively, in clustered Data ONTAP, aggregates are given back to the partner node in sequence: First the aggregates with volumes having the CFO policy are given back, all in parallel. It is a best practice that the node root aggregate have no user data contained within it. The process for bringing the root aggregate online involves syncing certain information with other nodes within the cluster and may take some time; therefore, the root aggregate is unavailable to serve data while it is in transition. Providing that all user data is on aggregates with SFO policy, the partner will continue to serve data to all aggregates having volumes containing user data until the node root volume is brought online and the node is assimilated back into the cluster and ready to receive the remaining aggregates. At this point, each aggregate is returned to the partner node, serially, so that each aggregate incurs a short period of transition as it is brought online on the partner

node (original home node). The same process is applied to planned take-over events for HA pair controllers for clustered Data ONTAP starting in Data ONTAP 8.2.

The following table outlines the failover and giveback sequences for clustered Data ONTAP.

Table 2) Storage failover and cluster failover events

HA Event	Event Description
Unplanned Event	All aggregates fail over to partner node in parallel.
Planned Event (Clustered Data ONTAP 8.1)	All aggregates fail over to partner node in parallel.
Planned Event (Clustered Data ONTAP 8.2, 8.3)	Each aggregate is failed over serially, the root aggregate is failed over once all user data containing aggregates is failed over to the partner node.
Giveback	Root aggregate is given back first; once a node is assimilated back into the cluster each data-containing aggregate is given back serially to the partner node.

2.6 Clustered Data ONTAP, HA Pairs, Cluster Quorum and Epsilon

A cluster can be made up of a single node, two nodes, or more than two nodes. The availability of the cluster for a single node depends on the single node staying up. If the single node experiences a planned or unplanned event that takes the node down, then there is no availability to any data residing on the attached storage.

For a cluster of two or more nodes, the availability of each node increases from the HA capabilities associated with failover and giveback. In the event of node failure, a failover happens and data continues to be accessed via the partner. Communication between the nodes is an important part of clustered Data ONTAP. If a certain number of nodes—which depends on the size of the cluster—are down, the ability of cluster communication and data to be accessible is compromised. The concept of a quorum is introduced to control the state of the cluster for any such consideration and which available actions can be performed during each state.

A cluster can either be in quorum or out of quorum. If a cluster is in quorum, configuration changes to the cluster can be made to any of the nodes that are up, meaning that they have not been taken over by their partner node. Epsilon can be reassigned to another node within the cluster. This may be desired if the node containing epsilon will be failed over to its partner node for a planned event, such as a clustered Data ONTAP NDU/ANDU.

Moving epsilon increases the odds of the cluster remaining in quorum. Consider, for example, if you have a four-node cluster (node-1, node-2, node-3, node-4) and node-1 is assigned epsilon. Node-1 will be failed over to node-2 while the NDU is being executed; however, epsilon does not fail over. At this point quorum is still maintained; however, the failure of any other node in the cluster will take the cluster out of quorum since the node containing epsilon is already down. Starting in clustered Data ONTAP 8.3, if the node holding epsilon experiences a disruption, epsilon automatically moves to another node of an healthy HA pair within the cluster. With espsilon assigned to node-3, or node-4, then the failure of another node while node-1 is down allows the cluster to remain in quorum, providing that the node that was reassigned quorum is not the node that goes down. The following pictorial representation details this scenario.

Figure 4) HA pair controllers and Cluster quorum

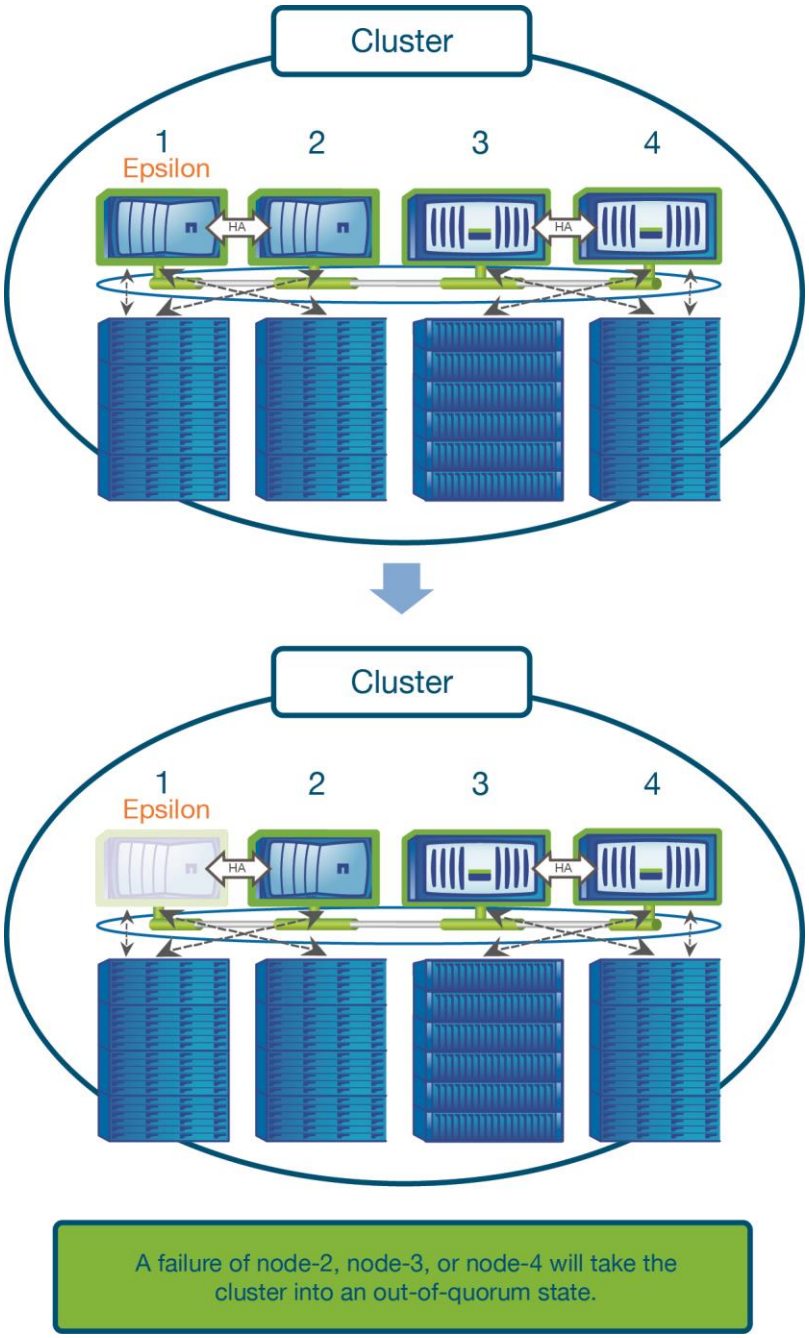
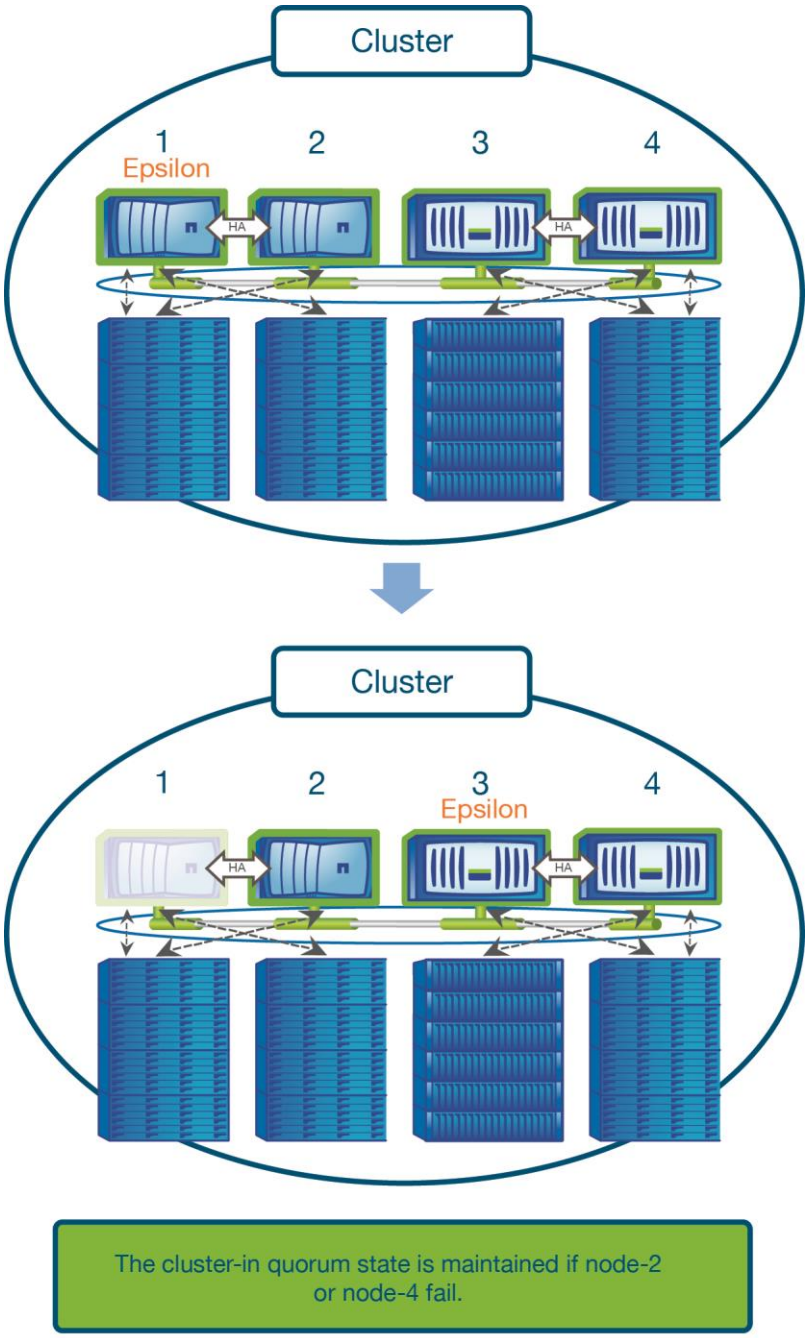


Figure 5) Moving epsilon to maintain Cluster quorum



Having the cluster in quorum means that data is available as expected providing one partner of the HA pair is able to service I/O requests. For example, node-1, node-2, node-3, and node-4 are a four-node cluster; node-1 and node-3 are failed over to their partner-nodes and node-4 holds epsilon. The cluster maintains quorum and all data is available. Network interfaces that are configured with failover groups will fail over to the respective surviving node.

A cluster in an out-of-quorum state has a varied degree of expected behavior. The configuration becomes locked and no changes can be made to the cluster configuration until it is back in quorum. Data is available for storage that is taken over (if attached to a node that has been failed over).

Clusters that have two nodes do not have the concept of epsilon. A special option called cluster HA must be enabled. This allows continuous monitoring of the nodes' state; if either node goes down, the remaining node has full read and write access for all the storage, logical interface, and management capabilities.

HA Pair Controller Configurations and Infinite Volume

Clusters that are using Infinite Volumes utilize the same HA pair controller configurations. HA pair controllers continue to provide HA failover resiliency for both planned and unplanned events for clusters using Infinite Volumes.

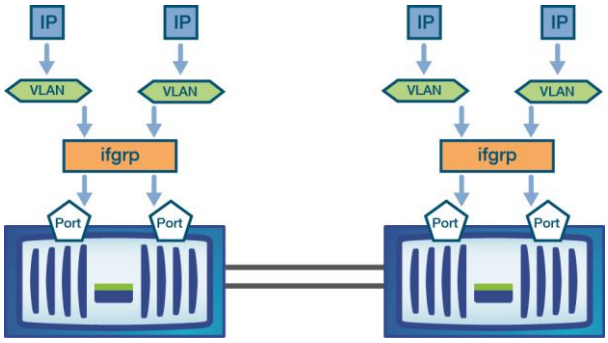
2.7 Network Overview for HA Pair Controllers

Each of the nodes in an HA pair has a separate network interface. The physical ports service requests for both file and block protocols. During normal operation, each node services requests through the network layer to the storage layer and returns data back to the clients independent of the partner node. This section covers the components of the networking layer and failover of the ports when a controller is taken over.

7-Mode Network Port Failover

An HA pair in a system operating in 7-Mode has the physical ports on each node. The network layer can consist of layering interface groups and VLANs. The interface group consolidates several physical ports to act as single port. VLANs can be created for the interface group with an assigned IP address. An IP address can be assigned to a VLAN, an interface group, or a physical port. A corresponding destination on the HA pair is defined for failover. If a VLAN is being used, then the VLAN on each node of the HA pair must have corresponding tags. For more information on 7-Mode networking concepts refer to TR-3802.

Figure 6) Networking hierarchy for system operating in 7-Mode



Clustered Data ONTAP Network Port Failover

One of the architectural values of clustered Data ONTAP is the virtualization of the network interfaces from the storage resources. Each network interface on a node in a cluster can receive an incoming I/O request and forward that request to the respective storage resource.

Each physical port has logical interfaces assigned to it. The logical interface (for NAS) has an IP address. This allows the LIF to be migrated to other physical ports in the cluster without incurring a disruption, referred to as nondisruptive LIF migrate. Each node has several logical interface types: data, cluster, node management, intercluster, and cluster management. The definition for each LIF type is described in the following list.

- Data LIFs are required on a node that serves data over logical interfaces (LIFs).
- Cluster LIFs connect each node to the cluster interconnect network.
- Node management LIFs provide management over a network interface.
- Intercluster LIFs service replication communication between clusters.

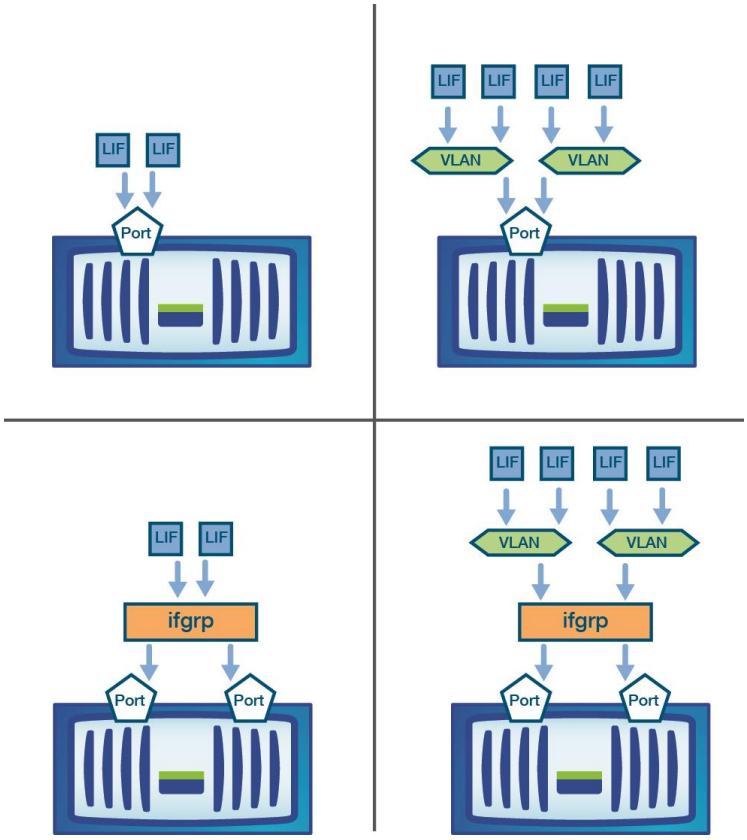
- Cluster management LIFs are used to manage the cluster.

LIFs are logical IP interfaces tied to a home port. The cluster management LIF as well as data LIFs use a cluster-wide failover group by default. Cluster-wide failover allows LIFs to fail over to any of the ports in this group that are still available. In addition, data LIFs have failover groups that are two-node system defined. Data and cluster management ports can be migrated to other ports within the cluster. SAN LIFs cannot be migrated and do not have failover groups. Intracluster LIFs do not have the ability to fail over to another node, but they can fail over to ports on the same node. In general, LIFs fail over automatically under the following conditions:

- A port containing a LIF is set to `down`.
- A node becomes out of quorum.
- Automatic revert is configured on a LIF and the home port status returns to `up`.
- Automatic revert is configured on a LIF and the node returns to quorum.

A physical port can have additional layers of resiliency built onto it by using VLANs and/or ifgrps. The following diagram shows the layer hierarchy.

Figure 7) Networking hierarchy for Clustered Data ONTAP



2.8 HA Pairs and Infrastructure Resiliency

Although the HA pair controller configuration is designed primarily to protect against storage controller failure, it also addresses numerous other single-point-of-failure (SPOF) conditions found in standalone storage controller environments. The following table summarizes components that can trigger failover in an HA pair controller configuration. The SyncMirror® and multipath HA storage features mentioned here are described in later sections.

Table 3) The hardware components that can trigger failover in HA pair controller configurations

Hardware Components	SPOF		How HA Pair Controller Failover Eliminates the SPOF
	Single Storage Controller	HA Pair Controller	
Storage controller	Yes, if not multipath	No	If a storage controller fails and it's not multipath, HA pair controllers automatically fail over to their partner node and serve data from the surviving storage controller.
NVRAM	Yes	No	If an NVRAM card fails, HA pair controllers automatically fail over to their partner node and serve data from the surviving storage controller.
Both CPU fans	Yes	No	If both CPU fans fail, the affected storage controller shuts down gracefully. HA pair controllers automatically fail over to the partner node and serve data from the surviving storage controller.
Numerous NIC cards with virtual interfaces (VIFs)	No	No	If a single network link fails, network traffic is automatically routed over the remaining network links on the storage controller. No failover is required in this situation. If all NIC cards or network connections on a node fail, the HA pair controller automatically fails over to the partner node and serves data from the surviving node. (Applies to FAS systems running Data ONTAP 7.1 and later.) If all NIC cards or network connections on a node fail, the operator can initiate a failover to the partner node and serve data from the surviving storage controller. (Applies to FAS systems running Data ONTAP prior to version 7.1.) **Note:** Customers are advised to use numerous NIC cards with VIFs to improve networking availability for both standalone storage controllers and HA pair controller configurations.
Single NIC card	Yes	No	If a single NIC card or network connection fails, the HA pair controller automatically fails over to the partner node and serves data from the surviving storage controller. (Applies to FAS systems running Data ONTAP 7.1 and later.) If a single NIC card or network connection fails, the operator can initiate a failover to the partner node and serve data from the surviving storage controller. (Applies to FAS systems running Data ONTAP prior to version 7.1.)
Disk shelf (including backplane)	No	No	NetApp disk shelves incorporate dual power supplies and dual fans. A single controller can optionally be configured with dual LRCs/ESHs to provide dual active-active FC-AL loops. HA pair controllers are configured with dual LRC/ ESH modules, which provide redundant FC-AL loops: an active primary path and a failover path to the partner node. In ESH2 or AT-FCX shelves configured with multipathing, there is an active primary path, an active secondary path, and two failover paths to the partner node. Disk shelves are the single most reliable component in a FAS

			system, with an MTBF rating exceeding 2 million hours (228 years).
FC-AL adapter	Yes	No	If an FC-AL adapter connected to the disks owned by the local storage controller node fails and neither SyncMirror nor multipath HA storage is configured, the storage controller initiates a failover to the partner node, which then serves data. (Failover is unnecessary with SyncMirror or multipath HA storage.) If the FC-AL adapter connected to the disks owned by the partner storage controller node fails and multipath HA storage is not configured, failover capability is disabled, but both storage controllers continue to serve data to their respective applications and users with no impact or delay. (Failover would not be disabled and could still occur if needed in a multipath configuration, provided that the other active secondary path was still available.)
FC-AL cable (storage controller to shelf, shelf to shelf)	Yes*	No	If an FC-AL loop breaks and neither SyncMirror nor multipath HA storage is configured, failover to the partner node is initiated. (Failover is unnecessary with SyncMirror or multipath HA storage.) *Applies only to single loop configuration. The multipath (redundant loop) configuration eliminates this SPOF.
LRC/ESH storage controller module	Yes*	No	If an LRC/ESH module fails and SyncMirror is not configured, failover to the partner node is initiated. *Applies only to single loop configuration. Multipath HA storage is supported only on ESH2 (or higher) and AT-FCX-R5 (or higher) and cannot be configured on loops with other controller modules.
Shelf I/O module Required for multipath HA storage solutions	Yes*	No	If a storage controller module fails and neither SyncMirror nor multipath HA storage is configured, failover to the partner node is initiated. (Failover is unnecessary with SyncMirror or multipath HA storage.) *Applies only to single loop configuration. The multipath (redundant loop) configuration eliminates this SPOF.
Power supply (storage controller or disk shelf)	No	No	Both the storage controller and the disk shelves include redundant power supplies. If one power supply fails, the second power supply automatically compensates. No failover is required in this situation.
Fan (storage controller or disk shelf)	No	No	Both the storage controller and the disk shelves include redundant fans. If one fan fails, the second fan automatically provides cooling. No failover is required in this situation.
Interconnect adapter	N/A	No	If an interconnect adapter card fails, failover capability is disabled, but both storage controllers continue to serve data to their respective applications and users.

Interconnect cable	N/A	No	The interconnect adapter incorporates dual interconnect cables. If one cable fails, the heartbeat and NVRAM data are automatically sent over the second cable without delay or interruption. If both cables fail, failover capability is disabled, but both storage controllers continue to serve data to their respective applications and users. The Cluster Monitor then generates warning messages.

3 HA Pairs and Cluster Scalability

Clustered Data ONTAP is a scalable storage architecture that allows a customer to grow the cluster as business needs require, rather than buy and overprovision a complete system. There are several transitions that a customer may encounter when growing the cluster over time. This section looks at each of these transitions and the corresponding concepts as they relate to HA pair controller configurations.

3.1 Single-Node to Two-Node Switchless Cluster (TNSC)

A single-node cluster consists of a single node and does not benefit from HA pair failover or giveback resiliency. A two-node cluster may become necessary for the customer as levels of resiliency are needed. Growing the single-node to a two-node cluster requires the additional cable to be added to the cluster as well as a reboot to the controllers to enable HA and prepare the NVRAM for mirroring of both the node's and partner node's data. This reboot is only required when going from a single-node to a two-node HA pair based cluster.

3.2 Two-Node Cluster (Switched or Switchless) to Four-Node Cluster

A two-node cluster can grow to four or more nodes without any disruption. Prior to your joining additional nodes to the cluster, cluster HA would need to be removed from the two-node system by using the `cluster HA enable` command set. Epsilon will be assigned to the node with the lowest valued site identification number or in the case where a new cluster is being assembled, epsilon will be assigned to the first node in the cluster. Starting in Data ONTAP 8.2 , when nodes are added to the cluster, they are in HA mode; therefore, when HA is enabled on the HA pair there is no required reboot. The nodes are capable of failover and giveback once HA is enabled.

4 HA Pair Solutions for Addressing Individual Business Needs

Best practices evaluate business needs before implementing a new solution. Therefore, equipped with an understanding of how an HA pair controller configuration provides higher data availability, let's analyze the business requirements.

- What are the business needs of the environment?
- What are the business needs in terms of the available HA pair controller solutions?
- How long is the timeout window for application servers?
- How long is the timeout window for clients?
- What single-point-of-failure conditions exist within the data infrastructure?

The following sections assist in choosing appropriate availability solutions based on the needs of the customer's environment.

4.1 Selecting an HA Pair Solution That Fulfills Business Needs

Several different HA pair controller solutions provide various degrees of resiliency. Evaluating each of the solutions with the business needs of the customer provides clarity when choosing an HA pair controller

solution. The following subsections define four tiers to help customers identify their business needs. HA pair controller solutions, features, limitations, and client interaction are covered in section 3.2, “Types of HA Pair Controller Solutions.”

Although HA pair controllers can address many business needs for data availability, other solutions in combination with HA pair controller technology are necessary for complete nondisruptive operations. This is a critical planning step in evaluating any high-availability solution and something to consider when architecting a high-availability system.

Tier 1: Mission Critical

Mission-critical environments enable services that are in high demand and that cost the customer significant loss of revenue when an outage occurs. Examples include online transaction processing (OLTP), batch transaction processing, and some virtualization and cloud environments. This tier of data availability prioritizes I/O response to foreground (client application) traffic to facilitate having dependent applications remain functional. Prioritizing foreground I/O over corrective I/O in degraded situations increases the time needed to complete corrective actions. This increases the risk of encountering additional failures in the system before completing a corrective action; for example, encountering an additional drive failure before an existing reconstruction operation can complete.

Tier 2: Business Critical

Business-critical environments are often subject to compliance requirements, and although maintaining client access to the storage system is important, the loss of data would be severely detrimental to the customer. No customer likes to lose data, but these customers are under legal obligations and are subject to significant penalties if they are found to be noncompliant. This could also be a configuration that protects a company’s intellectual property. Examples include medical records, software source code, and e-mail. This tier prioritizes corrective I/O while balancing foreground I/O. Prioritizing corrective I/O over foreground I/O in degraded situations increases the impact on foreground I/O performance.

Tier 3: Repository

Repository environments are used to store collaborative data or user data that is not critical to business operations. Examples include scientific and engineering compute data, workgroup collaborations, and user home directories. This tier is the middle ground that balances foreground operations with corrective actions if they are needed. Defaults are normally appropriate for these configurations.

Tier 4: Archival

Archival environments are subject to a large initial ingest of data (write), which is seldom accessed. System utilization is not expected to be very significant. Because the data is seldom accessed, it is important to fully leverage subsystem features that exercise that data for continued integrity. Given that the priority is maintaining data integrity, these configurations prioritize corrective I/O and minimize completion time for corrective actions. Examples include backup and recovery, archiving, near-line, and reference data.

Table 4) Recommended HA pair solutions based on business needs

Recommended					
HA Pair Solution	**Resiliency Features**	**Tier 1**	**Tier 2**	**Tier 3**	**Tier 4**
HA	Use HA pairs to maintain resiliency during hardware and software system failures.	Yes	Yes	Yes	Yes

Multipath HA	Use multipath HA to maintain resiliency during cable, port, or HBA failures.	Yes	Yes	Yes	Yes
SyncMirror	Use multipath HA to maintain resiliency during cable, port, or HBA failures.	Yes	Yes	Yes	No
MetroCluster (fabric)	Use MetroCluster to maintain resiliency during a site failure caused by external disasters or circumstances.	Yes	Yes	No	No

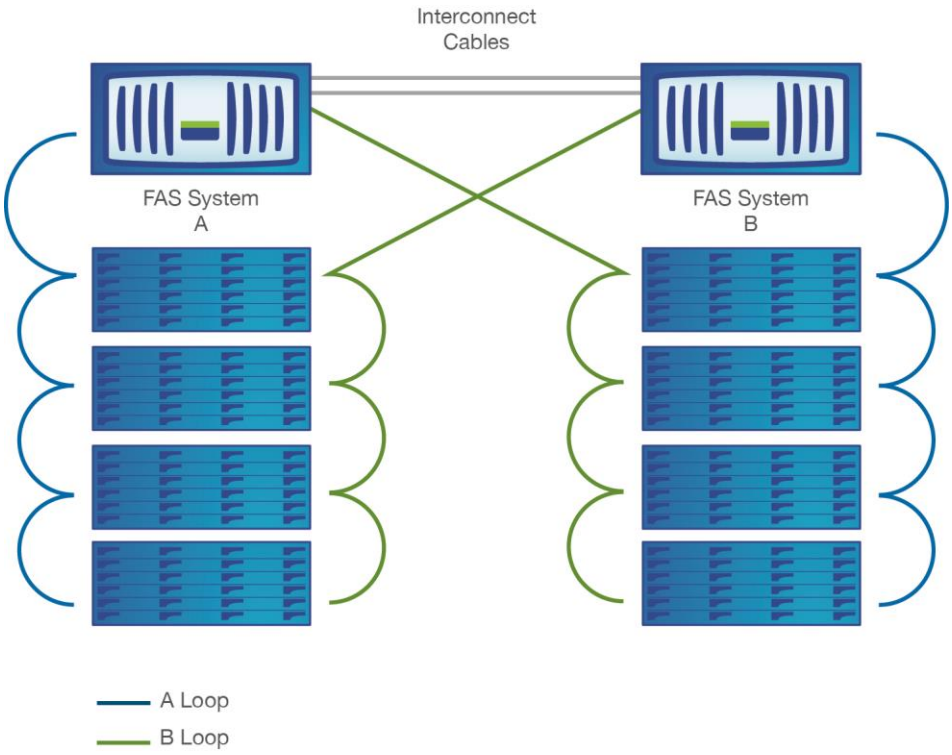
4.2 Standard HA Pair Controllers

A standard HA pair controller configuration contains two sets of Fibre Channel, SATA, and/or SAS disk shelves: one set for the local storage controller (local node) and the other set for the partner storage controller (partner node). In this solution, a single copy of data is used in serving clients. The two nodes are connected to each other through matching InfiniBand (IB) and NVRAM adapter cards. (Older FAS systems may use either Troika or ServerNet adapters.) This pair of interconnect cards enables the surviving node to serve data on the disks belonging to its failed partner node. Each node continually monitors its partner, mirroring the other's NVRAM data.

Configuration Requirements for This HA Pair Controller Configuration

- Identical InfiniBand (IB) and/or NVRAM adapter cards, along with their appropriate cables, must be installed on each storage controller.
  - The interconnect cards are responsible for transmitting the heartbeat signal and NVRAM synchronization between the partner nodes, so the firmware version on the cards must be identical, and they must be installed in the proper slots in the storage controller. Please refer to the appropriate system configuration guide on the NetApp Support site for slot assignments.
- Both nodes in the configuration must be attached to the same network and the network interface cards on each storage controller must be configured correctly.
  - If the nodes are installed in different networks, takeover cannot take place when one of the partner nodes fails. Attaching HA pair controllers to an unstable network causes significant delay when the surviving node attempts to take over the failed partner. Details concerning downtime analysis are covered in section 4.1, "Best Practices to Minimize Client Disruption."
- The `cluster` service must be licensed and enabled on both nodes.

Figure 8) Hardware and cabling overview for standard HA pair controller configuration



Advantages

- Data processing continues if one of the nodes fails and cannot reboot.
- Data processing continues if an FC-AL adapter on a node fails.
- There is a smaller data processing disruption due to a panic situation when the `cf.takeover.on_panic` option is enabled on the storage controllers. This option is covered in section 4.1, “Best Practices to Minimize Client Disruption.”
- System maintenance can be performed with minimal interruption to the data service.
- This system maintenance may include both hardware and software. Section 5, “Nondisruptive System Upgrade,” covers the Data ONTAP nondisruptive upgrade process in detail.

4.3 Multipath HA Pair Controllers

Standard nonswitched HA pair controller configurations employ a single path from the controller to the storage shelves. Consequently, a cable break, Fibre Channel adapter/port failure, or shelf I/O module failure triggers a failover, which can affect system availability and/or performance consistency. Similarly, when a failover occurs, the surviving partner node relies upon a single secondary path. If that path fails, controller takeover isn’t possible.

The multipath HA storage solution significantly improves data availability and performance consistency by providing multiple redundant paths to the storage shelves. This not only prevents unnecessary failovers caused by storage-related faults, it also increases bandwidth to the disk shelves. Multipath support is provided with Data ONTAP 7.1.1, 7.2.1, and later; so no additional software is necessary.

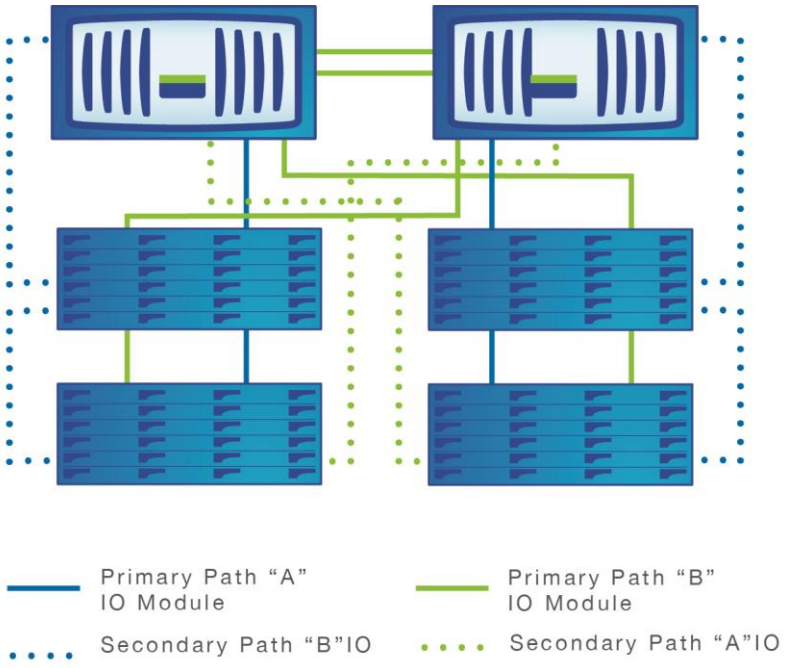
By providing independent primary and secondary paths from each controller to its storage shelves, if a shelf controller or cable fails on one path, the node’s data service automatically fails over to the redundant path and the node remains up without disruption.

Configuration Requirements for This HA Pair Controller Option

- Supported and highly recommended for all nonswitched HA pair controller deployments.
  - The fabric-attached MetroCluster configuration inherently provides multipath functionality.
- Each storage controller must be populated with sufficient Fibre Channel and/or SAS adapters/ports and cables to configure four loops: primary and secondary local, and primary and secondary partner.
  - Primary and secondary loops must use separate Fibre Channel and/or SAS adapters.
- FAS6000 and FAS3070 controllers are supported with Data ONTAP 7.2.1 and later.
- FAS3100 controllers are supported in Data ONTAP releases that support these controller families.
- Storage shelves must be equipped with controller modules providing autotermination functionality.
  - ESH2 and AT-FCX (RoHS compliant).
- Requires software-based disk ownership (SANOWN).

Instead of terminating a disk shelf at the end of the loop, connect the redundant Fibre Channel cable.

Figure 9) Hardware and cabling overview for multipath HA pair configuration



Advantages

- Can dramatically improve reliability with minimal additional investment.
- Avoids unnecessary failovers by providing controller-to-storage-shelf data path redundancy.
- Can increase performance by providing secondary path to storage.
- Dual primary paths help prevent failover due to storage issues; if a storage failure blocks a storage path, the redundant primary path provides path availability.
- Dual secondary paths help enable successful takeovers; if a storage failure blocks a secondary path during failover, the redundant secondary path enables path availability.
- Allows shelf controller or cable replacement maintenance operations without incurring failover.
- Ideal for failover-sensitive environments such as CIFS, FCP, and iSCSI.
- Complements the stretch MetroCluster configuration.

Table 5) Comparison of multipath HA pair configuration options

Avoids Storage Controller Failover Due to	Multipath HA Pair	Multipath HA Pair with SyncMirror
Single controller-to-shelf cable failure	Yes	Yes
Single intershelf cable failure (primary)	Yes	Yes
Dual intershelf cable failure	Yes	Yes
Shelf module hardware or firmware failure	Yes	Yes
Disk HBA or port failure	Yes	Yes
Several disk failures (3+)	No	Yes
Shelf failure	No	Yes
Failure of both loops in the dual-active configuration, entire shelf failure, or multishelf failure	No	Yes

4.4 HA Pair Controllers with SyncMirror

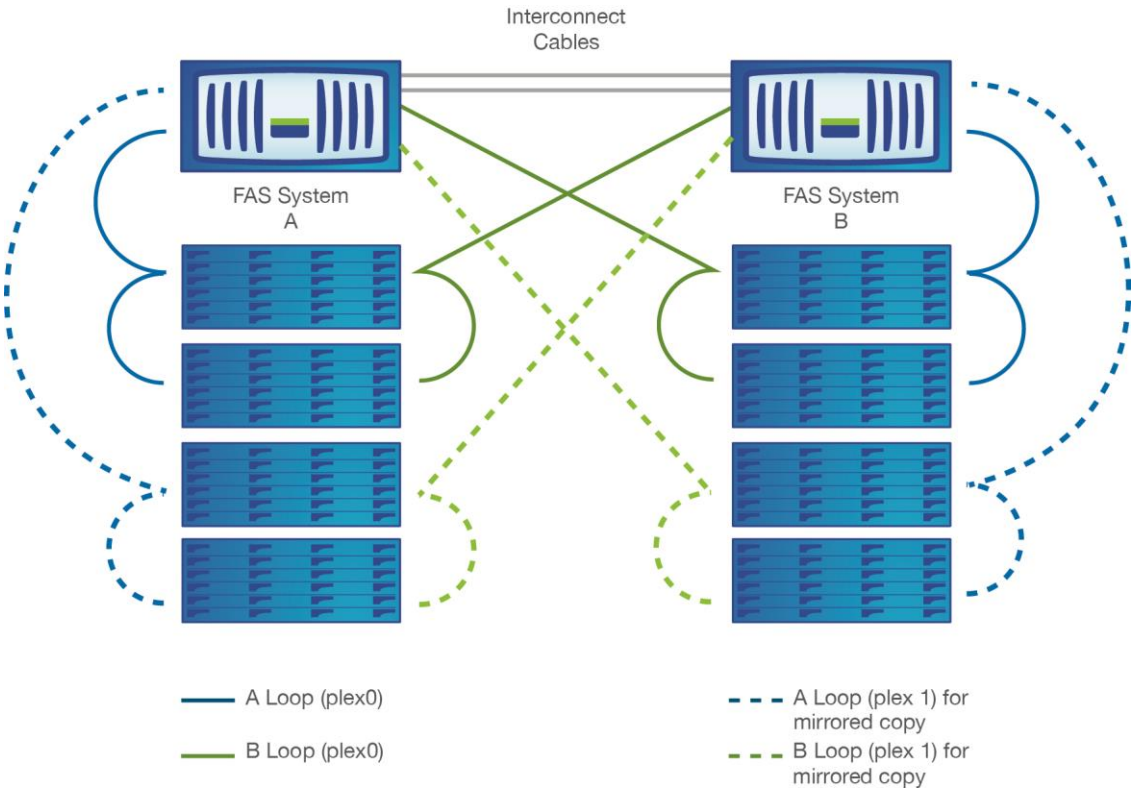
The mirrored HA pair controller configuration includes all of the basic features provided in the standard HA pair controller configuration. In addition, the mirrored HA pair controller contains two complete copies of data volumes, aggregates, or file systems specified as mirrored volumes or file systems within the HA pair controller. These copies are called *plexes* and are continuously and synchronously updated with SyncMirror each time Data ONTAP writes data to disk. Plexes are physically separated from each other across different groupings of disks and Fibre Channel and/or SAS adapters. They should also be connected to different power circuits and in different racks.

Configuration Requirements for This HA Pair Controller Option

- To support failure separation of plexes, spare disks are divided into two pools, pool0 and pool1. One plex of the mirrored volume is always created from disks in pool0, the other plex from disks in pool1.
- The FAS2000, FAS3000, and FAS6000 controller families use only software-based disk ownership and pool selection.
- The controllers also feature configuration checking functionality that detects and validates slot assignments as well as provides proper separation between disk pools.
- Software-based disk ownership can be used, if enabled, for the other FAS systems as well.
- The `syncmirror_local` software must be licensed and enabled on both nodes within the HA pair controller configuration.
  - All software licenses on both nodes in the HA pair controller configuration must be identical to provide uninterrupted service availability when failover occurs.

Sufficient spares must exist in each pool to accommodate disk failure.

Figure 10) Hardware and cabling overview for HA pair controller configuration with SyncMirror



Advantages

- Mirrored data on the HA pair controller configuration survives multiple disk failures in a RAID group.
  - Any number of disk failures, if the failure is isolated to a plex
  - Any combination of three disk failures when using RAID 4
  - Any combination of five disk failures when using NetApp RAID-DP® technology
- Mirrored data on the HA pair controller configuration survives loop failure to one plex. The other plex continues to serve data without the need for failover.

4.5 Fabric MetroCluster

The fabric MetroCluster configuration is an extended version of the mirrored HA pair controller option, achieved by adding two pairs of Fibre Channel switches between the partner FAS systems: One pair is placed on the local side and the second pair is placed on the remote side. The local FAS system is connected to the local switches; the remote FAS system is connected to the remote switches.

The fabric MetroCluster solution offers the highest data resiliency. This HA pair controller configuration solution provides site protection. When a site outage occurs, data processing service continues uninterrupted at the remote location.

MetroCluster can also be configured without switches. Please contact your NetApp sales representative for the cabling requirements of this MetroCluster configuration.

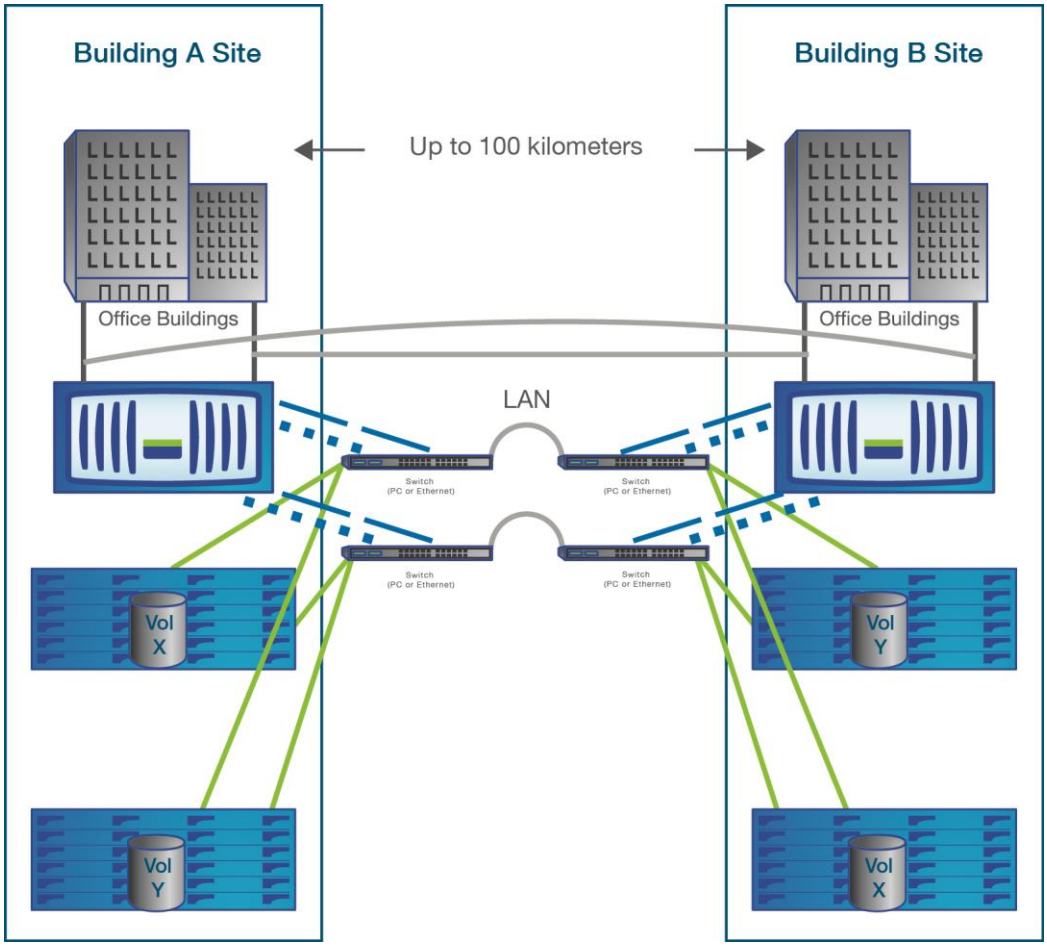
Configuration Requirements for This HA Pair Controller Option

- Data ONTAP 7-Mode and clustered Data ONTAP 8.3
- Each failover pair of nodes must use four switches, two per cluster side.
- Switches must be Brocade model 5100, 300E, 200E, 3200, 3800, 3250, or 3850.

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.

- Each node must have sufficient Fibre Channel ports to connect to both switches at its location. This can be achieved either through embedded ports (3xxx or 6xxx) or through redundant Fibre Channel HBAs. Each node must have the VI-MC cluster adapter installed.
- The `syncmirror_local` and `cluster_remote` services must be licensed and enabled on both local and partner nodes. The `cluster_remote` service is the key to MetroCluster, as compared with a SyncMirror-only solution. It enables the `cf forcetakeover -d` command, which allows a takeover to occur without a quorum of disks available. This is not possible in a SyncMirror-only environment.
- 7-Mode deployment and now available starting with clustered Data ONTAP 8.3

Figure 11) Hardware and cabling overviewfor fabric MetroCluster configuration



Advantages

- Enables significantly extended distances between the HA pair controllers' nodes.
- Fabric MetroCluster configurations support distances up to 100 kilometers.

- Nonswitched, or “stretch,” MetroCluster deployments configured with currently shipping disk shelves employing 2Gb connections are limited to 500 meters with OM3 cables, 300 meters with standard FC cables.
- A key single point of failure is eliminated because disk shelves are dual-attached to the Fibre Channel switches, providing several data paths. Each path is connected to a separate switch, eliminating the switch as a single point of failure.
- Because it can survive site outage, the fabric MetroCluster solution is an ideal disaster recovery option.
- The standard HA pair controller configuration can be easily upgraded to either the fabric or the stretch MetroCluster configuration.

Table 6) Characteristics and distance limitations of HA pair controller interconnect adapters

Interconnect Adapter Type	General Characteristics
InfiniBand (IB) adapter	Employs an optical cable up to 500 meters long. Requires two SC/SC cables and GBICs for distances over 30 meters (purchased separately).
VI adapter (Revision B)	Employs an optical cable up to 500 meters long. Requires an LC/LC cable for distances over 30 meters (purchased separately).
VI-MC adapter	Provides an interconnect solution up to 100 kilometers. Connects directly to Fibre Channel switches. Requires two SC/SC cables (purchased separately).

5 Understanding Client Impact

Client disruption, although minimal, may still occur in the HA pair controller environment during the takeover and giveback process. This section addresses known issues surrounding client disruption and reviews potential mitigations.

When one node in an HA pair controller configuration encounters an error and stops processing data, its partner detects the failed or failing status of the partner and takes over all data processing from that controller. The amount of time it takes a node to detect a failure on the partner is defined by using the `cf.takeover.detection.seconds` option. Hardware-assisted takeover can happen within a second of a failure being detected.

If the partner is confirmed down, the surviving storage controller immediately initiates the failover process to assume control of all services from the failed storage controller. This period is referred to as *takeover time* for clients.

After the failed storage controller is repaired, the administrator can return all services to the repaired storage controller by issuing the `cf giveback` command on the surviving storage controller serving all clients. This command triggers the giveback process, and the repaired storage controller boots when the giveback operation completes. This process is referred to as *giveback time for clients.*

Therefore, the takeover/giveback period for clients is simply the sum of the takeover time plus the giveback time. The following describes the process in equation format.

- Takeover time = time to detect controller error (mailbox disks not responding) and initiate takeover + time required for takeover to complete (synchronize the WAFL logs).

- Giveback time = time required to release the partner's disks + time to replay the WAFL log + time to start all services (NFS/NIS/CIFS, etc.) and process export rules.
- Total time = takeover time + giveback time.

For clients or applications using stateless connection protocols, I/O requests are suspended during the takeover/giveback period, but they can resume when the takeover/giveback process completes. For CIFS, sessions are lost, but the application may—and generally will—attempt to reestablish the session.

This takeover/giveback time is critical and has been decreasing with newer Data ONTAP releases. In some instances, total time can be very long if the network is unstable or the storage controller is configured incorrectly. Consequently, minimize the total takeover/giveback time by adopting the general best practices described in section 4.1.

5.1 Best Practices for Minimizing Client Impact

Monitor Network Connectivity and Stability

Unstable networks not only affect total takeover/giveback times, they adversely affect all devices on the network in various ways. NetApp storage controllers are typically connected to the network to serve data, so if the network is unstable, the first symptom is degradation of storage controller performance and availability. Client service requests are retransmitted many times before reaching the storage controller, thereby appearing to the client as a slow response from the storage controller. In a worst-case scenario, an unstable network can cause communication to time out, and the storage controller will appear to be unavailable.

During takeover and giveback operations within the HA pair controller environment, the storage controllers will attempt to connect to numerous types of servers on the network, including Windows® domain controllers, DNS, NIS, LDAP, and application servers. If these systems are unavailable or the network is unstable, the storage controller will continue to retry establishing communications, thereby delaying the takeover or giveback times.

Network monitoring is one of the most critical tasks for the mission-critical environment. Various tools are available to validate network stability between clients, servers, and storage controllers. Maintain a robust network infrastructure by implementing a policy of regular review of network health.

Validate VLAN Configuration

Verify that both storage controllers can communicate to all required VLANs within the network environment. When a node failure occurs and the partner takes over the failed node, the VLANs on the failed node are failed over to the partner. For clustered Data ONTAP, a VLAN may be failed over to other nodes in the cluster, depending on the defined failover groups. Because the surviving node is serving the failed node's clients, it requires connections to the same VLANs the failed node accessed prior to the failover. This step also involves confirming the VLAN configuration on the switch into which the storage controllers are connected.

Use Interface Groups to Provide Redundancy and Improve Network Availability

The interface group (IFGRP) configuration enables network interface failover and can prevent takeover from occurring at the system level when a single network interface or connection fails. The interface group option provides network layer redundancy; compared to system-level failover, network-layer failover causes almost no impact on clients. HA pair controllers without interface groups experience system-level failover every time a single network interface fails, so NetApp highly recommends interface groups in the HA pair controller environment.

There are two types of interface groups: single-ifgrps and multi-ifgrps. Using both interface group configurations together gives the highest degree of resiliency. Configure all interface groups and VLAN assignments symmetrically on both storage nodes, including all vFiler® unit instances, to provide

uninterrupted network resource availability. For clustered Data ONTAP, an interface group may fail over to a partner node or to other nodes within the cluster. Make sure the interface groups are configured symmetrically with the ports defined in the failover group.

Monitor Disk Performance and Health

Data ONTAP automatically performs background disk health checking when the AutoSupport™ feature on the storage controllers is enabled. AutoSupport is a mechanism that proactively monitors the health of your system and, if enabled, automatically sends e-mail messages to NetApp Technical Support, your internal support organization, and a support partner.

AutoSupport is enabled by default and can be disabled at any time.

If the AutoSupport feature must be disabled, review the `/etc/messages` file on a regular basis using the keyword “error” to monitor the health of the storage controller and disk drives. If any disks appear suspicious, contact NetApp Support. Failed disks can cause giveback operations to fail or to work incorrectly. As a general rule, NetApp strongly recommends removing all failed drives from the system at the earliest opportunity.

Monitor Storage Shelf Module Performance and Health

As stated previously, Data ONTAP automatically performs comprehensive background system health checking, so NetApp strongly recommends enabling the AutoSupport feature on all storage controllers. AutoSupport provides automated remote system health monitoring by the NetApp Global Support Center.

Verify That All Settings in the /etc/rc Files Are Correct and Consistent

Confirm that all relevant DNS, NIS, VIF, network interface, VLAN, and Ethernet settings are identical where appropriate between the partner storage controllers. An automated script can be written based on your environment settings to verify the `/etc/rc` files. If these settings are inconsistent between the partner nodes, clients may not be able to communicate with the surviving storage controller in takeover mode, thereby causing extended client I/O suspension.

Use Data ONTAP commands such as `setup` and `create` or use the FilerView® tool to change the system configuration, allowing Data ONTAP to safely and properly modify the `/etc/rc` file. Changing the file manually introduces the potential for improper storage controller configuration, so tools such as FilerView simplify the process of modifying system settings.

Use the Multipath HA Storage Configuration Option

The multipath HA storage configuration dramatically improves system reliability by providing a secondary path to storage. This not only eliminates unnecessary controller failovers by providing controller-to-storage shelf data path redundancy, it also increases performance.

MultiStore Instances Created in the HA Pair Controllers

For an HA pair controller configuration, as the number of vFiler instances increases for a system, the period of time for a failover and giveback to complete may increase. However, the number of vFiler unit instances will not prevent the successful completion of a planned or unplanned failover or giveback event on an HA pair. Also, it is important to confirm that both nodes within the controller can communicate with the network servers required by all vFiler unit instances, including DNS servers, NIS servers, and Windows domain controllers. This enables all vFiler unit instances to continue to function properly in takeover mode.

Client and Application Timeout Windows

Whenever possible, adjust the application and client timeout windows to exceed the failover/giveback time.

Enable the cf.takeover.on_panic Option

The `cf.takeover.on_panic` option is enabled by default. This option allows the partner to take over during an event that causes a panic. Disabling this option may incur a pause in I/O to clients while the core is dumped.

The autogiveback after cf.takeover.on panic is enabled by default on Data ONTAP 8.1.1 (and later releases). This allows a giveback to be done promptly once both nodes are in a state for giveback to proceed. Allowing the giveback to proceed automatically returns the HA pair controllers into an active-active HA state.

For All Storage Controllers and Disk Shelves, Connect the Individual Redundant Power Supplies to Separate Power Circuits

NetApp storage controllers and disk shelves are equipped with redundant power supply modules. By connecting each of the two power supplies within a device to separate power circuits or PDUs (power distribution units), devices are protected from single circuit power failures. NetApp strongly recommends this practice, which applies to all mission-critical equipment.

Verify That Both Storage Controllers Provide Identical Service Licenses

This enables all active service functionality to be maintained following failover.

Properly Terminate All Fibre Channel Adapters on Both Nodes in the HA Pair Controller

Unterminated or improperly terminated Fibre Channel adapters prolong boot time. When a storage controller boots, it validates the connectivity and termination of all Fibre Channel connections. Improperly terminated Fibre Channel adapters delay this validation process, thereby extending the total boot time. Boot times can be reduced by as much as 25 seconds by installing loopback adapters in all unused Fibre Channel ports on each controller. Although unterminated ports do not affect actual takeover and giveback times, the prolonged boot time they cause delays the stage at which giveback can begin. However, clients are not affected during this period. Similarly, if a node reboots without a takeover occurring, boot time can also be delayed by as much as 25 seconds.

Note: Do not terminate open Fibre Channel ports on live production controllers. Add loopback adapters either during initial storage controller installation or within a window of scheduled downtime.

Thoroughly Test Newly Installed HA Pair Controllers Before Moving Them into Production

General best practices require comprehensive testing of all mission-critical systems before introducing them to a production environment. HA pair controller testing should include not only takeover and giveback or functional testing, but performance evaluation as well. Extensive testing validates planning.

Test Production HA Pair Controllers on a Regular Schedule

Implement a regularly scheduled test of all critical functionality of the active-active controller configuration. Design a test scenario, document test procedures, and then map out the expected results. A well-designed test plan validates system recovery procedures. Follow the plan carefully during test execution and use the test as a training exercise for the IT staff. Use test results to modify the recovery procedures

as necessary. Practice makes perfect; not only does this testing exercise enable preparedness of both staff and systems when failure occurs, it also sets expectations for the user community.

Enable the PortFast Option

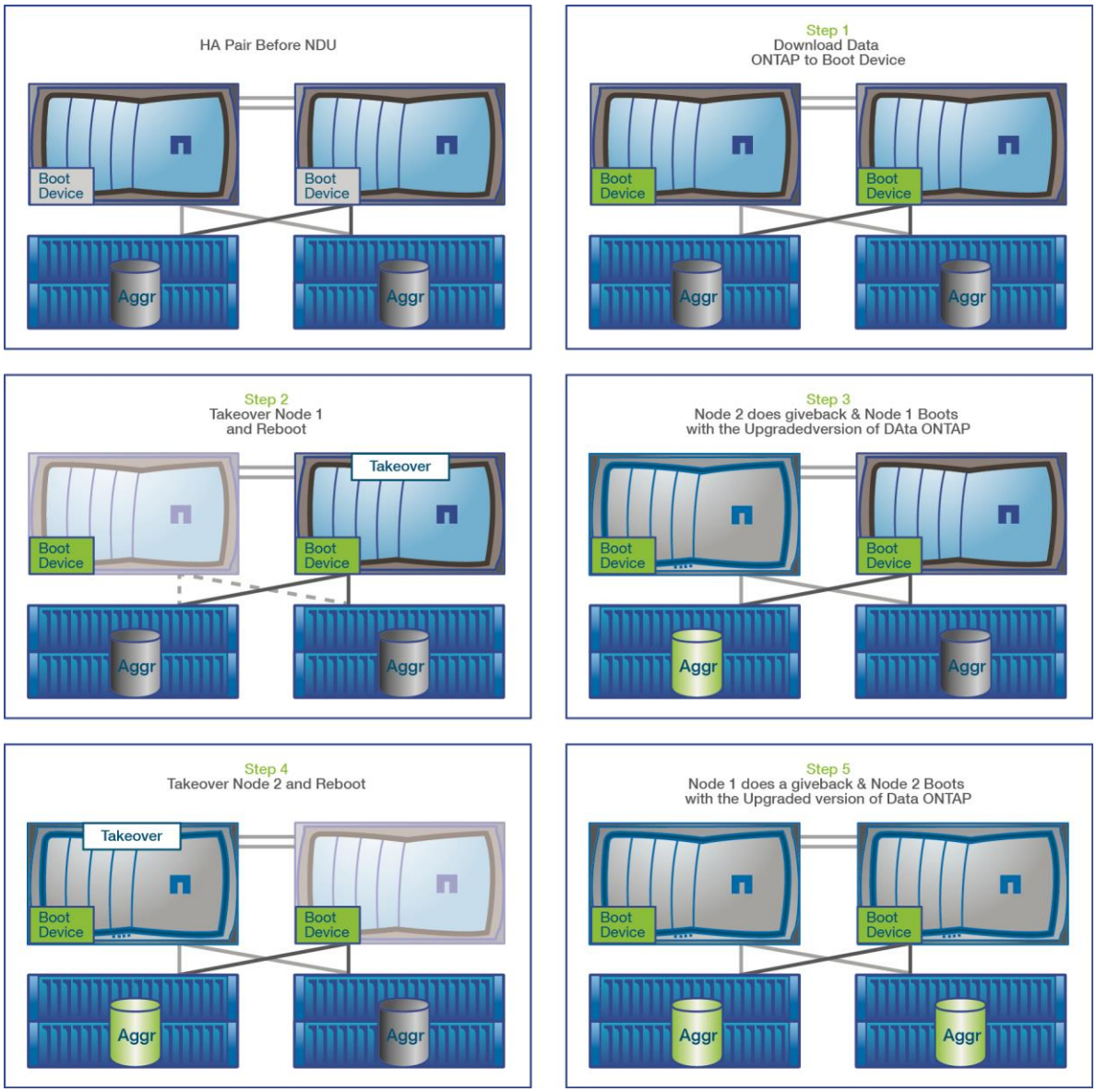
Whenever possible, enable the `PortFast` option on the clients and the storage controller ports on the switch. This option significantly reduces the spanning tree time and in turn reduces the takeover and giveback time.

6 Nondisruptive Upgrade (NDU) for HA Pair Controller Configurations

6.1 Nondisruptive Upgrade Overview

A nondisruptive system upgrade (NDU) is a mechanism that utilizes HA pair controller technology to minimize client disruption during an upgrade of Data ONTAP or controller firmware. This procedure allows each node of HA pair controllers to be upgraded individually to a newer version of Data ONTAP or firmware. Minor-release NDU was first supported in Data ONTAP 6.5.3. Major-release NDU is supported from Data ONTAP 7.0.6 or 7.1.2 to 7.2.3 and higher. Details of version compatibility and steps entailed in major-release upgrades are discussed in the following sections. Refer to the “Data ONTAP Upgrade Manual” for additional information. Upgrading Data ONTAP on a FAS system involves these key steps: install the system files, download the new operating system to compact flash, and reboot. With NDU, a takeover of the data service process belonging to the node being upgraded is performed prior to reboot, thereby minimizing disruption of client I/O. Following the reboot a giveback is initiated, returning the data service to the newly upgraded node.

Figure 12) NDU steps for an HA pair



Client I/O requests are suspended during takeover and giveback operations, which may lead to client or application disruption, depending on the factors outlined in the next two subsections. Although all clients incur a suspension of I/O during takeover/giveback, some may encounter application or protocol disruption, depending on the protocol, the length of the takeover/giveback, and the characteristics of the particular application. The rest of this section examines the effect on client connectivity to the storage controller with various protocols.

Client Considerations

- **CIFS.** Leads to a loss of session to the clients, and possible loss of data. Consequently, upgrades should be scheduled and preceded by requests for CIFS users to drop their connections voluntarily. Any remaining CIFS sessions can be ended by issuing the `cifs terminate -t` command prior to the reboot, or by using `reboot -t`. Loss of CIFS sessions is a problem common to Windows clients for all storage vendors.

- **NFS hard mounts.** Clients will continue to attempt reconnection indefinitely; therefore, controller reboot does not affect clients unless the application issuing the request times out waiting for NFS responses. Consequently, it may be appropriate to compensate by extending the application timeout window.
- **NFS soft mounts.** Client processes continue reconnection attempts until the timeout limit is reached. Although soft mounts may reduce the possibility of client instability during failover, they expose applications to the potential for silent data corruption, so they are advisable only in cases in which client responsiveness is more important than data integrity. If TCP soft mounts are not possible, reduce the risk of UDP soft mounts by specifying long retransmission timeout values and a relatively large number of retries in the mount options (for example, `timeo=30, retrans=10`).
- **FTP, NDMP, HTTP, backups, restores.** State is lost and the operation must be retried by clientA.
- **Applications (for example, Oracle®, Exchange).** Generally, if timeout based, application parameters can be tuned to increase timeout intervals to exceed Data ONTAP reboot time as a means of avoiding application disruption. See the application's best-practices guide for details.

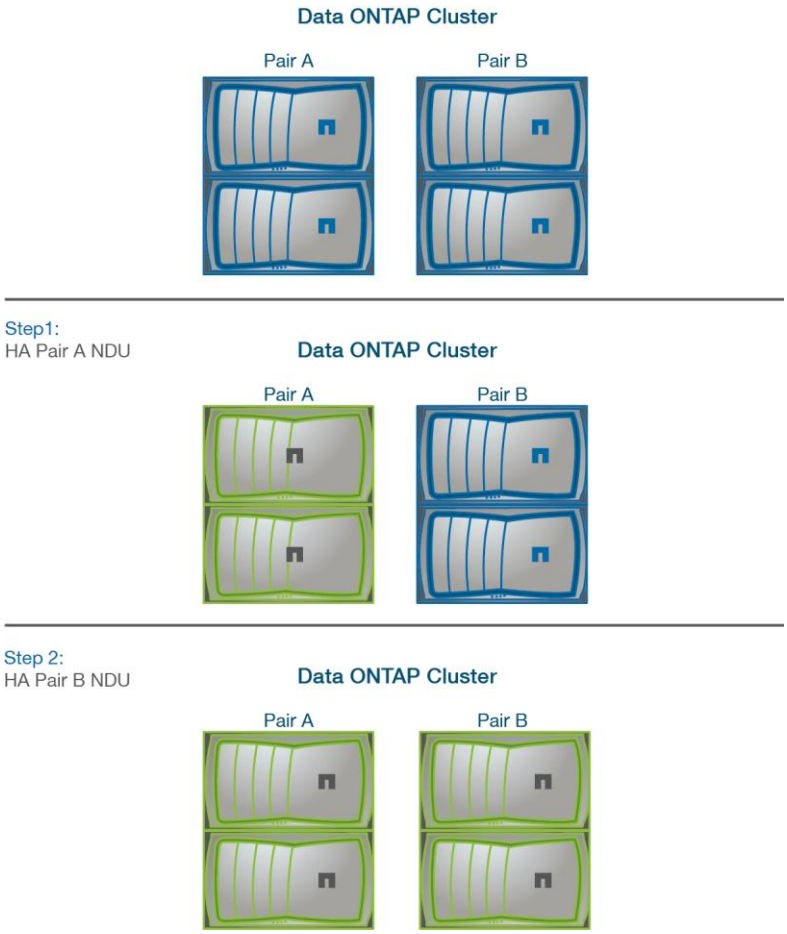
Considerations for FCP or iSCSI Environments

- For major-version nondisruptive upgrades within the SAN context, refer to the instructions listed on the NetApp Support site.
- Nondisruptive upgrade is not currently supported with Solaris or HP-UX hosts using Symantec™ Veritas™ Storage Foundation 3.5 in an FCP or iSCSI environment.
- Major-version upgrades are not supported for Windows client machines using DSM integrated with NetApp SnapDrive® data management software. The version of DSM shipping with the WAK 4.0 or later is necessary.

Rolling Upgrades

Rolling upgrades allow a single HA pair to be upgraded while still keeping the cluster in quorum. The current standard to keep the cluster in quorum allows only one HA pair to be upgraded at a time during a rolling upgrade. For example, a cluster having two HA pairs, consisting of four nodes, requires the NDU procedure to be executed on a single HA pair to maintain quorum. The entire rolling upgrade process on a four-node cluster requires upgrading each of the HA pairs individually until all nodes in the cluster are running the upgraded version of Data ONTAP. Rolling upgrades also allow the cluster to continue serving data while the nodes in the cluster are running mixed versions of Data ONTAP. Clusters with mixed versions of Data ONTAP are not intended to run for long periods of time. NetApp recommends completing a rolling upgrade within a few hours.

Figure 13) Rolling upgrade steps for systems operating in clustered Data ONTAP



The Upgrade Advisor tool, available on the NetApp Support site, should be used before performing an NDU for any system that has AutoSupport enabled. Upgrade Advisor uses the latest AutoSupport logs from the system to be upgraded. A customized step-by-step NDU procedure is created based on the system configuration and the version of Data ONTAP planned for the upgrade.

6.2 Requirements for Nondisruptive Upgrade

- Applicable only to HA pair controllers.
- Supported only for flash-booted FAS systems.
- Must perform a clean shutdown of the storage controller using the lower version of Data ONTAP.
  - The administrator must follow NDU procedures exactly.
  - Clean controller shutdown forces a WAFL consistency point and empties the NVRAM to enable data synchronization.
- Both the lower and the higher Data ONTAP releases must provide major-release NDU support.
- The `cluster` service must be licensed and enabled on both nodes throughout the NDU process.
- Use the `cf takeover -n` command (be sure to include the `-n` option).
  - Use this command on the higher-release storage controller to the lower-release storage controller.

6.3 Data ONTAP Support Matrix for Nondisruptive Upgrade

- Minor-release NDU (starting with Data ONTAP 6.5.3)
  - Supported for Data ONTAP 6.5.1 (and later)
  - For example: Data ONTAP 7.0 to 7.0.1, or 7.2 to 7.2.1, or x.y to x.y.z
- Major-release NDU
  - Data ONTAP 7.0.6 to 7.2.3 (and later)
  - Data ONTAP 7.1.2 to 7.2.3 (and later)

6.4 Limits for Nondisruptive Upgrade

Nondisruptive upgrades have a set of defined limits to optimize the necessary takeover or giveback, allowing it to complete in the shortest period of time. NDU success is based on meeting the client expectation for I/O being returned. Therefore it is essential for a planned takeover or planned giveback to complete in the shortest amount of time, and also for the client to have the appropriate settings configured to account for the worst-case time.

The defined limits for NDU on the controller are verified limits to aid success of the NDU. An increase in any one of these limits may not result in a definite failure, but it does introduce additional risk that may result in increased outages during a takeover or giveback.

6.5 Nondisruptive Upgrade Best Practices

- Confirm that the Data ONTAP upgrade path supports the NDU process.
- Verify that the storage controllers meet all requirements.
- Confirm that the identical version of Data ONTAP is downloaded to both storage controllers.
- Use the `cf status` or `storage failover show` command to confirm that the cluster is enabled before beginning the NDU process.
- The CIFS sessions will be disrupted, so inform CIFS clients that an outage will occur when performing the NDU process. End any remaining CIFS sessions on the storage controllers before starting the NDU process. Remember to restart CIFS sessions when the NDU operation has completed.
- The takeover and giveback operations enable continuous data service throughout the upgrade process. However, when one node is halted and the partner node serves data for both, NVRAM is not mirrored and automatic takeover is disabled, thereby introducing the possibility of loss of data logged in NVRAM if a controller failure event occurs during this time. Therefore, once the first controller is upgraded, perform all remaining operations as quickly and carefully as possible.
- Perform the NDU operation during an off-peak period of system activity. Note that following the successful NDU process, disk firmware is upgraded automatically in the background while disk health check monitoring continues to be performed. The disk firmware upgrade process is a low-priority background thread that uses idle CPU time. All failed disks should be physically removed prior to the NDU. As a general rule, NetApp strongly recommends removing failed disks from the system at the earliest opportunity.

6.6 Nondisruptive Upgrade Caveats and Considerations

CIFS is disruptive for NDU due to the nature of the protocol. For HA pairs in clustered Data ONTAP, there are nondisruptive solutions for volumes servicing CIFS clients during NDU. The use of volume move and

LIF migrate is nondisruptive for CIFS. Nondisruptive LIF migrate is supported for SMB version 2.0 and above. For customers using SMB 1.0, there may be a disruption during the LIF migrates. By relocating data traffic and volumes to alternate nodes in the cluster, the data servicing the CIFS clients remains online and available while an NDU is executed on the original home node of the volumes. This process requires available space and processing power to be available on other nodes in the cluster. The GUI is limited in providing an interface for the necessary steps for NDU. System Manager can be used for some parts of the NDU process, but it may be necessary to interface directly with the command line interface (CLI) to complete the NDU procedure.

7 Command Line Interface (CLI)

Clustered Data ONTAP introduces a more modular CLI with varying command sets from 7-Mode. Table 5 outlines some common user scenarios for command line management of HA pairs. The table shows the basic command set for each user scenario without all optional fields. For a complete list of commands, see the HA Configuration Guide or the CLI to review the available options for each command set.

Table 7) Common CLI usage for HA pair configurations

User Scenario	Command Line Interface	
	7-Mode	Clustered Data ONTAP
Monitor the HA pair status	`cf status`	`storage failover show`
Enable storage failover on the HA pair	`cf enable`	`storage failover modify –node` *`node`* `–enabled true`
Disable storage failover on the HA pair	`cf disable`	`storage failover modify –node` *`node`* `–enabled false`
Initiate a takeover on the HA pair	`cf takeover`	`storage failover takeover`
Initiate a giveback on the HA pair	`cf giveback`	`storage failover giveback`
Initiate a takeover with version mismatch (for NDU process)	`cf takeover –n`	`storage failover takeover –option allow-version-mismatch`
Enable storage failover on the HA pair under certain conditions	`options cf.hw.assist.enable` `Cf hw_assist status`	`storage failover hwassist`
Cluster image is downloaded and installed on all nodes in the cluster	`8.3 only`	`cluster image package get –url`
Validate cluster components and configuration	`8.3 only`	`cluster image validate` `–version 8.3.x`
Execute upgrade of all nodes within the cluster	`8.3 only`	`cluster image update` `–version 8.3.x`

8 Automated NDU (ANDU) for HA Pair Controller Configurations

7.1 Automated NDU Overview

Clustered Data ONTAP 8.3 adds support for automated, nondisruptive software upgrades (ANDU) The ANDU method validates the cluster components to ensure that the cluster can be upgraded nondisruptively, installs the target Data ONTAP image on each node, and then, based on the number of nodes in the cluster, executes either a rolling or batch upgrade in the background. All core NDU commands and routines required to nondisruptively upgrade the cluster are imbedded in the ANDU process and the administrator can monitor the progress, pause or resume an upgrade, and see the cluster update history.

7.2 Requirements for ANDU

- The cluster must be running Data ONTAP 8.3 or higher.
- Must read Clustered Data ONTAP 8.3 Upgrade and Revert/Downgrade Guide

7.3 ANDU Process

Three commands simplify upgrading your cluster by first bringing the clustered ONTAP package into the cluster(first obtained from support.netapp.com), validate the cluster to see if it is properly configured and ready to be upgraded, and then perform the actual upgrade.

Step 1: Download the software package

cluster image package get -url *location*
The software package contains the target Data ONTAP image and firmware, and the set of upgrade validation rules. This package is downloaded to the cluster package repository.

Step 2: Verify the cluster is ready to be upgrade

cluster image validate -version *package_version_number*
This command checks the cluster components to validate that the upgrade can be completed nondisruptively, and then provides the status of each check and any required action you must take before performing the software upgrade.

Step 3: Perform the software upgrade

cluster image update -version *package_version_number*
This command validates that each cluster component is ready to be upgraded, installs the target Data ONTAP image on each node in the cluster, and then performs a nondisruptive upgrade in the background.
If the cluster consists of 2 through 6 nodes, a rolling upgrade is performed.
If the cluster consists of 8 or more nodes, a batch upgrade is performed by default.

7.4 ANDU Caveats and Considerations

You can only use ANDU to upgrade from 8.3 to a subsequent release.
Because the code to run the automated upgrades is in 8.3, you must use a traditional approach to upgrade from 8.2 to 8.3.
Always perform post-upgrade checks manually as they are not automated.

7.5 Clustered Data ONTAP Support Matrix for ANDU

- Minor-release ANDU (starting with Data ONTAP 8.3.0)
  - Supported for Data ONTAP 8.3.0 (and later) ie; ONTAP 8.3.0 to 8.3.1, or 8.3.x to 8.4.x
- Major-release ANDU
  - Data ONTAP 8.3.x to 8.4 (and later) ie; ONTAP 8.4 to 8.5

Conclusion

This set of best practices offers general guidelines for deploying HA pair controllers. Every business environment is unique; therefore, it's crucial to understand the data infrastructure, network, and application requirements for your environment and then choose an appropriate solution based on those needs.

After implementing any high-availability solution, be sure to test production systems on a regular basis. Document recovery procedures, train the IT personnel to manage failures, and educate users on what to expect if failure occurs. In summary, be prepared for system outages.

Disclaimer

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.

References

The following references were used in this TR:

- TR-4847: Best Practices for Clustered Data ONTAP Network Configurations
  https://fieldportal.netapp.com/DirectLink.aspx?documentID=84837&contentID=108569
- TR-3975: DataMotion for Volumes Overview for clustered Data ONTAP
  https://fieldportal.netapp.com/DirectLink.aspx?documentID=63909&contentID=71149
- TR-4075: Data Motion for Volumes for clustered Data ONTAP 8.2 / 8.3
  http://www.netapp.com/us/system/pdf-reader.aspx?pdfuri=tcm:10-134838-16&m=tr-4075.pdf
- MetroCluster Compatibility Matrix
  http://support.netapp.com/NOW/products/interoperability/MetroCluster_Compatibility_Matrix.pdf

Version	Date	Document Version History
Version 1.0	March 2006	Initial Version
Version 1.0.1	June 2006	Content Update
Version 1.0.2	January 2007	Content Update
Version 2.0	August 2010	Content and Template Update
Version 2.0.1	September 2011	Content Update
Version 2.0.2	September 2012	Clustered Data ONTAP Update
Version 3.0	April 2013	Content and Template Update
Version 4.0	May 2015	Content and Template Update
Version 4.1	February 2016	Jay Bounds: Minor updates

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.

© 2013 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, DataMotion, Data ONTAP, FilerView, MetroCluster, NOW, RAID-DP, SnapDrive, SyncMirror, vFiler, and WAFL are trademarks or registered trademarks of NetApp, Inc. in the United States and/or other countries. Windows is a registered trademark of Microsoft Corporation. Oracle is a registered trademark of Oracle Corporation. Symantec and Veritas are trademarks of Symantec Corporation. All other brands or products are trademarks or registered trademarks of their respective holders and should be treated as such. TR-3450-0413

NetApp
www.netapp.com