



Technical FAQ

Nondisruptive Operations (NDO) Technical FAQ

Jay Bounds, NetApp

March 2015

Abstract

The nondisruptive operations technical FAQ is a compilation of frequently discussed topics and questions on NDO for clustered Data ONTAP.

Version History

Version	Date	Document Version History
Version 1.0	August 2012	Initial version
Version 2.0	June 2013	New template update; Data ONTAP 8.2 content
Version 2.1	October 2013	Terminology and other revisions
Version 2.1.1	December 2013	Minor updates for Data ONTAP 8.2.1 content
Version 3.0	March 2015	Major updates for Data ONTAP 8.3.0 content

TABLE OF CONTENTS

1	Architecting Clustered Data ONTAP Systems for Nondisruptive Operations	5
1.1	Configuration Requirements for Nondisruptive Operations	5
	What are the requirements for a clustered Data ONTAP system to be capable of nondisruptive operations?	5
	Hardware and Software	6
	Best Practices for Maintenance and Lifecycle Operations	9
1.2	NDO Minimum Configuration	10
	What is a the NDO Minimum Configuration”	10
2	Lifecycle Operations (DataMotion for Volume, LIF Migrate, and DataMotion for LUN)	11
2.1	Capacity and Performance Optimization using DataMotion for Volumes	11
	Is volume move nondisruptive for all protocols?	11
	Can a node root volume be moved using volume move?	11
	Are limits validated as part of the volume move process?	11
	Are there limitations when moving volumes between aggregates on controllers with different version of clustered Data ONTAP?	11
	Where can I find additional resources for DataMotion for volumes?	11
	Is LIF migrate nondisruptive for all protocols?	11
	Is LIF failover nondisruptive when an unplanned failover occurs?	12
	Are LIF failover groups required for nondisruptive operations?	12
	Are additional resources available for LIF migrate?	12
2.2	Capacity and Performance Optimization using DataMotion for LUNs	12
	What is DataMotion for LUNs?	12
	How is DataMotion for LUNs different from DataMotion for Volumes?	12
	Can I use DataMotion for LUNs to move or copy a LUN to a volume where other LUNs already exist? ..	12

Does DataMotion for LUNs use the same underlying engine developed to support VMware® Virtual Volumes (VVols)?	13
Can I throttle DataMotion for LUN operations?	13
Can I move LUNs between SVMs?	13
What happens to the Storage Efficiency Policies when a LUN is moved to another volume?	13
What interactions does a LUN move operation have with a vol move operation?	13
What happens to the Snapshot copies of a LUN that is moved to another volume?	13
What is meant by “promote early” or “promote late” when using LUN copy or move commands?	13
What is the OnDemand engine?	14
3 Maintenance Operations (Storage Failover and Aggregate Relocate)	14
3.1 Nondisruptive Upgrade (NDU)	14
How long does a nondisruptive upgrade take to complete on a single HA pair within a clustered Data ONTAP system?	14
How is a nondisruptive upgrade different from a rolling upgrade?	15
What is a mixed-version cluster?	16
What is the effective cluster version?	16
Can I add nodes to a cluster with a different version of Data ONTAP from the existing nodes?	16
Are there restrictions for load-sharing mirrors and data protection mirrors in mixed-version clusters?	17
What are best practices for upgrading nodes that have SnapMirror relationships set up?	17
Are there restrictions for doing a volume move or volume copy in conjunction with an NDU?	17
Is NDU disruptive to applications accessing SMB shares?	17
Where can I find additional NDU resources?	18
3.2 Automated Nondisruptive Upgrade (ANDU)	18
What is Automated Nondisruptive Upgrade (ANDU)?	18
How is ANDU different from NDU?	18
How does ANDU work?	18
What NetApp ONTAP versions support ANDU?	18
3.3 Hardware Upgrade and Replacement	19
Can controller hardware be upgraded nondisruptively?	19
How do cluster node limits affect controller upgrades?	19
Can I combine an in-place controller upgrade using aggregate relocate with a Data ONTAP upgrade?	19
Do the OnCommand management tools support NDO functions?	20
Can storage shelves be removed nondisruptively?	20
Can nodes in a cluster be nondisruptively physically moved to a different part of the data center?	20
3.4 Storage Failover (SFO)	21
Is failover in clustered Data ONTAP different from failover in 7-Mode?	21
Can I have data volumes in my node root aggregate? What about an SVM root volume?	21
How do high-availability (HA) pairs work in a clustered Data ONTAP configuration?	21
Does hardware-assisted takeover use the cluster interconnect or the HA interconnect?	22
Does a single-node cluster have HA? What about other NDO capabilities?	22
Is HA pair failover completely nondisruptive?	22

What is cluster quorum, and how does it affect data availability?	23
Does a two-node cluster use epsilon?	24
What happens in a two-node cluster if the cluster interconnect fails?	24
Can I add an additional or dedicated node to host epsilon?	25
Does epsilon move to another node if the original node holding epsilon goes down?	25
In what circumstances should epsilon be manually moved to another node?	25
Are there additional resources available for HA pairs and storage failover?	25
4 Cluster Scalability.....	25
4.1 Cluster Growth and Operations.....	25
Can a single-node cluster upgrade to a two-node cluster nondisruptively?	25
Can a two-node switchless cluster be upgraded to a switched cluster nondisruptively?	26
Is a switch required to grow a cluster beyond two nodes?	26
Can nodes be joined to the cluster nondisruptively?	26
Can nodes be unjoined from the cluster nondisruptively?	26
Can two separate clusters be merged together as a single cluster nondisruptively?	26
Can a cluster be split into two separate clusters nondisruptively?	27
References.....	28
Version History	2

LIST OF TABLES

Table 1) Customer business analysis overview.....	5
Table 3) NDO overview for NAS and SAN environments.	10
Table 3) NDU preprocess steps.	14
Table 4) NDU steps.....	14
Table 5) NDU postprocess steps.....	15
Table 6) Examples for understanding epsilon in a four-node cluster.....	24

LIST OF FIGURES

Figure 1) SVM root with load-sharing mirror on each node in the cluster.	7
Figure 2) LIF failover group example.....	8
Figure 4) Hardware resiliency diagram.....	9
Figure 8) Example four-node cluster for epsilon.....	23

1 Architecting Clustered Data ONTAP Systems for Nondisruptive Operations

1.1 Configuration Requirements for Nondisruptive Operations

What are the requirements for a clustered Data ONTAP system to be capable of nondisruptive operations?

Answer: Nondisruptive operation (NDO) is the capability of systems operating in clustered Data ONTAP® to continuously serve data without disruption during maintenance and lifecycle operations as well as unplanned failure events. NDOs are based on certain architectural requirements, enabling necessary operations that maintain the resiliency of the infrastructure. The highest degree of infrastructure resiliency is needed for customers requiring zero downtime; low levels of resiliency cater to customers with the ability to schedule small windows of downtime to execute operations.

The following section discusses the different business-level objectives of customers and the overall goal of each in terms of resiliency. A tier 1 customer requires the highest degree of resiliency, while a tier 4 might not. As customers architect their clustered Data ONTAP solution, they should identify the tier of business this solution addresses and then apply the necessary levels of resiliency. Each is discussed within this section.

Table 1) Customer business analysis overview.

Tier	Description
1. mission critical	These types of environments enable services that are in high demand and cost the customer significant loss of revenue when an outage occurs. Online transaction processing (OLTP), batch transaction processing, and virtualization/cloud environments are examples of environments that fit into this tier of data availability. This tier is tuned toward prioritizing I/O response to foreground (client application) traffic so that dependent applications remain functional. By prioritizing foreground I/O over corrective I/O in degraded situations, you increase the time necessary 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.
2: business critical	These types of environments are likely subject to compliance requirements. Although maintaining client access to the storage system is important, the loss of data would be severely detrimental to the customer. This could also be a configuration that is protecting a company's intellectual property. Medical records, software source code, and e-mail are examples of environments that fit into this tier of data availability. This tier is tuned toward prioritizing corrective I/O while balancing foreground I/O. By prioritizing corrective I/O over foreground I/O in degraded situations, you increase the impact on foreground I/O performance.

Tier	Description
3: repository	Repository environments are used to store collaborative data or user data that is not critical to business operations. Scientific and engineering compute data, workgroup collaboration, and user home directories are examples of environments that fit into this tier of data availability. This tier is the middle ground that balances foreground operations with corrective actions (should they be needed). Defaults are normally appropriate for these configurations.
4: archival	This type of environment is subject to a large initial ingest of data (write) that is then seldom accessed. System utilization on average is not expected to be very significant. Since 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 are tuned toward prioritizing corrective I/O and minimizing completion time for corrective actions. Backup and recovery, archiving, near-line, and reference data are examples of environments that fit into this tier of data availability.

For each of the customer storage tiers defined in Table 1, consider the following recommendations for each logical, physical, and operational component.

Hardware and Software

Platform Mixing Rules

The composition of the cluster must conform to the platform mixing rules at all times. A cluster may not be temporarily out of compliance in order to perform a technology refresh operation. In particular, it is recommended that clusters consisting of entry-level (22x0) platforms using internal drives should be limited to 2-nodes, in order to allow nodes to be added for a platform refresh. Section 3.2 contains more details on this.

See the References section of this document for information on the platform mixing rules.

Cluster Interconnect Configuration

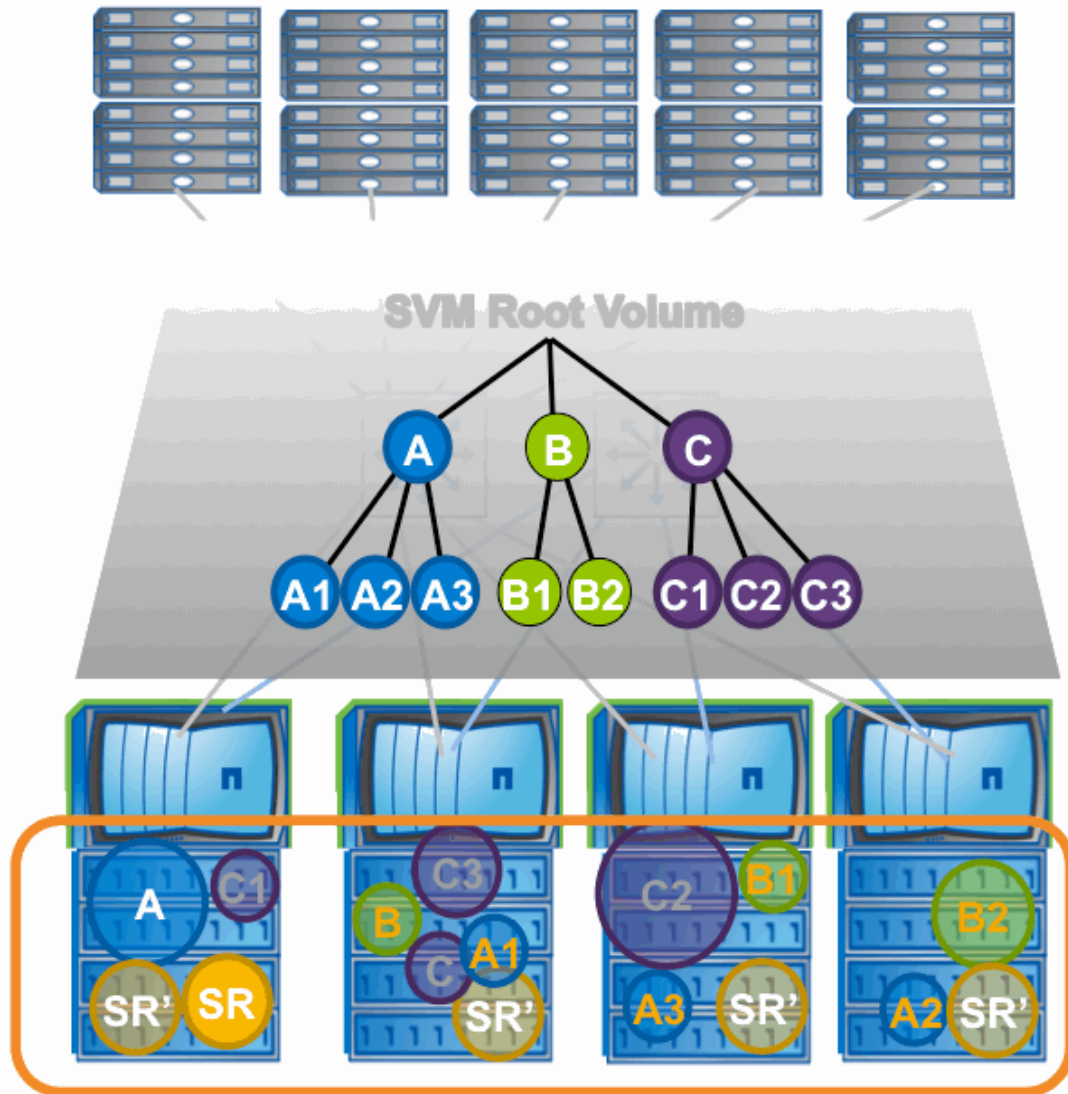
The cluster interconnect switch is a critical cluster resource for performance and resilience. It must be installed and maintained according to the reference configuration. Dual cluster interconnects (LIFs) are required to establish no single point of failure for the cluster interconnect and for compliance with the NDO Minimum Configuration. The one-cable two-node switchless cluster (FAS2220 and 22240 only) has a single point of failure in the cluster interconnect and has very limited technology refresh options. See section 3.4 and section 4.1 for more notes on the one-cable two-node switchless cluster.

SVM Resiliency

The Storage Virtual Machine (SVM) root is the gateway for accessing volumes in NAS environments. If the SVM root is located on the storage of an HA pair where both nodes become unavailable, access to the root of the SVM will be disrupted, which might disrupt access to other volumes in the SVM. Configuring load-sharing mirrors of the SVM root adds a layer of resiliency for this scenario by allowing the promotion of the load-sharing mirror and resuming NAS access to the SVM. It is a requirement for basic SVM resiliency to have at least one load-sharing mirror on another HA pair in the cluster completely separate from the HA pair on which the SVM root resides. Since the SVM root volume is typically small in

size, best practice is to configure load-sharing mirrors on very cluster node, including the node hosting the SVM root, for maximum resiliency.

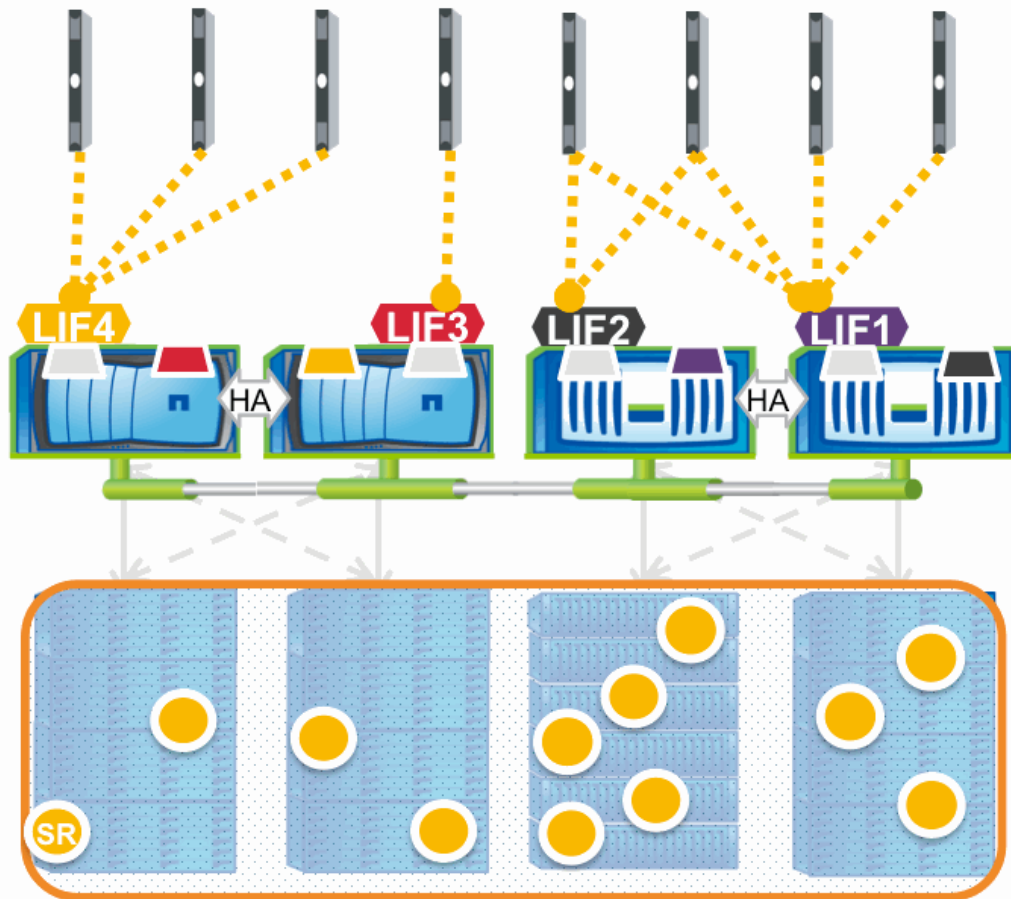
Figure 1) SVM root with load-sharing mirror on each node in the cluster.



Logical Interface (LIF) Resiliency

The logical interfaces that are created on each node should have failover groups defined for each LIF. In the case of a node failure or a port failure, the LIF can fail over to another network port or ifgroup hosted on a different controller elsewhere in the cluster and continue to serve data. Having failover groups defined appropriately for a LIF that contain the same layer 2 network as the home port/interface group of the LIF is critical for achieving transparency for the workload during and after the LIF migration or failover.

Figure 2) LIF failover group example.



Node (Controller) Resiliency

In the case of a planned or an unplanned event that takes a node down, HA pair controllers are configured to maintain access to the data. A node within an HA pair does a failover of its storage to the partner node to allow client requests to be processed during the time the node is down.

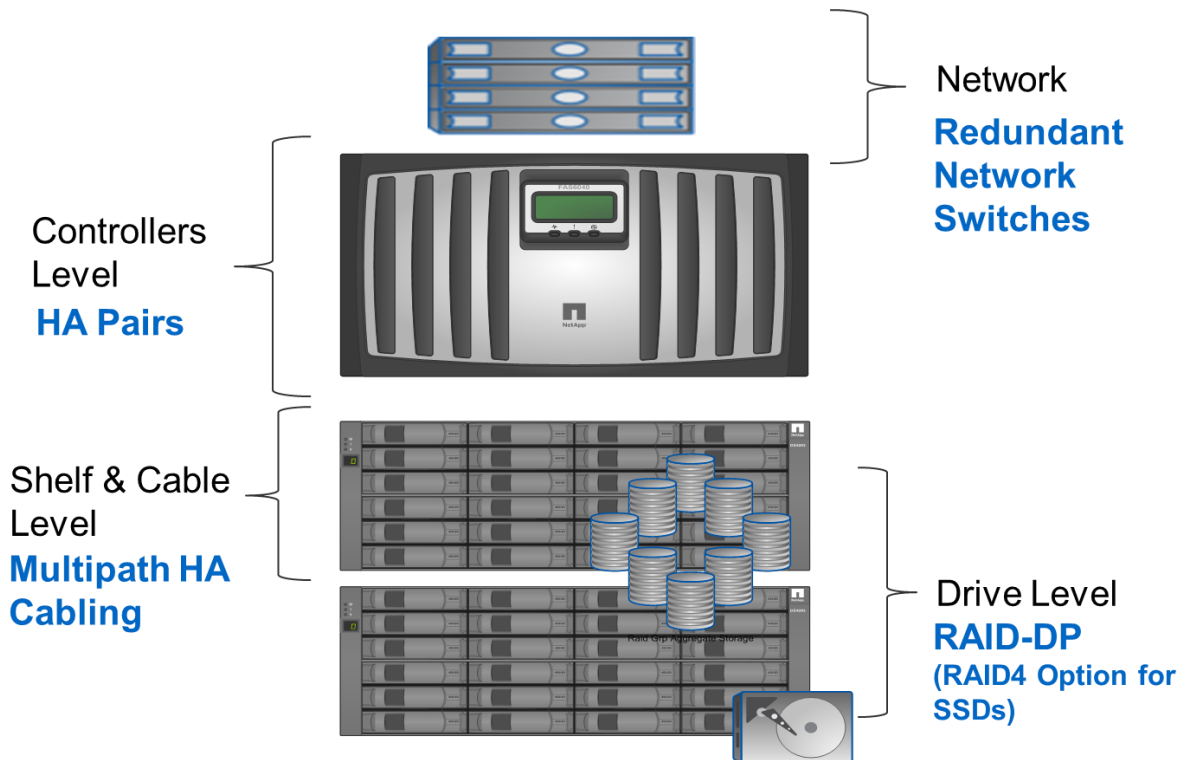
Each cluster node contains a root aggregate which contains the node root volume. The root aggregate must not contain any data volumes. The reason for this is that, during giveback, the root aggregate will be returned to its home node first, assimilating the node back into the cluster and syncing necessary cluster information. The process during which the root aggregate completes giveback can take time; therefore, any data residing in the root aggregate would not be available during this giveback period. Data aggregates are not returned to their home node until the root aggregate is online and the node is back in quorum. Each data aggregate continues to be served by the partner node while the root aggregate is brought online. When the data aggregates are to be returned to their home node, each data aggregate is sent home serially, minimizing the window during which each data aggregate will be unavailable during the giveback transition. Generally, a data aggregate takes up to 30 seconds to complete a giveback and resume servicing data requests. The same process occurs in reverse during a planned failover in clustered Data ONTAP 8.2 and higher: data aggregates are taken over serially by the partner node, and the root aggregate is failed over last.

As discussed in the previous section, LIF failover groups should be defined for network connections to continue processing I/O requests nondisruptively in case of a node failure by migrating to an alternate node within the cluster.

Hardware Resiliency

The storage subsystem has several layers of resiliency that can be built into the architecture. These are some of the minimum requirements for optimal data availability.

Figure 3) Hardware resiliency diagram.



Best Practices for Maintenance and Lifecycle Operations

Follow Recommended NDO System Limits

For clustered Data ONTAP systems, the customer should stay within defined limits for NDU and other planned maintenance operations. The important limits are documented in the [Nondisruptive Upgrade Technical FAQ](#).

Reserve Space in the Cluster for Cluster Growth and Optimization

Allowing space within the clustered Data ONTAP system to move volumes and execute activities, such as hardware maintenance, will streamline activities requiring a volume move.

Perform Regular System Monitoring

Finally, tracking node CPU utilization and monitoring potential hot spots within the cluster are important for maintaining overall cluster health. A node that is running with CPU utilization greater than 50% might experience an impact on takeover. It is a best practice to maintain CPU utilization of no more than 50% on each node throughout the cluster.

Review Support for Protocols and Associated Best Practices for Network and Application Resiliency

A general recommendation for all applicable configurations is to have multipath client/host connections. Client and/or host application networks should be connected to multiple nodes in the cluster to support LIF failover.

Clustered Data ONTAP systems servicing SMB clients must run SMB 2.0 to take advantage of current NDO solutions for CIFS environments. SMB 1.0 clients might incur unexpected disruptive events. For SMB 3.0, nondisruptive storage failover and aggregate relocate are supported only for Continuously Available shares in Hyper-V environments. Refer to [TR-4100: Nondisruptive Operations and SMB File Shares for Clustered Data ONTAP](#) for more information.

Table 2) NDO overview for NAS and SAN environments.

Nondisruptive Operation	NAS				SAN
	SMB 1.0	SMB 2.x and SMB3.0	SMB 3.0 (Hyper-V Only)	NFS	
LIF migrate	Possibly disruptive (Refer to section 2)	Nondisruptive*	Nondisruptive	Nondisruptive	N/A
Volume move	Nondisruptive	Nondisruptive	Nondisruptive	Nondisruptive	Nondisruptive
LUN Move	Nondisruptive	Nondisruptive	Nondisruptive	Nondisruptive	Nondisruptive
Aggregate relocate	Disruptive	Disruptive	Nondisruptive	Nondisruptive	Nondisruptive
Storage failover	Disruptive (Refer to Section 3)	Disruptive (Refer to Section 3)	Nondisruptive	Nondisruptive	Nondisruptive

- Applications connected to SMB 2.x file shares may incur a disruption when the application allows multiuser access to the same file or for applications that do not pick up an open file handle within 15 minutes from when the LIF migrate was executed.

1.2 NDO Minimum Configuration

What is a the NDO Minimum Configuration”

Answer: The NDO Minimum Configuration consists in summary, of the following points discussed above:

- Clustered Data ONTAP 8.1 or higher (8.2 or higher recommended, and required for SMB3)
- Clustered Data ONTAP 8.3 for LUN move and SMB3(CA)
- Clusters consist of HA pairs with multipath HA disk cabling, storage configured to best practices
- Two-node cluster must have cluster HA enabled
- Load-sharing mirrors of all SVM root volumes
- Dedicated root aggregate on each node, containing no other volumes
- Appropriate and correct LIF failover groups

- Reference configuration for the cluster interconnect switch
- Dual cluster interconnects

2 Lifecycle Operations (DataMotion for Volume, LIF Migrate, and DataMotion for LUN)

2.1 Capacity and Performance Optimization using DataMotion for Volumes

Is volume move nondisruptive for all protocols?

Answer: Yes. Volume move is nondisruptive for all protocols. However, there are differences in the way volume move is handled compared to other NDO utilities such as HA failover. During the cutover phase of a volume move operation, necessary volume information that is stored in memory on the controller is replicated to the controller memory of the node that owns the source aggregate of the volume move. The cutover phase allows file handling information to remain persistent for stateful protocols after the volume has been moved.

Can a node root volume be moved using volume move?

Answer: No. The only way to move a root volume is to use the disk repair feature. The disk repair feature will copy the data from each drive in the root volume to a new physical spare drive. This might be needed when upgrading controllers that have internal drives in the controller chassis that contain the root volume. Disk copy is nondisruptive; however, it differs from volume move in the way it utilizes disk space on the destination drive. The destination drive of the disk copy will be sized down to be the same size as the source drive. If the destination drive is larger than the source drive, the additional space will be lost.

Are limits validated as part of the volume move process?

Answer: Some limits are validated, such as FlexVol® limits per node. Many limits are not validated and might be exceeded when a volume is moved. Some limits that should be considered are limits on LUNs, max volume, and aggregate size between dissimilar controllers. Many of these limits will be incorporated into the volume move validation phase; however, at this time there is a small subset of checks that are carried out.

Are there limitations when moving volumes between aggregates on controllers with different version of clustered Data ONTAP?

Answer: Yes. A cluster that is being upgraded from a version of Data ONTAP 8.0 to a version of Data ONTAP 8.1 does not support a volume move or volume copy operation between nodes running Data ONTAP 8.0 and 8.1. A volume move or volume copy operation is supported for Data ONTAP 8.1 and beyond when executing the volume operation from the lower version of Data ONTAP to the higher version of Data ONTAP. A volume move or copy operation is not supported when executing the volume operation from the newer version of Data ONTAP to the older version of Data ONTAP.

Where can I find additional resources for DataMotion for volumes?

Answer: For more information on volume move, see [TR-3975: DataMotion for Volumes Overview](#) and [TR-4075: DataMotion for Volumes Best Practices and Optimization](#).

Is LIF migrate nondisruptive for all protocols?

Answer: No. LIF migrate is nondisruptive for all commonly supported protocols with the exception of SMB 1.0 for CIFS. There might be a disruption for clients running SMB 1.0 during a LIF migration. SMB 2.0 and beyond are nondisruptive for LIF migration. For SMB 3.0 with continuously available shares for Hyper-V™ environments, LIF migrate is completely nondisruptive.

Is LIF failover nondisruptive when an unplanned failover occurs?

Answer: Yes, for all protocols except SMB 1.0. SMB 2.0 supports persistent file handles, and SMB 3.0 supports continuous file handles, both allowing for the open session to resume when the LIF is migrated. A LIF can fail over to another node within the cluster during an event in which SFO occurs on the node currently hosting the LIF. As a best practice, failover groups should be defined for a LIF in the cluster. Refer to [TR-4100: Nondisruptive Operations for SMB File Shares](#) for details.

Are LIF failover groups required for nondisruptive operations?

Answer: Yes. Failover groups should be configured as the first line of redundancy in the event of a failure; the LIFs need to have some kind of direction to manage to where they fail over. Additional resiliency can be combined with failover groups by using features such as ifgrps, VLANs, and so on.

Are additional resources available for LIF migrate?

Answer: Yes. For additional information on clustered Data ONTAP networking and LIF migrate, refer to [Clustered Data ONTAP Networking FAQ](#).

2.2 Capacity and Performance Optimization using DataMotion for LUNs

What is DataMotion for LUNs?

Answer: DataMotion for LUNs allows individual LUNs to be moved and copied nondisruptively between Server Virtual Machine(SVM) in a cluster. DataMotion for LUNs permits host(s) to access the LUN on its new destination immediately while data is being transferred over the cluster interconnect network in the background.

How is DataMotion for LUNs different from DataMotion for Volumes?

DataMotion for LUNs is similar to DataMotion for volumes in that the operation takes place nondisruptively in the background with ALUA changing path status in order to reflect a LUN's new location. Unlike DataMotion for volumes - which initiates a cutover to the destination only when all data has been transferred to the destination, DataMotion for LUNs performs the cutover instantly - making the new destination LUN immediately available for write access.

There are two operations associated with DataMotion for LUNs: move and copy:

- Move operations can only occur within the same SVM, though the source and destination volumes can be on any node within the cluster. If necessary, the SVM signals a path state change using ALUA as soon as the destination LUN is prepared to accept data. The move is nondisruptive to hosts accessing the LUN.
- Copy operations can occur between any two volumes within an SVM. LUN copy can also specify a different SVM's volume as a destination if the operation is performed by a cluster administrator. LUN copy is nondisruptive insofar as the LUN is available for access at the destination as soon as it has been created.

Can I use DataMotion for LUNs to move or copy a LUN to a volume where other LUNs already exist?

Answer: Yes. A destination can have other LUNs.

Does DataMotion for LUNs use the same underlying engine developed to support VMware® Virtual Volumes (VVols)?

Answer: Yes. However, DataMotion for LUNs commands such as `lun move` and `lun copy` should not be used on VVol endpoints.

Can I throttle DataMotion for LUN operations?

Answer: Yes, the throughput (that is, MB/s, GB/s) can be throttled for each individual LUN move/copy operation. The throttle can be set to another value while the operation is in flight.

Can I move LUNs between SVMs?

Answer: A LUN move is not possible between SVMs. A LUN mapping must be created on the destination SVM before data can be accessed. Additionally, other SVMs will have a different Fibre Channel target WWN or iSCSI target portal identifier and will be seen by the host as an independent target.

A LUN copy between SVMs is permitted if performed by a cluster administrator.

What happens to the Storage Efficiency Policies when a LUN is moved to another volume?

Answer: After LUN(s) arrive at the destination, they will not be storage efficient on disk, and deduplication or compression must be run to recoup space in the destination volume. Destination volumes that have storage efficiency enabled will log incoming blocks as needing a deduplication or compression scan during the next background run.

What interactions does a LUN move operation have with a vol move operation?

Answer: Vol move is not allowed on the destination volume while a LUN copy or move operation is active. However, the source volume can be relocated by vol move during a LUN copy or move operation, but it will be paused during the vol move cutover.

What happens to the Snapshot copies of a LUN that is moved to another volume?

Answer: Subvolume moves or copies will not copy volume-level data such as Snapshot® copies or deduplication and compression. After the move or copy is completed, compression and deduplication scans can be started to regain any storage efficiencies lost during the move or copy.

Until new Snapshot copies have been initiated on the destination volume, it is a best practice to retain existing Snapshot copies on the source volume for disaster recovery and backup purposes. It is also possible to use LUNs in the source volume's Snapshot copies as LUN copy sources. If multiple copies of a LUN are all part of the same LUN copy operation, they will be transported efficiently to the destination and after a deduplication scan will share blocks in the same way that SIS-based LUN clones do.

What is meant by “promote early” or “promote late” when using LUN copy or move commands?

Answer: When moving or copying a LUN, the notion of early or late promotion has to do with when the cluster advertises the new LUN. In the case of LUN move, the default is “promote early.” This means that

the new LUN is advertised to hosts as soon as the destination LUN is created. All reads and writes are routed to the LUN's new location while a background operation copies existing LUN data from the source to the destination. Snapshot operations will be blocked until the new LUN is fully framed and the move/copy reaches "moving data" status.

The alternative is late promote (default for LUN copy), in which the cluster promotes the new LUN location after the LUN has been fully framed. Fully framed means that the new LUN exists as a fully functional LUN, but its data is missing/hasn't arrived yet. A consequence of being fully framed is that Snapshot copies are no longer blocked.

When moving or copying a LUN, the promoting "late" or early is controlled by the command line LUN move/LUN copy command using the argument "-promote-late" for LUN move and "-promote-early" for LUN copy. If not specified, then the default for the operation will be taken:

What is the OnDemand engine?

Answer: The OnDemand engine enables actions, such as LUN move, to appear to complete instantly. This allows traffic to begin on the newly created destination and allows ALUA to signal path changes, while clustered Data ONTAP transfers the actual data in the background.

3 Maintenance Operations (Storage Failover and Aggregate Relocate)

3.1 Nondisruptive Upgrade (NDU)

Nondisruptive upgrade is the ability to nondisruptively upgrade system software and firmware without having to take downtime.

How long does a nondisruptive upgrade take to complete on a single HA pair within a clustered Data ONTAP system?

Answer: The NDU procedure includes several different steps. The breakdown of steps includes preprocess and postprocess steps as well as the actual execution of the NDU. The length of the NDU includes the preprocess, NDU, and postprocess steps.

Table 3) NDU preprocess steps.

Preprocess Steps	
Steps	Approximate Duration
Download disk/shelf firmware	~3 minutes per disk (BDFU enabled)
Data ONTAP software installation	10–30 minutes

Table 4) NDU steps.

NDU Steps	
Steps	Approximate Duration
Takeover #1	Less than 2 minutes

Update BIOS	11 minutes
Reboot	3–4 minutes
Giveback #1	4–10 minutes
Stabilize MP host software	8 minutes
Takeover #2	Less than 2 minutes
Update BIOS	11 minutes
Reboot	3–4 minutes
Giveback #2	4–10 minutes

Table 5) NDU postprocess steps.

Postprocess Steps	
Steps	Approximate Duration
Upgrade service processor	5–25 minutes
Upgrade RLM	5 minutes
Upgrade BMC	45–90 minutes

The giveback duration is the summation of the time it takes for the root aggregate to complete giveback plus the time it takes for each aggregate to complete giveback serially. For example, on a system with 10 data aggregates, a giveback might take about 3 minutes for the root aggregate to come online and assimilate the node back into the cluster and 30 seconds for each aggregate to complete giveback. Therefore, the total time it would take for giveback to complete would be (3 minutes) + (30 seconds * 10 aggregates), equaling 6 minutes. The important point, especially in regard to NDO, is that the time any one aggregate would be unavailable to serve data would be 30 seconds or less.

The complete time it takes to perform an NDU on an HA pair will vary depending on the configuration. Systems with a smaller number of drives and fewer aggregates will take less time than systems with a larger number of drives and more aggregates.

For example, a system with 24 drives and 2 aggregates would have a total NDU time of (assuming the lowest duration for any of the preceding steps that have duration ranges):

Preprocess = (24 drives * 3 minutes) + (10 minutes) = 82 minutes

NDU process = (2 + 11 + 3 + 4 + 8 + 2 + 11 + 3 + 4) = 48 minutes

Postprocess = (5 minutes + 5 minutes + 45 minutes) = 65 minutes

How is a nondisruptive upgrade different from a rolling upgrade?

Answer: A rolling upgrade is the complete upgrade of a cluster as opposed to a nondisruptive upgrade, which is the upgrade of a single HA pair within a cluster. A rolling upgrade consists of one or more nondisruptive upgrade procedures being executed on the nodes of the cluster. The number of NDU procedures is dependent on the number of HA pairs in the cluster. For example, a rolling upgrade for a two-node cluster would be complete when a single nondisruptive upgrade is completed on the HA pair that makes up the two-node cluster. Alternatively, a rolling upgrade on a four-node cluster would not be complete until a nondisruptive upgrade was first executed on one HA pair and then the second HA pair, in a serialized fashion. A standard rolling upgrade allows only one HA pair to be upgraded at a time.

In Data ONTAP 8.2, batch upgrades are introduced, similar to rolling upgrades; however, a batch upgrade allows a set of HA pairs, referred to as a batch, to be upgraded at the same time. Batch upgrades reduce the overall time it takes to upgrade the cluster to a new version of Data ONTAP and are available for clusters of 8 nodes or higher.

During a rolling or batch upgrade, the cluster will have two versions of Data ONTAP running in the cluster. This is considered a mixed-version cluster. During the time a cluster is in a mixed-version state, it is recommended to complete the rolling or batch upgrade in the shortest period of time, preferably within a seven-day period. During the period of time when the cluster is in a mixed-version state, there is limited support for cluster configuration changes and cluster operations such as volume moves. This is due to the effective cluster version (ECV) that is defined by the lowest version of Data ONTAP in the cluster at any given time. The effective cluster version defines what features are supported in the cluster and what support for those features is enabled. Feature incompatibility becomes an issue in mixed-version clusters, and therefore limited use is enforced during an upgrade.

What is a mixed-version cluster?

Answer: A mixed-version cluster is a cluster that has a subset of nodes running one version of Data ONTAP and another subset of nodes running a different version of Data ONTAP. This applies to both major and minor releases, so a cluster containing nodes running clustered Data ONTAP 8.1 and 8.1.1 or 8.1.1 and 8.2 would be operating in mixed-version. A cluster running 8.2 and 8.2P4 would not be considered a mixed-version cluster; patches within the same release do not count as a mixed-version cluster.

A mixed-version cluster:

- can have no more than two sequential versions of Data ONTAP running on the nodes within the cluster. A patch release of a Data ONTAP release is not considered to be a separate version of Data ONTAP.
- is for the purpose of completing a rolling upgrade. Other scenarios that introduce a mixed-version cluster such as adding nodes to a cluster with different versions of Data ONTAP might affect the effective cluster version. If new nodes are running a higher ONTAP release than the current effective cluster version, it is recommended to either first upgrade the existing nodes to match the new node version, or to install the effective cluster version release on the new nodes before adding them to the cluster. In this way, a mixed-version cluster will be avoided.
- is not considered a normal or steady state of the cluster. The mixed-version state should be limited to the time it takes to complete the rolling upgrade, preferably within a seven-day period.

What is the effective cluster version?

Answer: A cluster always has an effective cluster version (ECV) of clustered ONTAP. The ECV is considered the lowest version in the cluster and does not change until all nodes in the cluster have been upgraded and are running the same version. The ECV determines the cluster-wide features that are available. Cluster-wide features of an ONTAP release higher than the ECV will not be available for deployment in the cluster. If nodes at a higher release than the cluster's ECV are added to a cluster, the ECV remains at the lower release.

Can I add nodes to a cluster with a different version of Data ONTAP from the existing nodes?

Answer: Yes, however it is recommended to avoid this if possible. If a new platform requires a higher release of ONTAP than the current ECV, a mixed-version cluster will be avoided if the existing nodes are first upgraded to the new release before joining the new nodes. If this is not possible, then the existing nodes should be upgraded to the higher release within the recommended seven-day period. Alternatively, if it is intended to evacuate and remove the older nodes, this should also be completed as soon as possible. The ECV will continue to be defined by the lower, existing version of Data ONTAP. Any new features available on the new version of Data ONTAP on the newly added nodes might not work until the entire cluster is upgraded to the latest version of clustered Data ONTAP, thereby incrementing the ECV.

If a node is running a lower version of Data ONTAP than the nodes in the cluster, the ECV would be decremented to the ECV of the lower version of Data ONTAP. The lower version of Data ONTAP would redefine the ECV, and some features might become unavailable or lose certain functionality depending on when they were implemented in clustered Data ONTAP. For example, if a cluster is running Data ONTAP 8.1.1 and an HA pair running Data ONTAP 8.1 is joined to the cluster, the two new nodes would cause a change in the ECV, thereby imposing restrictions on new features enabled by the Data ONTAP 8.1.1 ECV. The two nodes would need to be nondisruptively upgraded to Data ONTAP 8.1.1 for the ECV to be incremented to the ECV defined by Data ONTAP 8.1.1. The preferred method is to install the correct version of ONTAP on the new nodes to match the existing ECV before joining them to the cluster.

Are there restrictions for load-sharing mirrors and data protection mirrors in mixed-version clusters?

Answer: Yes. A cluster that is being upgraded from a version of Data ONTAP 8.0 to a version of Data ONTAP 8.1 requires that all load-sharing and data protection mirrors be deleted prior to the upgrade. The replication engine in Data ONTAP 8.0 is not compatible with Data ONTAP 8.1 and beyond.

Additionally, during a rolling upgrade, load-sharing and data protection mirrors only work from the older release to the newer release. Mirrors from newer to older releases will not be updated. After the nodes running the older release are updated to the newer release, the mirrors will continue to operate.

What are best practices for upgrading nodes that have SnapMirror relationships set up?

Answer: SnapMirror® relationships create a complex strategy for upgrading nodes in a cluster while maintaining active mirrored relationships. In general, the destination node of the SnapMirror relationship must be running an equivalent or higher version of Data ONTAP than the source node. SnapMirror relationships (both data protection and load sharing mirrors) must be quiesced on each HA pair as it is upgraded. The destination node of the SnapMirror relationship should be upgraded before the source node.

Are there restrictions for doing a volume move or volume copy in conjunction with an NDU?

Answer: Yes. A cluster that is being upgraded from a version of Data ONTAP 8.0 to a version of Data ONTAP 8.1 does not support a volume move or volume copy operation between nodes running Data ONTAP 8.0 and 8.1. A volume move or volume copy operation is supported for Data ONTAP 8.1 and beyond when executing the volume operation from the lower version of Data ONTAP to the higher version of Data ONTAP. A volume move or copy operation is not supported when executing the volume operation from the newer version of Data ONTAP to the older version of Data ONTAP.

Is NDU disruptive to applications accessing SMB shares?

Answer: Yes, with the exception of SMB 3.0 configured with Continuously Available shares for Hyper-V. For a clustered Data ONTAP configuration servicing SMB 1 and 2 clients, there might be a disruption to the CIFS clients during an NDU. The statefulness of the protocol is not persistent across controller failover; therefore, when an NDU and a failover are required during controller reboot, the necessary information required by the client about the open file is lost.

NDU in the traditional sense is disruptive to SMB 1 and 2 versions, but there are alternative solutions in clustered Data ONTAP using volume move and LIF migration in conjunction with NDU to maintain data availability for CIFS clients and upgrade controllers within the cluster. Refer to [TR-4100: Nondisruptive Operations for SMB File Shares](#) for details.

NDU is disruptive to CIFS clients that are accessing data on a volume owned by the node that is rebooted during the NDU procedure. CIFS clients accessing data on other volumes in the cluster not owned by the node being rebooted are not affected if one of the following conditions is true:

- The LIF is migrated from the node being rebooted to another node in the cluster.
- The LIF has defined failover rules and is failed over when the node is rebooted.
- The LIF is hosted by another node within the cluster.

For CIFS client applications that are sensitive to an outage, a volume move can be used to move the volumes from the node that is being upgraded to another node within the cluster that has sufficient aggregate space available for the duration of the upgrade.

Where can I find additional NDU resources?

Answer: Nondisruptive upgrade information is available in several different formats. For guidance on a specific NDU plan, refer to Upgrade Advisor, available on the NetApp® Support site as part of the My AutoSupport™ tool. General procedural information is available in the product documentation for the “Upgrade and Revert Guide,” available on the NetApp Support site. In addition, the [Nondisruptive Upgrade Technical FAQ](#) is available on the Field Portal for other important NDU information.

3.2 Automated Nondisruptive Upgrade (ANDU)

What is Automated Nondisruptive Upgrade (ANDU)?

Answer: ANDU simplifies the nondisruptive upgrade process by reducing the number of user input commands associated with performing software upgrades. ANDU does not require the user to be present during the upgrade process.

How is ANDU different from NDU?

Answer: Reduces Time...

The existing Non-Disruptive Upgrade procedure involves executing a long list of commands to upgrade to the next Data ONTAP Version. The user is required to sit in front of the console till the upgrade completes. The number of commands to be executed and the duration of the upgrade increases drastically with the increase in cluster size.

Answer: Reduces Errors...

ANDU reduces the number of processes (manually entered commands) a user would be required to complete - thus significantly reducing the risks associated with data entry. If ANDU encounters an error, it pauses the upgrade and waits for the user to take appropriate measures to resolve the error. Once corrective measures are taken, the user can resume the upgrade procedure.

How does ANDU work?

Answer: There are three phases to ANDU:

Package Management - this involves getting the new Data ONTAP image into the cluster repository. The image is stored in the mroot of two selected repository nodes in the cluster.

Validation - this phase validates the health of the cluster and its readiness for update. It returns warnings, errors or OK based on the validation checks. Users must fix errors before proceeding and may ignore warnings, based on their discretion.

Update - this phase starts with validation, prompts the user to proceed when validation completes with OK or Warning(s), and starts the process of upgrading the cluster.

What NetApp ONTAP versions support ANDU?

Answer: ANDU is only available starting in Data ONTAP version 8.3.0
ANDU can be used to upgrade systems going from 8.3.0 to any version greater than 8.3.0.

3.3 Hardware Upgrade and Replacement

A growing trend for IT infrastructure is being able to upgrade or replace the hardware in the storage network infrastructure without taking downtime. Specific to NetApp environments, this mainly includes storage controllers and shelves. However, in some situations it is also required to update HBAs, cluster interconnect switches, cables, or drives. In all these cases, a nondisruptive solution for upgrading or replacing (meaning a like-for-like replacement) is preferred.

Can controller hardware be upgraded nondisruptively?

Answer: Yes. Nondisruptive platform (controller) upgrades are supported only when running clustered Data ONTAP 8.0 and later releases. This is possible because you can relocate data from the controllers you want to upgrade to other controllers in the cluster using DataMotion™ for volumes to nondisruptively migrate the storage resources.

Clustered Data ONTAP 8.2 introduces aggregate relocate, which is an enabling capability for nondisruptive in-place controller hardware upgrades. Aggregate relocate is a “no-copy” based reassign of the drives in data aggregates to the partner node of an HA pair without causing disruption to the client or host applications. For more information on ARL, refer to [TR-4146](#) and product documentation available on the NetApp Support site.

LIF migrate is used in conjunction with aggregate relocate to nondisruptively migrate the network resources (LIFs) to alternative nodes within the cluster while the controller is being upgraded.

How do cluster node limits affect controller upgrades?

Answer: The cluster platform mixing rules were put in place to define a cluster size that is optimized for customer usage and facilitates predictable and continuous availability of the cluster. Aggregate relocate can be used to do an in-place upgrade of a controller without affecting the node limits. In the in-place controller upgrade procedure, the original node is cleanly shutdown and removed after its data aggregates are assigned to its HA pair. The new node is installed to replace the original node and booted to commission it and join the cluster. The replacement node accesses the same node root volume used by the original node and in this way, the cluster size never changes. For example, in a four-node cluster (limited to four nodes), node 1 and node 2 are intended to be upgraded to node 5 and node 6. The aggregates owned by node 1 are relocated to node 2 (HA partner node), and then node 1 is shutdown and physically removed. Node 5 is installed to the cluster, completes setup, and assumes ownership of the node root volume and spare drives that were assigned to node 1. The same is done to replace node 2 with node 6 by relocating all aggregates to node 5. During the period of the upgrade, the HA interconnect is temporarily disabled.

If new controllers will be added to the original cluster alongside the original nodes, for example because internal drives are in use which prevents an aggregate relocate-based in-place upgrade, or because shelves are being upgraded in conjunction with controllers, the maximum cluster size cannot be exceeded. DataMotion for volumes can be used to evacuate the original nodes to the replacement hardware.

Can I combine an in-place controller upgrade using aggregate relocate with a Data ONTAP upgrade?

Answer: No. Both the original controllers and the replacement controllers must be running the same version of clustered Data ONTAP 8.2.x or later. If introducing replacement controllers running Data ONTAP 8.2.1, upgrade the original controllers to 8.2.1 before starting the controller upgrade.

Can I exceed the maximum cluster size temporarily in order to perform a hardware refresh?

Answer: No. The maximum number of nodes supported cannot be exceeded to perform lifecycle operations such as a technology refresh of an HA pair. DataMotion for volumes (nondisruptive volume movement, or “vol move”) and the new functionality of aggregate relocate (ARL) in clustered Data ONTAP 8.2 can be used to nondisruptively upgrade controllers within the maximum number of supported nodes.

For FAS2200 series controllers with aggregates utilizing internal drives, the volumes hosted in those aggregates would have to be relocated to another aggregate in the cluster using DataMotion for volumes (not ARL). It is recommended that clusters consisting of FAS2200 series controllers that have aggregates utilizing internal drives consist of no more than two nodes, in order to facilitate nondisruptive controller hardware upgrades using DataMotion for volumes.

For FAS2200 series controllers with aggregates located only on external storage, ARL can be used to conduct in-place controller upgrades. Therefore, FAS2200 series controllers with aggregates on external storage will support nondisruptive controller upgrades on clusters consisting of up to four nodes (provided no internal storage is used).

In general, if DataMotion for volumes is the intended solution for a controller hardware upgrades, then the cluster should reserve space for two additional nodes. If ARL is the intended solution for a controller hardware upgrade, then the cluster can consist of the maximum number of nodes. Refer to [Clustered Data ONTAP Platform Mixing Rules](#) for more information.

For further information regarding ARL, see [TR-4146: Aggregate Relocate Overview and Best Practices for Clustered Data ONTAP](#).

Do the OnCommand management tools support NDO functions?

Answer: OnCommand System Manager 3.0 provides the following NDO capabilities:

- volume move wizard and manual volume move
- LIF send-home
- HA monitoring.

OnCommand System Manager 3.1 adds the following capabilities:

- LIF migrate (for nodes running clustered Data ONTAP 8.2.1 and higher)
- HA takeover and giveback.

OnCommand Workflow Automation 2.1 (WFA) provides the following NDO workflows:

- volume move
- technology refresh for controller and shelf upgrade of an HA pair using volume move to evacuate data from nodes.

Can storage shelves be removed nondisruptively?

Answer: Yes, at Data ONTAP 8.2.1 and higher, shelves containing no data can be removed nondisruptively from nodes. The critical prerequisite is that all aggregates containing disks on the shelves to be removed must be deleted, and disk ownership of all drives in the shelves must be removed, so that the shelves to be removed contain only unowned spare drives. Data volumes can be relocated to other shelves using DataMotion for volumes. If the shelves to be evacuated contain the root volume, it can be copied to disks in other shelves using the `disk replace` command.

Nondisruptive shelf removal is not supported in releases earlier than 8.2.1. Data can be migrated to new shelves within the node or within the cluster nondisruptively using volume move; however, the nodes attached to the shelves being removed would incur a disruption when the shelves are removed.

Can nodes in a cluster be nondisruptively physically moved to a different part of the data center?

Answer: Yes. A cluster can be moved to another physical location within the data center nondisruptively if the distance between the cluster network switch and the controller remains within the 300-meter limit. The nodes being moved would need to be evacuated of all data before the move. The nodes would then be shutdown, moved to their new location, then rebooted to rejoin the cluster. The data volumes can then be redistributed using volume move once again.

3.4 Storage Failover (SFO)

Is failover in clustered Data ONTAP different from failover in 7-Mode?

Answer: Yes. The HA pair controller architecture is very similar in both 7-Mode and clustered Data ONTAP; however, there are a few differences. During a giveback in 7-Mode, all aggregates are returned to the home node at the same time, similar to the way takeover is executed. In clustered Data ONTAP, each data aggregate is returned home serially after the root aggregate has completed giveback and the aggregate's home node has been assimilated back into the cluster. In clustered Data ONTAP 8.2 and higher, a planned failover sends the data aggregates over serially to the partner, followed lastly by the root aggregate.

Can I have data volumes in my node root aggregate? What about an SVM root volume?

Answer: No. The root aggregate should contain only the node root volume, with all data volumes (including SVM root volumes, which are no different in this aspect from any other data volumes) hosted in aggregates separate from the root aggregate.

There are several reasons for this. When an SFO event planned takeover or giveback occurs, the root aggregate is considered separately from other aggregates. In a planned takeover, the non-root aggregates are taken over one by one, while the root aggregate is taken over last. In a giveback operation, the reverse occurs – the root aggregate is given back first, followed by the non-root aggregates one by one. As part of the root aggregate giveback process, the node which had been taken over is assimilated back into the cluster. This synchronization process can take several minutes – well beyond acceptable client or host timeouts. Any data volumes contained on the node root aggregate would be unavailable for access until the synchronization was complete. Volumes in the data aggregates remain online and serve data through the partner node until it their individual aggregate's turn to be given back. Ensuring reliable giveback times is therefore the most critical reason for avoiding data volumes in the node root aggregate.

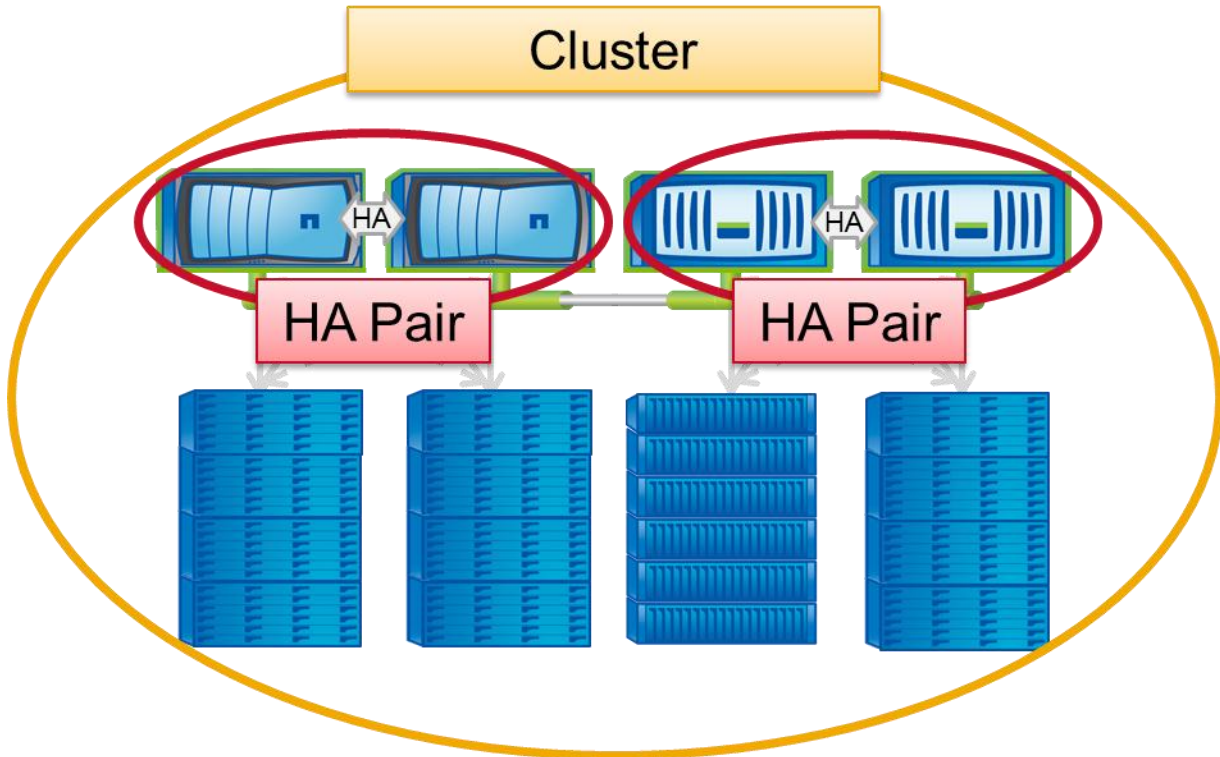
The node root volume contains critical cluster configuration information which must be up to date among all cluster nodes for correct cluster operation, Logs, coredumps and other diagnostic information are also stored in the node root volume. It is vital that there be sufficient space and performance for the node root volume to operate correctly. Data volumes in the same aggregate would compete for storage space and I/O resources with the node root volume, which is a critical system resource. For more information on requirement for a dedicated root aggregate on each node, see <https://kb.netapp.com/support/index?page=content&id=3013563>.

How do high-availability (HA) pairs work in a clustered Data ONTAP configuration?

Answer: Clustered Data ONTAP 8.0 enables the scale-out of multiple HA pairs into a single cluster. Previously, the term cluster was synonymous with a 7-Mode HA pair, but clustered Data ONTAP clearly delineates a difference between an HA pair and a cluster. An HA pair consists of the two partner nodes. A cluster is made up of multiple HA pairs. The HA pairs are joined by a single network common to the cluster over an internal cluster interconnect.

The HA pair controller configuration maintains the same resiliency as a 7-Mode system. 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.

A cluster can consist of a single HA pair. Such clusters, with only two nodes, require the enablement of cluster HA, a command line option. This allows continuous monitoring of the nodes' state; if either node goes down and a failover occurs, the remaining node has full read and write access for all the storage, logical interface, and management capabilities. A cluster consisting of more than two nodes does not use the cluster HA feature.



Does hardware-assisted takeover use the cluster interconnect or the HA interconnect?

Answer: Hardware-assisted takeover in both Clustered Data ONTAP and 7-Mode uses the HA interconnect.

Does a single-node cluster have HA? What about other NDO capabilities?

Answer: No. In clustered Data ONTAP 8.2 and higher, it is possible to configure a single-node cluster with no cluster LIFs and no HA partner. A single node cluster by definition is not part of an HA pair and does not have SFO capability. It does not meet the minimum requirement for NDO. Data ONTAP cannot be nondisruptively upgraded, ARL is not supported, vol move is obviously restricted to aggregates within the single controller and LIFs can only migrate to controller's own ports.

Is HA pair failover completely nondisruptive?

Answer: It depends. Failover is intended for resiliency during planned and unplanned events such as NDUs, a hardware failure, or a software failure. In general, HA pair controller configurations are architected for failover to prevent disruption to client and host applications. However, there are a few scenarios to consider that are disruptive in the event of an HA failover:

- For the majority of CIFS environments, failover of the storage causes disruption to CIFS clients. The exception to this Data ONTAP 8.2, SMB 3.0 Continuously Available File Shares that are servicing requests from a server running Hyper-V where takeover and giveback are nondisruptive. See [TR-4100: Nondisruptive Operations and SMB File Shares for Clustered Data ONTAP](#) for more information.
- A failover when the partner node is already down or HA has been disabled is disruptive.

For clustered Data ONTAP systems that service applications using the CIFS protocol (except for SMB 3/Hyper-V Continuously Available Shares), disruption occurs when any of the following is true:

- The file being accessed resides on a node that has to fail over to its partner node due to a planned or an unplanned event.

- The file being accessed resides on the storage of a node that has been failed over, and a subsequent giveback operation is required.
- The clients are running SMB 1.0, and a LIF is migrated or failed over to another node within the cluster.

See [TR-4100: Nondisruptive Operations and SMB File Shares for Clustered Data ONTAP](#) for detailed scenarios of client behavior in the event of node failover.

What is cluster quorum, and how does it affect data availability?

Answer: Cluster quorum is the state of the cluster when the majority of the nodes in the cluster are active. If a node in the cluster becomes inactive, it might affect cluster quorum. A cluster which is out of quorum will not serve data.

The rules for quorum to be maintained in a cluster of greater than 2 nodes are:

- More than half the nodes are active in the cluster.
- At least half the nodes in the cluster are active, including the node holding epsilon. Epsilon is assigned to one node and counts as a fractionally higher vote in the cluster compared to the nodes not holding epsilon. This is to prevent split-brain scenarios in a cluster. Epsilon is not used in a two-node cluster – see the following question.

When a cluster is in quorum, configuration changes can happen without limitation. All data is available, unless an HA pair is down and storage failover has not occurred. For example, in a cluster of six nodes, it is possible (though unlikely) to have both nodes in an HA pair go down while cluster quorum is still maintained because there are still four active nodes in the cluster. Although all quorum-related activities might continue, the storage attached to the HA pair that is down is not available.

Alternatively, consider the following four-node cluster and what would happen if one node or two nodes went down. The scenarios and their affect on cluster quorum in a four-node cluster, are detailed in Table 6.

Figure 4) Example four-node cluster for epsilon.

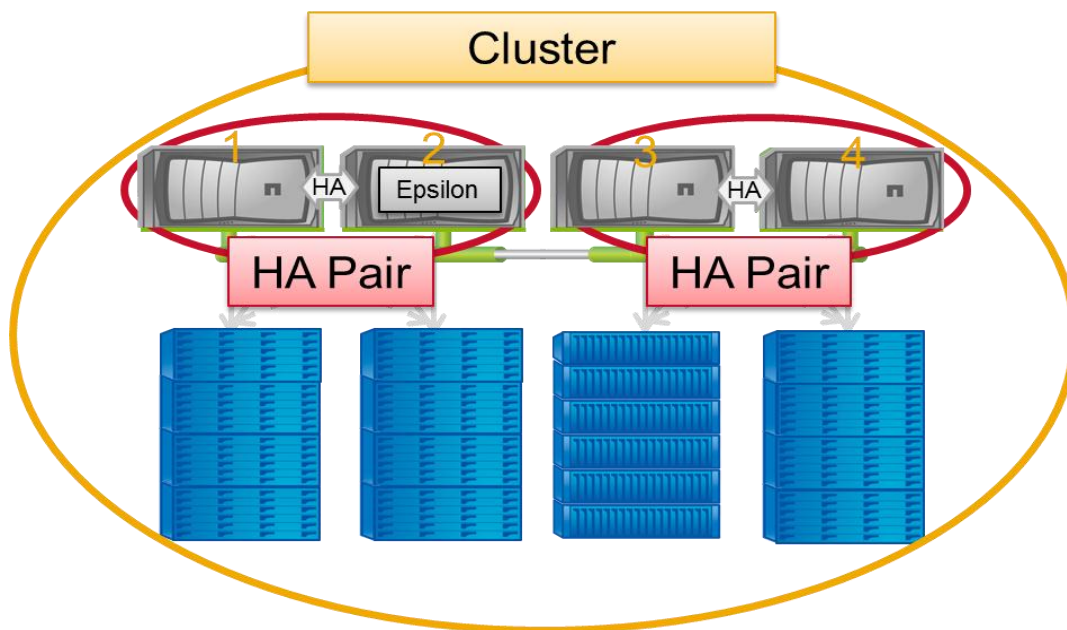


Table 6) Examples for understanding epsilon in a four-node cluster.

Scenario	Is Quorum Maintained?	Rationale
One node is down in the cluster.	Yes	More than half of the nodes in the cluster are up and participating in the cluster.
Two nodes in the cluster are down, but the node containing epsilon is participating in the cluster.	Yes	Half the nodes in the cluster are active, and epsilon is held by a node still active in the cluster.
Two nodes in the cluster are down, but the node containing epsilon is NOT participating in the cluster.	No	Half the nodes in the cluster are active, but epsilon is gone because the node holding epsilon is down.
Three nodes in the cluster are down, one node up – regardless of whether the up node holds epsilon.	No	More than half the nodes in the cluster are not participating in the cluster, epsilon is irrelevant

Does a two-node cluster use epsilon?

Answer: No, epsilon is not used in a two-node cluster, since if one node was weighted greater than the other, then takeover could only occur in one direction. In a two-node cluster, “cluster HA” must be enabled. Cluster HA is only used in a two-node cluster and must be disabled if the cluster grows beyond two nodes. Cluster HA enables the two nodes to continuously poll one another to maintain quorum. If either node goes down, the remaining active node takes over and a quorum of one node is formed. The surviving node assumes read-write access to all the storage until the down node returns and the storage is given back. Storage failover (SFO) is enabled in all two-node and higher clusters.

What happens in a two-node cluster if the cluster interconnect fails?

Answer: The cluster interconnect is a critical component of the cluster architecture and is architected for resilience and reliability. Requirements for supported NICs and ports, cluster inter-connect switches and switch configuration must be strictly adhered to when ordering and configuring a cluster to ensure no single point of failure in the infrastructure

As previously stated, in a 2-node cluster, there is no epsilon – cluster HA must be enabled to allow either node to takeover for the other and maintain quorum if only one of the nodes is up. Takeover is required to maintain quorum.

CLAM (Cluster Liveliness Availability Monitor) introduced with SAN support in clustered Data ONTAP 8.1, monitors heartbeats over the cluster interconnect, regardless of whether SAN protocols are licensed and configured.

If heartbeats are dropped because of cluster interconnect failure (switch failure, cables cut etc) or one node becomes isolated (all its cluster ports or LIFs fail), both nodes drop out of quorum. CLAM detects the out of quorum condition and logs appropriate messages . The timeout period for lost interconnect heartbeats is around 22.5 seconds. In a cluster with 2 cluster interconnect LIFs, this would only occur *after dual or more failures*. A one-cable two-node switchless cluster (switchless FAS2220 or FAS2240 with clustered Data ONTAP 8.2 only) however has a *single point of failure which should be carefully positioned if proposing this configuration to customers*. It does not meet the NDO Minimum Configuration.

Regardless of the number of interconnects, there are two possibilities for what occurs next.

1. If at least one SAN object is configured (LUN, portset, etc), then CLAM induces a takeover by one of the nodes after the cluster interconnect timeout. The taking over node must be otherwise healthy and capable of performing the takeover. The node which took over is now in a "quorum of one" as for any other takeover, and full data services, NAS and SAN, are resumed on this node. After fixing the issue with the cluster interconnect, the administrator can perform a giveback to resume normal two-node operation.
2. If the cluster has NO SAN objects, then no automatic takeover occurs, both nodes remain out of quorum and neither will serve data. Manual intervention will be required to either restore the cluster interconnect or perform a takeover.

Can I add an additional or dedicated node to host epsilon?

Answer: No, clusters of two or more nodes must consist of HA pairs. A single node as an epsilon tie-breaker is not supported.

Does epsilon move to another node if the original node holding epsilon goes down?

Answer: No, and epsilon does not transfer if a node is taken over. However this is not typically a concern for a temporary takeover of short duration such as during an NDU. However, particularly for 4-node clusters, moving epsilon can provide additional resilience in the event of a second node failure when the original node holding epsilon is down.

In what circumstances should epsilon be manually moved to another node?

Answer: If the node holding epsilon will be down for an extended period of time, it is recommended to manually move epsilon to its HA partner. This will provide additional resiliency in the event of a second node failure while the original node is down. Moving epsilon to the HA partner protects cluster quorum and preserves data access if a node in the other HA pair fails. Epsilon can be reassigned if the node holding epsilon is already down, but only if the cluster is still in quorum. Epsilon cannot be reassigned in a cluster which is out of quorum. It is recommended to move epsilon to the HA partner node when performing the following operations on the node holding epsilon:

- Unjoining the node from the cluster
- For any planned maintenance event including aggregate relocate or hardware reconfiguration
- If the node is down for longer than a normal reboot sequence.

Are there additional resources available for HA pairs and storage failover?

Answer: Yes. For more information on HA pair controller configurations and best practices, refer to [TR-3450: High-Availability Pair Controller Configuration Overview and Best Practices](#).

4 Cluster Scalability

4.1 Cluster Growth and Operations

Can a single-node cluster upgrade to a two-node cluster nondisruptively?

Answer: No. A single-node cluster does not meet the minimum configuration requirements for NDO. The absence of an HA partner node in a single-node cluster prevents any resiliency for planned or unplanned events such as reboot, Data ONTAP upgrade or other maintenance operations. If the single-node cluster needs to be expanded to scale out to two or more nodes, there are two factors which dictate a reboot of the node, with associated downtime.

- If the single node does not already have available supported 10GbE ports for the cluster interconnect, it will be need to be shutdown for hardware installation, provided the node is capable of installing the required NICs. A single-node FAS2240 with a FC card in the mezzanine

slot cannot add 10GbE ports and therefore cannot be upgraded to a two-node cluster. Similarly a single-node FAS2240 using both 10GbE ports in the mezzanine slot for client/host traffic also cannot be upgraded to a two-node cluster.

- A single-node cluster's NVRAM is optimized and dedicated for just one node. In order to add a second node, it needs to be rebooted to split the NVRAM for partner mirroring to enable storage failover.

Can a two-node switchless cluster be upgraded to a switched cluster nondisruptively?

Answer: Yes. A two-node switchless cluster can be upgraded to a switched cluster if it already uses two cluster interconnect connections between the two nodes. The integration of a switch into the cluster can be done using a series of LIF migrations that migrate all the LIFs from one of the cluster interconnect paths to the redundant cluster interconnect path. The switch is then cabled into the path that does not have any cluster interconnects hosted on it. After the switch is cabled into one path, all the LIFs are migrated from the redundant path, and the switch is cabled into the redundant path.

If the two-node switchless cluster has only a single interconnect (FAS2220 and FAS2240 configurations only – known as the one-cable two-node switchless cluster), a cluster interconnect LIF on the second 10GbE port must first be configured before upgrading to a two-node switchless cluster. Cluster interconnects require the ports to be dedicated, hence if the second 10GbE port is needed for data connections, then this configuration cannot be upgraded to a switched cluster or expanded beyond two nodes.

Is a switch required to grow a cluster beyond two nodes?

Answer: Yes. The switchless cluster configuration is available for two-node clusters only and a switch must be introduced in order to grow to four nodes and beyond.

Can nodes be joined to the cluster nondisruptively?

Answer: Yes. Nodes can be added to a cluster without disrupting client/host traffic, provided the new cluster size does not exceed the platform mixing rules. If a cluster is already at its maximum permitted size, nodes must be unjoined from the cluster (after first evacuating all data and deleting its resources) before adding the new nodes to the cluster. As each new node is added, it will synchronize its configuration from the existing nodes. Aggregates can then be created on the node's storage and volumes can be moved to these aggregates, or new volumes provisioned as required..

Can nodes be unjoined from the cluster nondisruptively?

Answer: Yes. Nodes can be nondisruptively unjoined from the cluster. Any data residing on the nodes should be nondisruptively moved, using volume move, to other nodes within the cluster. Aggregates will need to be deleted and LIFs either deleted or rehome to other nodes before removing the node from the cluster.

Can two separate clusters be merged together as a single cluster nondisruptively?

Answer: No. It is not possible to join two clusters together; only nodes with no data can join a cluster. To join nodes from cluster B to cluster A, the nodes from cluster B need to be unjoined and wiped clean, losing all the existing data. The clean nodes can then join cluster A as new nodes. Cluster A's original nodes and clients/hosts will not be affected but all data from cluster B will be lost. Data, LIFs and other configuration from the original cluster B need to be re-created and the clients and hosts then reconnect to the expanded cluster A.

Can a cluster be split into two separate clusters nondisruptively?

Answer: No. A cluster cannot be split into two separate clusters. Nodes would need to be unjoined from the cluster , and wiped clean. They can then be set up as a separate new cluster with a new cluster ID and license.

References

The following references were used in this FAQ:

- [Clustered Data ONTAP Platform Mixing Rules](#)
- [Nondisruptive Upgrade Technical FAQ](#)
- [TR-3975: DataMotion for Volumes Overview](#)
- [TR-4075: DataMotion for Volumes Best Practices and Optimization](#)
- [TR-3450: High-Availability Pair Controller Configuration Overview and Best Practices](#)
- [TR-4100: Nondisruptive Operations and SMB File Shares for Clustered Data ONTAP](#)

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®](#)