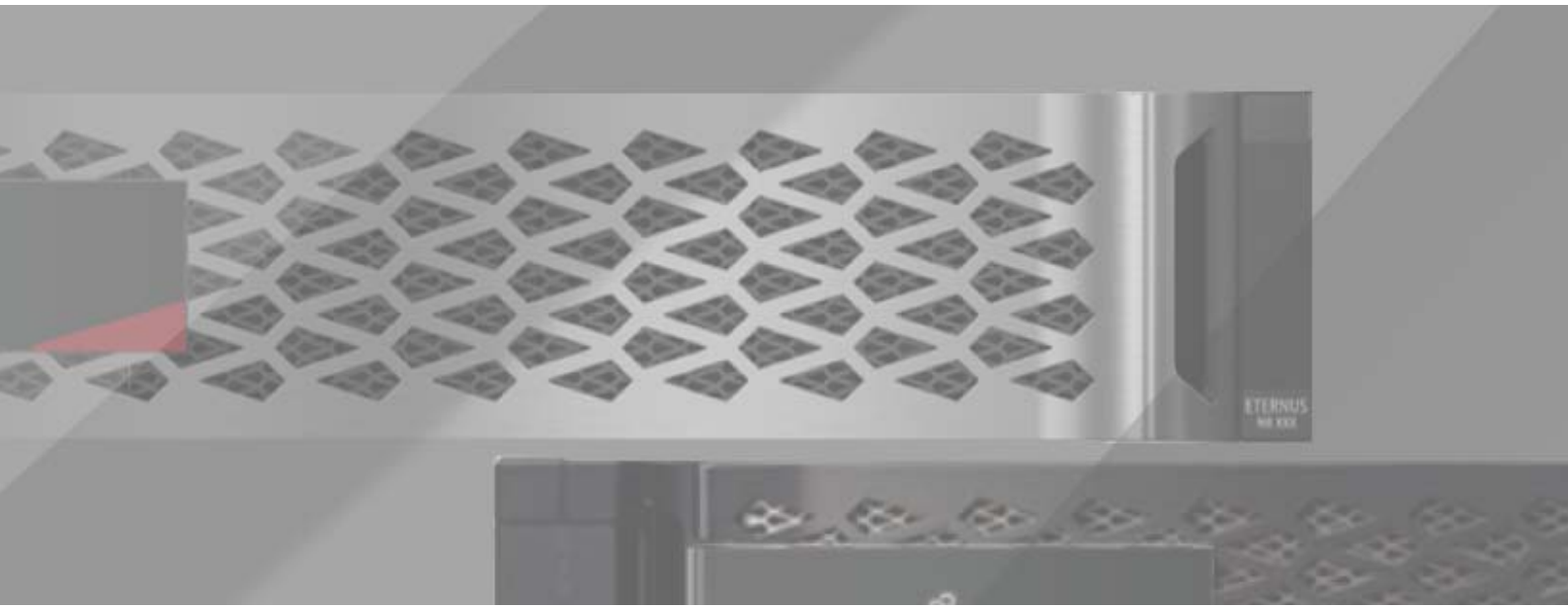
P3AG-6332-01ENZ0

FUJITSU Storage
ETERNUS AX series All-Flash Arrays,
ETERNUS HX series Hybrid Arrays

Best Practices for ONTAP SAN Configurations



ONTAP 9

FUJITSU

Table of Contents

9. Scalable SAN Key Value Propositions and Features 43

List of Figures

List of Tables

Preface

This document provides an overview of and best practices for the SAN of the ETERNUS AX/HX series.

First Edition
November 2021

Trademarks

Third-party trademark information related to this product is available at:
https://www.fujitsu.com/global/products/computing/storage/eternus/trademarks.html

Trademark symbols such as ™ and ® are omitted in this document.

About This Manual

Intended Audience

This manual is intended for system administrators who configure and manage operations of the ETERNUS AX/HX, or field engineers who perform maintenance. Refer to this manual as required.

Related Information and Documents

The latest information for the ETERNUS AX/HX is available at:
https://www.fujitsu.com/global/support/products/computing/storage/manuals-list.html

Document Conventions

Notice Symbols

The following notice symbols are used in this manual:

Caution	Indicates information that you need to observe when using the ETERNUS AX/HX. Make sure to read the information.
Note	Indicates information and suggestions that supplement the descriptions included in this manual.

1. Overview

This document presents an overview of the clustered SAN of the ETERNUS AX/HX series from the view point of SAN-attached hosts. It also describes the use of SAN configurations with the ETERNUS AX series to optimize performance. Additionally, it describes best practices for leveraging the high-availability and data mobility features of ONTAP.

Audience

This document is intended for system and storage architects who design iSCSI, FC, and other solutions with storage solutions running ONTAP 9.7 or later. It assumes that the reader:

- Has a general knowledge of hardware and software solutions
- Is familiar with block-access protocols such as FC and iSCSI

Caveats

This document is not meant to be a general introduction to ONTAP administration. An introduction is covered by the ETERNUS AX/HX Series System Administration Reference and the ETERNUS AX/HX Series SAN Administration Guide. SAN-related limits for ONTAP clusters that use SAN protocols can be found in the ETERNUS AX/HX Series SAN Configuration Guide. These documents can be found in the Fujitsu manual site.

2. Summary of Best Practices

For more information about each of the best practices, review the following links:

- Create the FCP or iSCSI service at the same time as creating an SVM.
  "Considerations for Optimizing SAN Performance"
- Combine LUNs that are related, have similar performance requirements, and management requirements into the same volume to reduce administrative effort, operate as a consistency group, and to maximize storage efficiency.
  "Considerations for Optimizing SAN Performance"
- Optimize performance by increasing the number of volumes and increase the number of LUNs. The ideal lay-out in most cases will be approximately 8 volumes and 8-16 LUNs.
  "Considerations for Optimizing SAN Performance"
- Volumes containing LUNs do not need to be junctioned to a namespace in order to serve data using FCP or iSCSI.
  "Volume Configuration"
- An SVM should have one LIF per Ethernet network or Fibre Channel fabric on every storage controller that is going to serve data using iSCSI or Fibre Channel.
  "Path Selection"
- Fibre Channel fabrics attached to a clustered ONTAP storage must have N_Port ID virtualization (NPIV) enabled.
  "FC and NPIV"
- Use only NPIV virtual worldwide port names (WWPNs) as targets in Fibre Channel fabric zoning configurations. The target ports' physical WWPNs should not be used.
  "FC and NPIV"
- Selective LUN mapping means that most LUNs have four paths, two direct and two indirect, corresponding to the storage controller and its high-availability (HA) partner, respectively. In this default case, change LUN mappings whenever moving a LUN to a new HA pair in the same cluster.
  "Management Interfaces"
- Create more paths as needed, either to facilitate data mobility operations or to leverage additional I/O resources, but do not exceed the maximum number of paths a host OS can support.
  "DataMotion and Selective LUN Mapping: Discovering and Discarding Paths"
- Follow a standard procedure on hosts when changing LUN mappings so that they discover new paths and discard paths that have been removed.
  "DataMotion and Selective LUN Mapping: Discovering and Discarding Paths"
- SVMs serving data with FCP or iSCSI need an SVM management interface.
  "Management Interfaces"
- When using LUN move on a LUN that is deduplicated or compressed, make sure that the destination volume also has these policies enabled.
  "Storage Efficiency Considerations"
- If LUN move is used to move a LUN out of a volume protected by software used to manage Snapshot copies, that software should be aware of the possibility of LUNs with nonlocal Snapshot copies.
  "Data Protection Considerations"
- Make use of LUN move's pause and throttle features for more granular control over LUN replication.
  "Scalability and Throughput Considerations"

- Use LUN move to shorten existing data mobility and replication workflows.
  "Data Management and Workflow Considerations"
- Do not exceed the cluster size limit for clusters serving SAN data, as specified in the ETERNUS AX/HX Series SAN Configuration Guide in the Fujitsu manual site.
- Use Active IQ Unified Manager to verify that all nodes in the cluster are at or below their performance capacities; use vol move to rebalance any workloads that are on performance capacity overprovisioned nodes.
  "Performance Capacity"
- Set guarantee In Order Delivery on all Fibre Channel switches in your fabrics.
  "FC In-Order Delivery"

3. SAN Features

ONTAP 9 has several SAN-related features as shown below:

- Cluster hardening.
- Improvements in quorum handling, including several auto-heal features.
- Enhancements to first failure detection and reliability, availability, and serviceability (RAS).
- Foreign LUN Import (FLI) to verify performance improvements. The imported LUN verification workflow has some performance improvements that allow verifications to complete more rapidly.
- FLI hardening. There has been some hardening in FLI to allow imports to survive and be restarted after a take-over or giveback, an event that previously would have required a restart of the import.
- These enhancements fall into two primary categories:

These enhancements fall into two primary categories:

- **Hardening and resiliency**
  These are enhancements made to ONTAP code to make ONTAP more resilient in the face of faults. In many cases ONTAP can encounter a fault and heal itself without any human intervention.
- **RAS**
  These enhancements are primarily categorized around creating messaging and gathering appropriate counters so that faults are more easily isolated, recognized, and diagnosed. This allows storage administrators and Fujitsu Support to more quickly diagnose and resolve fault conditions, often before there is any externally recognizable impact.

Additional features that are noticeable and likely to be compelling to storage architects, professional services, and storage administrators include:

- Fast failover
- Performance capacity
- Low-latency consistent performance
- Intercluster copy offload
- NVMe over Fabrics (NVMe-oF) protocol, NVMe over Fibre Channel (NVMe/FC) support
- SnapMirror Synchronous
- Asymmetric Namespace Access (ANA) support
- 512-byte block size support
- VMware Compare and Write support for NVMe
- NVMe nondisruptive volume move support
- NVMe QoS support

These features are described in the following sections.

Fast Failover

ONTAP 9 continues code optimizations and enhancements that have reduced the time ONTAP HA pairs require to take over and give back partner workloads. When operating in an ETERNUS AX series SAN configuration, planned takeovers and givebacks complete within 2 to 10 seconds. Unplanned takeovers and givebacks complete within 2 to 15 seconds. Both of these measurements are the time it takes for one node of the HA pair to fail over or give back from the other node in the HA pair. In testing, most operating system stacks resumed I/O from 4 to 7 seconds after the takeover or giveback event started.
Table 1 summarizes the failover times by platform type and version of ONTAP. All the numbers assume that the HA pairs are in a prescribed SAN configuration as defined by the ETERNUS AX/HX series Best Practices for Business-Critical Workloads in the Fujitsu manual site.

Table 1 Fast failover takeover and giveback timing guidelines

Platform	Planned takeover	Unplanned takeover
ETERNUS AX series ONTAP 9	10 seconds	15 seconds
ETERNUS HX series with Flash Pool or SSD aggregates	30 seconds	60 seconds

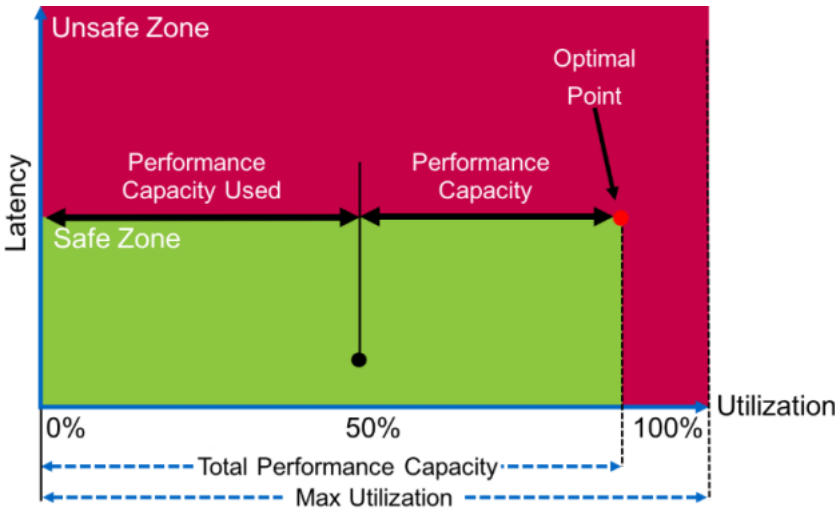
Performance Capacity

Performance capacity is a feature that uses counter manager statistics gathered by the controller, which are then consumed by Active IQ Unified Manager to analyze performance counters to dynamically optimize for the maximum IOPS that can be produced while maintaining consistent low latency. This means that storage administrators no longer have to guess whether there is sufficient remaining performance potential on a controller or HA pair to add additional workloads.

In an ETERNUS AX series SAN business-critical configuration, Active IQ Unified Manager can use the ONTAP generated performance capacity calculations to maximize IOPS while maintaining <1ms latencies.
Latencies might be somewhat higher when there is a failover; obviously, this needs to be taken into account in workload planning.

Performance capacity data and Active IQ Unified Manager visualizations can be used to optimize IOPS while keeping performance in takeover within acceptable latencies. For more information about performance capacity and using Active IQ Unified Manager to optimize storage performance while maintaining consistent low latencies, review the ETERNUS AX/HX series Best Practices for Business-Critical Workloads in the Fujitsu manual site. Figure 1 illustrates performance capacity. It shows the safe and unsafe zones (as defined by latency) and also identifies current performance capacity used and the optimal point. The optimal point is the point that maximizes IOPS while maintaining the latency target. From the graphic, it is easy to see that performance capacity is the amount of performance capacity remaining when capacity used is subtracted from the optimal point.

Figure 1 Performance capacity showing performance capacity used and optimal point



Low-Latency Consistent Performance

ONTAP 9 is able to achieve and guarantee consistent low-latency performance when ETERNUS AX series prescriptive SAN configuration guidelines are followed in conjunction with using performance capacity and Active IQ Unified Manager.

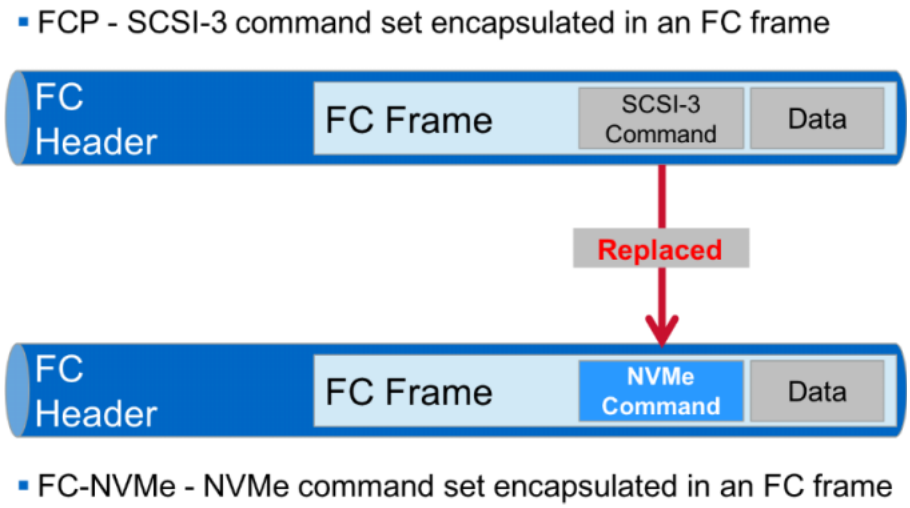
Intercluster Copy Offload

Intercluster copy offload is a feature of ONTAP 9 that allows Windows to nondisruptively offload copying blocks between LUNs, which significantly reduces CPU load, network bandwidth, and move durations, while maintaining access to the LUNs, which remain online throughout the move. This feature is an extension of Microsoft Windows Server's copy offload feature (ODX). In SAN, it is supported only with Windows LUNs.

NVMe over Fabrics (NVMe-oF) Protocol, NVMe over Fibre Channel (NVMe/FC) Support

An NVMe over Fabrics (NVMe-oF) protocol–NVMe over Fibre Channel (NVMe/FC) is supported. NVMe/FC encapsulates the NVMe command set inside an FC frame, replacing the existing SCSI-3 command descriptor block, as shown in Figure 2.

Figure 2 FC versus NVMe/FC frames



The new NVMe command set is:

- Command streamlining
- Removing all software locks
- Reduced context switches
- Increased multithreading-64K queues with a maximum queue depth of 64K

These optimizations have created a much more efficient, and therefore, high-performing protocol that improves throughput and reduces latencies for workloads by simply replacing the block protocol used. There is no need to rewrite applications to gain the benefits of the new protocol.

SnapMirror Synchronous

SnapMirror Synchronous (SM-S) enables synchronous replication. MetroCluster provides synchronous replication, however, it requires all data to be mirrored. The synchronous replication feature provides customers the ability to choose which workloads needed synchronous replication. SM-S has the following attributes and features:

- Volume granular, synchronous data replication for FC and iSCSI
- Zero recovery point objective (RPO) and very low recovery time objective (RTO). Not a business continuance solution.
- No additional external hardware, software, or networking is required.
- Application I/O to the primary volume is not disrupted if replication errors occur-automatically recovers and resyncs after replication failures are corrected.
- Guaranteed 100% synchronicity between primary and secondary volume in strict mode.
- SM-S is a better choice for customers who want to granularly manage replication and data protection at the volume level rather replicating the entire cluster (use MetroCluster in that case)
- SM-S is an additional license in addition to the required SnapMirror license.

Asymmetric Namespace Access (ANA) Support

ANA is introduced as part of the NVMe/FC target. As with asymmetric logical unit access (ALUA), ANA uses both an initiator-side and target-side implementation for it to be able to provide all the path and path state information that the host-side multipathing implementation to work with a storage high-availability multipathing software used with each OS stack. ANA requires both the target and initiator to implement and support ANA to function. If either side is not available or implemented, ANA isn't able to function, and NVMe/FC falls back to not supporting storage high availability. In those circumstances, applications must support high availability for redundancy.
NVMe/FC relies on the ANA protocol to provide multipathing and path management necessary for both path and target failover. The ANA protocol defines how the NVMe subsystem communicates path and subsystem errors back to the host so that the host can manage paths and failover from one path to another. ANA fills the same role in NVMe/FC that ALUA does for both FCP and iSCSI protocols. ANA with host OS path management such as multipath I/O (MPIO) or Device Mapper Multipathing (DM-Multipath) provide path management and failover capabilities for NVMe/FC.

512-Byte Block Size Support

In addition to the natively supported 4k block size by ONTAP, a 512-byte block size option is supported by all the items including NVMe namespaces. This feature can simplify integration with existing 512-byte configurations, such as VMware datastores and Oracle Automatic Storage Management (ASM) disk groups that use 512-byte blocks. 4096-byte (4k) blocks remain the default. However, a new block size argument and both 512-byte and 4096-byte values are available for both the `vserver nvme namespace create` command and the matching API.

VMware Compare and Write Support for NVMe

VMware Compare and Write/Atomic Test and Set (CAW/ATS) fused operation is supported for ONTAP to support features such as VMware Storage vMotion. The ATS primitive uses CAW to fuse the NVMe compare and the NVMe write commands to first perform a compare operation. If the compare operation is successful, then the write completes; if it fails, then the write is aborted.

NVMe Nondisruptive Volume Move Support

NVMe-oF adds volume move support, which allows storage administrators to nondisruptively move volumes that contain NVMe namespaces from one aggregate to another.

NVMe QoS Support

With NVMe QoS support, nvmf.qos.mismatched.policy, which is an event management service (EMS) event, can be used. The EMS warns a user if there is a policy mismatch. This event is soft enforced and only verified during a vserver nvme subsystem map add operation. The following limitations apply to this event:

- All namespaces in a subsystem must reside in volumes that have the same QoS policy.
- All namespaces in a subsystem must reside in volumes that have the same QoS throttling policy to work correctly.

4. Common Features of SAN and NAS

Minimum Quality of Service

QoS maximums allow the administrator to assign a maximum number of IOPs or the data rate that can be assigned to a workload. This prevents the workload from starving adjacent workloads when it spikes. QoS minimums operate in the opposite way. When a minimum is associated with a workload, I/O from that workload is prioritized ahead of other I/O so that it can meet its performance requirements.
Working together, minimums and maximums can effectively shape and prioritize workloads in a shared environment. QoS Minimums are available on SAN objects (volumes and LUNs) on ETERNUS AX series platforms.

Balanced Placement

Balanced placement enhances the existing simplified application provisioning. It optimizes initial workload placement by using two metrics to determine the optimal location for a newly provisioned workload by analyzing the amount of space available on each of the cluster nodes and the performance capacity remaining on all nodes to make a workload placement decision that optimizes based on available resources.
Here's a summary of the approach that is used.
Each aggregate in the cluster is assessed to determine currently available (free) IOPS (performance capacity remaining/headroom). This is done by starting with optimal maximum IOPS for the aggregate from QoS headroom calculations, then subtracting the maximum of either the current IOPS for the aggregate (again from QoS headroom) or modeled IOPS (the sum of modeled IOPS for all volumes in the aggregate). Modeled IOPS are either 10% of maximum IOPS (set by QoS policy on the volume) or 10% of the IOPS/TB times volume size based on the best-fitting storage service (when no QoS policy is set).
The aggregates are sorted in a list by drive type (favoring lower cost), by free IOPS, and finally by free capacity. The list is walked and the first aggregate that meets the IOPS and capacity requirements is selected.
After volumes are provisioned, QoS maximum is set based on the defined service level.

The following enhancements are introduced for balanced placement:

- Support for adaptive QoS.
- Increased number of volumes supported in the AppDM created application object from 10 to 16.
- Balanced placement introduces the ability to modify the storage service level commitments of an existing app-component, and reject the request if it can't be met in-place. If the service level commitment request can't be met based on available resources, then the change is blocked and the storage administrator must either add resources, manually adjust resource placement, or discard the requested change.

5. ONTAP 9.7 New Features

All SAN Array (ASA)

ASA is a SAN-only HA pair that is symmetric active/active. This means that the ASA advertises and uses paths through both controllers to any underlying LUN. The effect of this architectural change is to guarantee that hosts will always have an active path to any LUN hosted on either controller in the HA pair. This is important because it makes otherwise disruptive planned or unplanned failovers and givebacks virtually instantaneous. This brings frame array functionality to a modern modular array architecture without compromises such as requiring one of the two nodes being stranded in a standby mode. For more information, see the ETERNUS AX/HX series Best Practices for Business-Critical Workloads in the Fujitsu manual site.

6. ONTAP 9.8 New Features

ONTAP 9.8 adds several new features, some of which are available on both unified and ASA platforms. Other features are initially being introduced on the SAN-only ASA. The expectation is that these features will likely be added to unified ONTAP clusters in upcoming releases of ONTAP. ONTAP's new SAN features include:

- Virtual machine ID (VMID)—a virtual machine (VM) telemetry enhancement
- NVMe-oF protocol coexistence

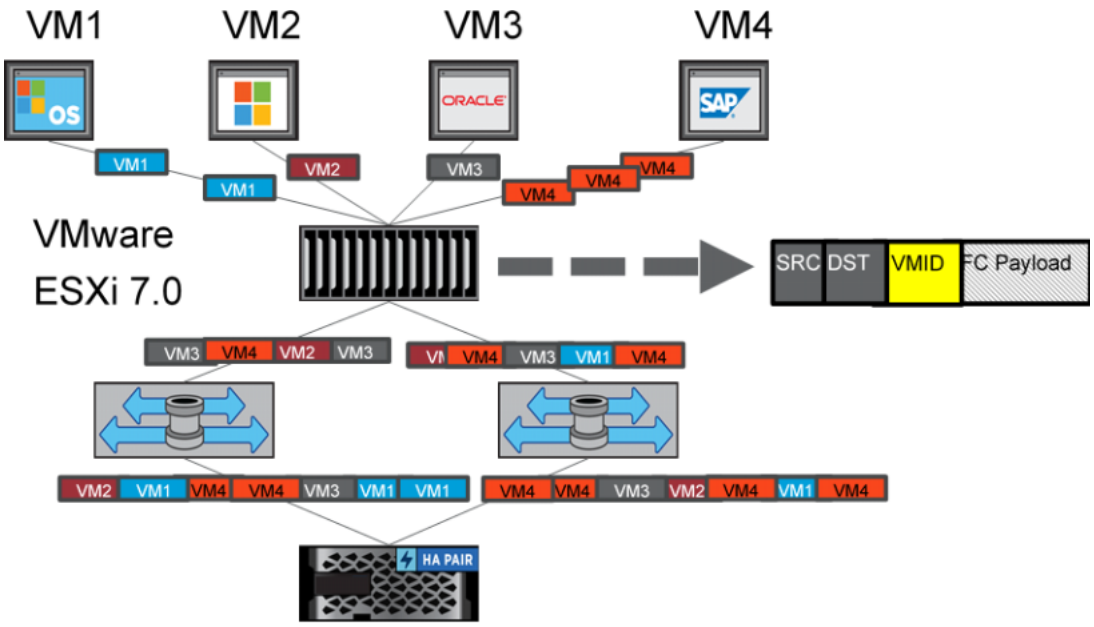
Other new features initially offered on the ASA only are:

- Larger maximum LUN size
- Persistent ports

Virtual Machine ID (VMID)

VMware generates a globally unique identifier for each of the VMs it is hosting. It writes these to the header field to tag each FC frame so that it can be correlated to a specific VM. This allows administrators to identify and follow I/Os from each VM using a shared FC LUN-backed datastore. Prior to the VMID feature, the highest possible level of granularity was the ability to identify the datastore with which the I/O was associated. With VMID, administrators can identify and track I/O characteristics for each of the VMs sharing a datastore individually. This functionality allows for end-to-end QoS, which allows for significantly better insight into traffic patterns, workload characteristics, substantially enhanced troubleshooting, and more detailed VM traffic analysis and reporting. This functionality is initially supported with Brocade switches (Gen 6 and later).

Figure 3 VMID



Tracking I/O from each VM in a shared datastore through an FC fabric includes these steps:

1. Hypervisor assigns a globally unique ID to each VM.
2. The VMID is tagged to each frame from the VM.
3. The switch and storage nodes propagate and reflect each frame and VMID.

NVMe-oF Protocol Coexistence

ONTAP 9.8 removed the requirement that NVMe/FC be segregated from other block and file protocols in its own SVM. This removal was initially done to speed the release of NVMe/FC by reducing the required QA regression testing required to release NVMe/FC. By segregating protocols, NVMe-oF protocols were released because engineering QA teams didn't need to test for NVMe-oF impacts on other protocols such as FCP, iSCSI, NFS, SMB, or S3. The regression testing necessary for protocol coexistence was added to QA testing plans and was performed for ONTAP 9.8 and for all subsequent releases of ONTAP.

Larger Maximum LUN Size

ONTAP 9.8 increased the maximum LUN size from 16TB to 128TB. This increase also has a corresponding increase in size for volumes from 100TB to 300TB. The expectation is that the most popular use of those larger LUNs is going to be for LUNs backing hypervisor data stores.

Persistent Ports

Persistent ports reduce the impact of takeovers. They are able to do this by creating a shadow LIF on the corresponding physical port on the HA partner. When a node is taken over, the shadow LIF on the corresponding partner node assumes the identity of the original LIF, including the WWPN.

Persistent ports work because they are able to advertise the shadow LIF as active optimized (AO) to the host MPIO stack before it changes the previous path status to down (faulty).
The host MPIO stack shifts I/O to the next AO path (formerly the shadow LIF) so that any I/O disruption is minimized. The host always sees the same number of paths to the target regardless of the target's state (steady state or in takeover).
Persistent ports are introduced with ONTAP 9.8 on ASA. When a node is upgraded to ONTAP 9.8, this feature is enabled by default.
Best practices for persistent ports require that FCP port characteristics must be identical within the HA pair:

- FCP port counts
- FCP port names
- FCP port speeds
- FCP LIF WWPN-based zoning
- Both the active and shadow LIF need to be in the same zone with the initiator

If any of these best practices are violated, an EMS message is generated with the following body:

```
EMS : scsiblade.lif.persistent.ports.fcp.init.error
```

The persistent ports feature is available with FC but not iSCSI. It requires that zone membership be identified by WWPN, since the WWPN is essentially spoofed on the shadow LIF also.
Figure 4 shows the persistent ports.

Figure 4 Persistent ports

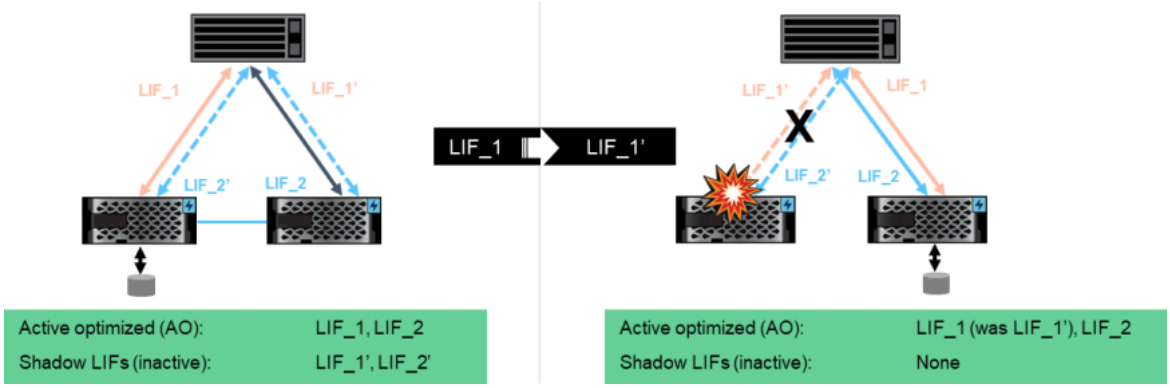
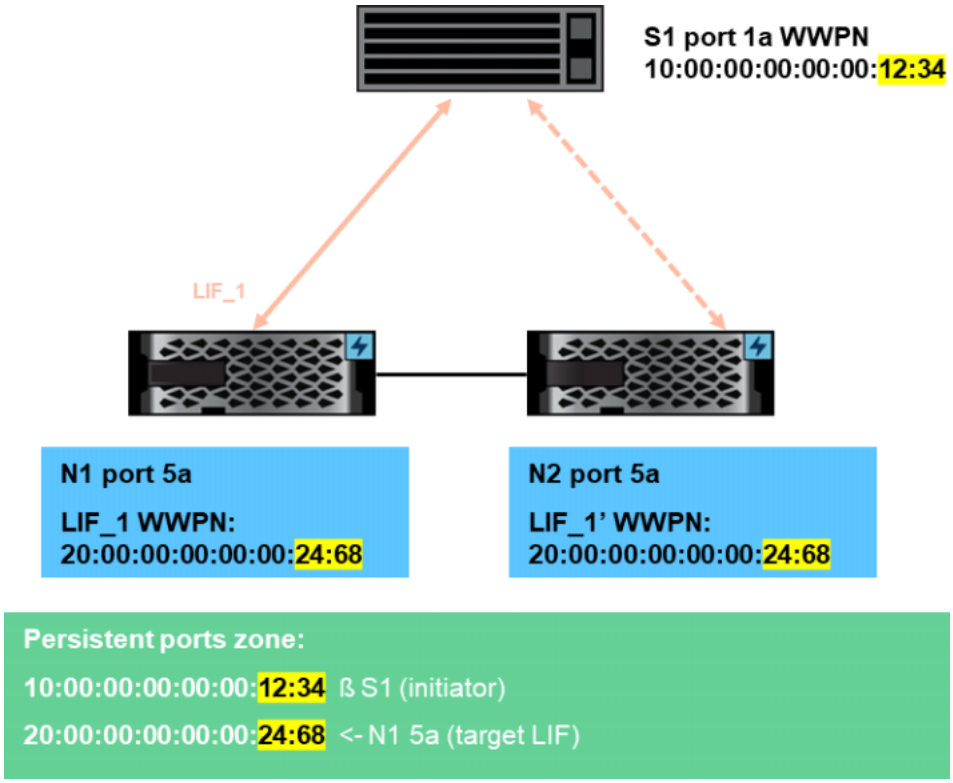


Figure 5 shows a persistent ports zoning example.

Figure 5 Persistent ports zoning example



7. ONTAP 9.9.1 New Features

ONTAP 9.9.1 adds a number of enhancements to ONTAP SAN, which include:

- System Manager
- SnapMirror Business Continuity (SMBC)
- ASA maximum cluster size grows from a single HA pair to 12 nodes
- Single LUN performance improvement
- Nested igroups
- NVMe/FC vVols support
- NVMe remote I/O support
- ASA adds NVMe/FC support

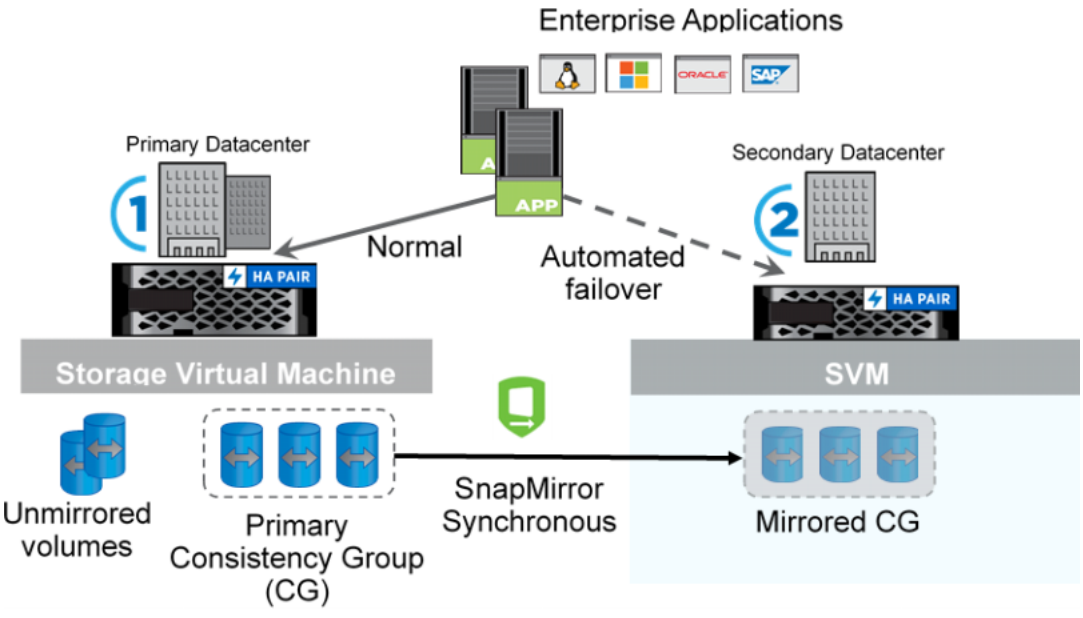
System Manager

System Manager can display the Rest API calls for every action performed. This functionality assists administrators who want to script workflows or check syntax. Operations commonly performed are enhanced and simplified; for example, when you are asked to map a new LUN to an igroup, all the potential igroups are displayed. Or when you are asked to provide WWPNs in an igroup, all the existing WWPNs seen are displayed and can be selected. This should save time and reduce input mistakes.

SnapMirror Business Continuity (SMBC)

ONTAP 9.9.1 introduces SMBC, which uses SnapMirror Synchronous to synchronously replicate applications using application consistency groups to manage and replicate all application objects between the two sites. SMBC enables automated failovers between two synchronously replicated sites. This reduces any outage durations and significantly lowers administrative costs associated with maintaining both mirrors and managing automated failovers. Figure 6 shows the SMBC topology.

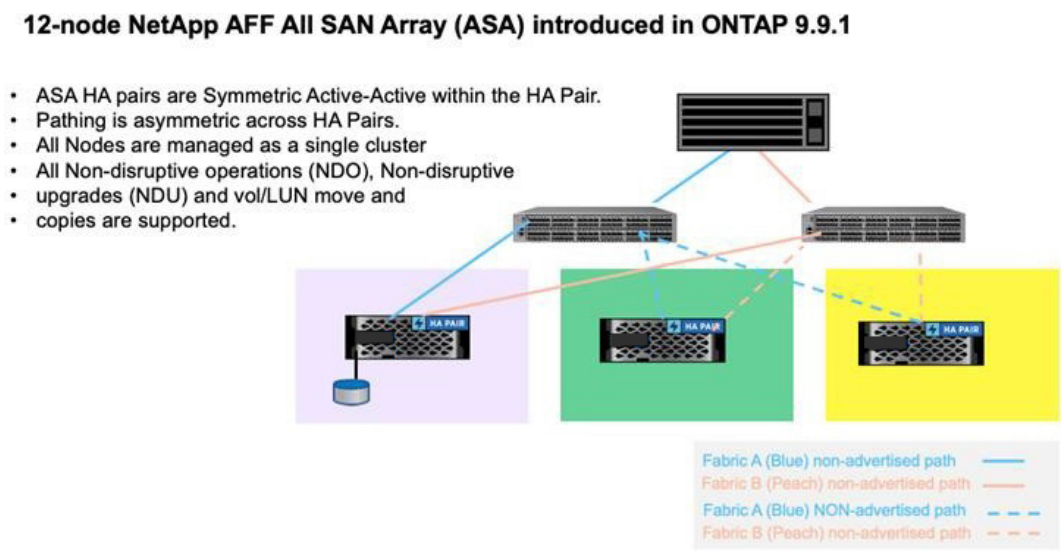
Figure 6 SMBC topology



ASA Maximum Cluster Size Grows from a Single HA Pair to 12 Nodes

In configurations larger than a single HA pair, it is important to understand that the ASA cluster is essentially a federation of ASA HA pairs that are clustered together and share a single management plan. This layout allows all of the usual nondisruptive operations (NDO) and other features of ONTAP clusters with the exceptions that there are no NAS protocols or features supported on those nodes and certain other ASA specifics. Because the primary feature that defines an ASA is that it offers symmetric active-active access, it is important to understand how access and pathing works when ASAs are combined into a larger cluster. ASA within a single HA pair advertises all paths through both nodes as AO or preferred paths, and therefore the host MPIO stack will use all AO paths. For ASA clusters that are larger then a single HA pair, each HA pair advertises all paths to LUNs hosted by that pair as AO. However, any paths through other controllers not part of the hosting HA pair would advertise paths as active nonoptimized (ANO or non-preferred paths). These paths are not used by host MPIO stacks. Additionally, by default, ONTAP has a feature called Selective LUN Map (SLM) that only advertises paths through the hosting HA pair, therefore hosts would not be aware of other nonoptimized or less preferred paths unless SLM was configured to advertise additional paths.
For more information about SLM, see the "Selective LUN Mapping" (page 36) section. For a full list of those specifics, see the ETERNUS AX/HX series Best Practices for Business-Critical Workloads in the Fujitsu manual site.

Figure 7 12-node ASA



Single LUN Performance Improvement

ONTAP 9.9.1 introduces significant single LUN performance improvements. These improvements are primarily made possible by parallelizing more of LUN I/O operations so that more processing can be done concurrently. The amount of the performance improvement tends to skew bigger on larger memory/CPU controllers. We anticipate that the bulk of the benefits to single LUN performance is likely to be most valuable in the following area:

- Virtualization where a datastore is being backed by an LUN.

In many cases, single LUN performance improvements won't have an effect on many customers because most LUNs are stripped together using logical volume managers or other application aggregations. Single LUN performance has no effect on the overall performance when many LUNs are used, which is the common usage pattern.

Nested igroups

ONTAP 9.9.1 adds nested igroups to simplify LUN masking. Nested igroups allow an existing igroup to be added to a new igroup. This capability grants greater simplicity in defining what initiators have access to which LUNs. It enables users to be more creative in terms of igroup naming and aliases that make sense to storage administrators. Some other enhancements to igroups include:

- igroups and initiators might have a comment assigned
- LUN igroup initiators show (new CLI-only command to view initiators)

The following is an example of adding a comment and showing initiators with the new comment:

```
tme-a700s-clus::> lun igroup initiator modify -initiator 10:00:00:10:9b:34:9f:34 -comment "This
is a comment about 10:00:00:10:9b:34:9f:34"

tme-a700s-clus::> lun igroup initiator show                                    Vserver
Initiator                                 Comment
---------- ---------------------------------------- ----------------------
svm0      10:00:00:10:9b:34:9f:34              This is a comment about 10:00:00:10:9b:34:9f:34
svm0      10:00:00:10:9b:34:9f:35              -
svm0      10:00:00:90:fa:d1:ea:f7              -
svm0      10:00:00:90:fa:d1:ea:f8              -
4 entries were displayed.
```

Igroups can now contain up to three levels of nesting:

- Grandparent igroup
- Parent igroup
- Child igroup

Some additional caveats for nested igroups are:

- The operating system and protocol should match and they might not be modified.
- `allow_delete_while_mapped = True` might unnest and delete the igroup if it is mapped to lun: `igroup_nested_delete, igroup_delete.`
- You cannot unnest or delete a child if one of its parents is mapped.
- If the child is mapped, you can unnest or delete the parent.
- Deleting (igroup delete) an igroup in a nested relationship is allowed and the nondeleted igroups persist (if unmapped).

NVMe/FC vVols Support

ONTAP 9.9.1 added NVMe/FC vVol support. This enhancement allows management and automation of the storage via VMware administration using vVols after vSphere supports that option within vVols/vCenter.

NVMe Remote I/O Support

With ONTAP 9.9.1, NVMe-oF adds remote I/O support. This changes NVMe-oF pathing from an active-inactive model to the AO/ANO model that all other ONTAP block protocols use.
Figure 8 illustrates the NVMe-oF without remote I/O support.

Figure 8 NVMe-oF without remote I/O support

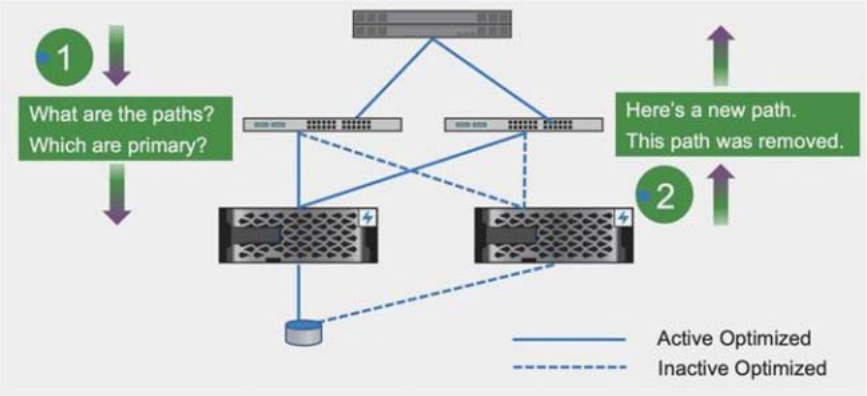
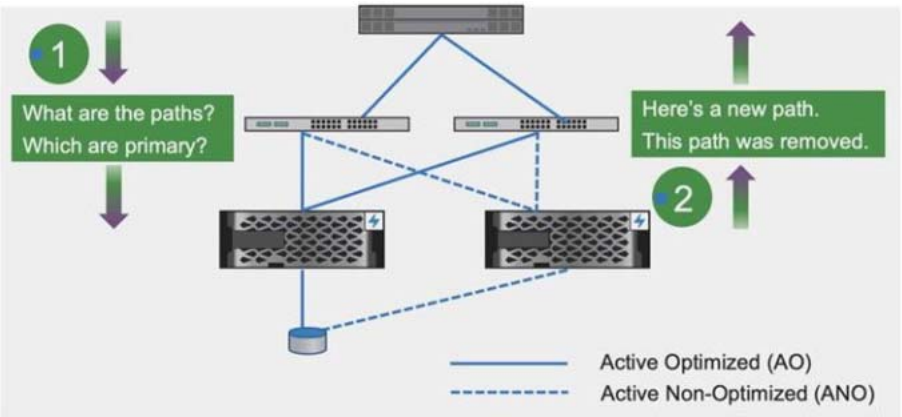


Figure 9 NVMe-oF with remote I/O support



Comparing Figure 8 and Figure 9, it doesn't seem like much of a difference because the difference is subtle. With remote I/O support in NVMe-oF, all paths are active, which means that I/O sent down any of those paths is acknowledged and responded to or answered. Previously, without remote I/O, inactive paths were unavailable and could not be used.

ASA Adds NVMe/FC Support

ONTAP 9.9.1 adds NVMe/FC as an additional block protocol. Unlike FC or iSCSI, NVMe/FC on ASA will continue to be asymmetric (AO/ANO), which is due to the differences in how NVMe-oF works with remote versus local paths.

8. ONTAP and SAN Protocols

ONTAP Overview

Storage controllers running an ONTAP are referred to as nodes. These nodes are aggregated into a clustered system. The nodes in the cluster communicate with each other continuously, coordinate cluster activities, and move data transparently from node to node by using redundant paths to a dedicated cluster network.

Although the basic unit of a cluster is the node, nodes are added to the cluster as part of a HA pair. HA pairs enable high availability by communicating with each other over an HA interconnect (separate from the dedicated cluster network) and by maintaining redundant connections to the HA pair's disks. Disks are not shared between HA pairs, although shelves can contain disks that belong to either member of an HA pair.

Clusters are administered on the whole cluster rather than on a per-node basis, and data is served from one or more SVMs. Each SVM is configured to own storage in the form of volumes (and LUNs) provisioned from a physical aggregate, and LIFs are assigned either to a physical Ethernet network or to FC target ports. LUNs are created inside an SVM's volumes and mapped to hosts to provide them with storage space. SVMs are node independent and cluster-based; they can make use of physical resources such as volumes or network ports anywhere in the cluster.

Considerations for Optimizing SAN Performance

ONTAP is optimized to use multiple processor cores in order to concurrently process as much work as possible. Many ONTAP operations can be distributed across available multiple processor cores. However, there are some operations that can't be split across multiple processor cores that can reduce the maximum performance that can be achieved. These nondistributable threads can limit the maximum performance that could be achieved if workloads are properly split across multiple objects. By using more objects, you can bring all available cores to bear to optimize performance by increasing the number of volumes and LUNs provisioned for a given workload.

Volumes

When considering how many volumes should be provisioned for a given application, it is important to first consider what a volume is used for in a SAN context. Depending on the storage vendor being considered, the word volume can be used to mean many things. When talking about volumes in an ONTAP context, it is important to understand what volumes are used for in the context of ONTAP SAN. In ONTAP, volumes provide the following functionality:

- Management container for any LUNs they host. This can be beneficial in cases where volumes hold more than a single LUN. This simplifies the management of multiple LUNs that are hosted within the volumes.
- ONTAP volumes can be used as a consistency group because a Snapshot copy is taken at the volume level and therefore captures all blocks in the volume. This means that multiple LUNs hosted by the same volume all create Snapshot copies at the same time. This greatly reduces maintain workload consistency across a group of LUNs.
- Storage efficiency features are primarily performed at the volume level. This means that the larger the dataset is the common blocks are likely to be found. In other words, larger volumes generally increase storage efficiencies.

While we have just covered how fewer larger volumes eases administration, can be used as a consistency group, and increases storage efficiencies, we also need to consider how many volumes should be provisioned to optimize performance.

As previously mentioned, some processing threads can't be distributed across multiple processor cores. Not being able to spread this work across multiple cores can be managed by spreading the workload across multiple volumes. By doing this, you can break up the amount of non-distributable work that has to be done by a single core.

Each volume has its own non-distributable threads, so the more volumes that are spread work across allows us to concurrently work many of these non-distributable threads across multiple processor cores, thus improving throughput.

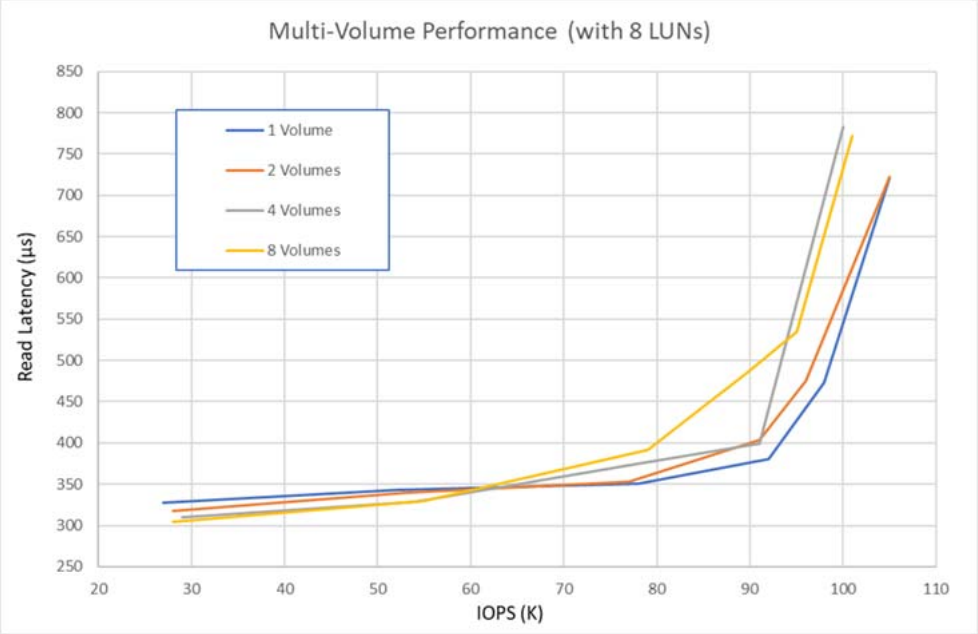
This means that there is a balance that must be drawn between grouping many LUNs in the same volume versus increasing the number of volumes. Generally, from considerable testing and years of customer experience, Fujitsu has been able to provide the guideline that, in most cases, you can optimize performance by increasing the number of volumes up to a point. After that point, it was discovered that more volumes don't increase performance, but they do increase complexity and reduce storage efficiency.

Fujitsu recommends using between 4–16 volumes, with the ideal being between 8–16 volumes.

Note

Figure 10 and Figure 11 show the effects of adding more volumes and LUNs; they are displayed for illustrative purposes only. They are not meant to guarantee or benchmark specific workloads or provide estimates of performance.

Figure 10 The effect of spreading work across more volumes



Best Practices
• In most cases, Fujitsu recommends using between 8–16 volumes to maximize performance. This recommendation assumes that these are the only volumes on a given controller. If other volumes are present, then you can generally consider using up to eight additional LUNs to maximize performance. • LUNs that are related to each other and have similar performance and management requirements can be hosted by a single volume. By using the same volume, organizations can realize the following benefits: - Reduced administrative complexity through using a common administrative container. - Snapshot copies and data protection, or replication offerings that build on Snapshot copies, are managed at the volume level. If all LUNs that a given application (or a given host) are hosted by a common volume, then the volume can effectively serve as a consistency group. A Snapshot copy context is the volume, and all items hosted by that volume are captured by all Snapshot copies. - Storage efficiency uses the volume as the organizing container; all storage efficiency objects and metadata are stored at the volume level. Therefore, the more LUNs a volume contains, the more efficiencies (such as blocks deduplicated, compression, and compaction) can be found. • There are cases where it might not make sense to combine multiple LUNs in the same volume; however, combining LUNs that are related, have similar performance requirements, and can benefit from being in a common consistency group, should be grouped together.

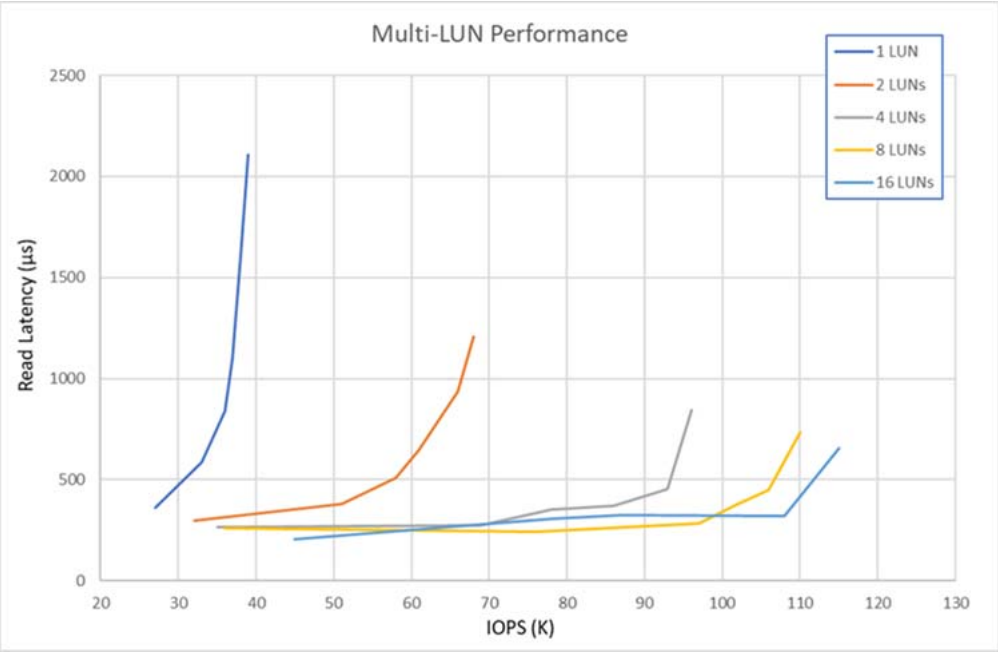
LUNs

LUNs (such as iSCSI or FC) have some threads that can't be distributed across multiple processor cores. As a result, this means that Fujitsu recommends using more smaller LUNs as opposed to fewer larger LUNs if performance is the primary consideration. Like volumes, workload performance can be optimized by spreading work across multiple LUNs, which allows more processor cores to be used to concurrently process workload I/O.
Some approaches that can be used to increase the number of LUNs an application uses are:

- Logical volume managers (LVM) combine multiple LUNs into a single volume that is presented to the host's OS or application. LVMs are commonly used with Linux and UNIX OSs. Oracle ASM can also aggregate multiple LUNs as a single storage object.
- In some cases, it might also make sense to advertise multiple LUNs to an OS or application if the OS or application essentially manages the LUNs presented.

Best Practices
Fujitsu recommends using more smaller LUNs versus fewer bigger ones. Ideally between 8-16 LUNs.

Figure 11 Effects of spreading work across multiple LUNs



FC In-Order Delivery

FC switches should be configured to ensure in-order delivery (IOD). While this step isn't necessary for ONTAP operations, it is a best practice because ONTAP will drop the exchange when an out of order or dropped frame is encountered. As a result, ONTAP must send a notification to the initiator (host) to retransmit the frame when the initiator hits its SCSI timeout threshold. This process might take 60 seconds. ONTAP will survive and recover from this situation but at a cost of the latency caused by the SCSI timeout and retransmit times.

If IOD is configured on all FC switches in the fabric, ONTAP won't receive any out of order frames and therefore, won't endure long host SCSI timeouts while awaiting frame retransmits.

Conclusions

There is essentially little performance benefit from additional volumes. Any variation is just a margin of error. Figure 10 and Figure 11 show a single volume with eight LUNs is delivering an easy 100K random IOPS at good latency. The inference that can be drawn from this is that you can increase application performance by increasing the number of LUNs used with a given application. While there are some performance improvements by increasing the number of volumes, those performance improvements are much smaller than those seen from increasing the number of LUNs. In both cases, returns from increasing the number of volumes tend to be both smaller. Furthermore, increasing the number of volumes can lead to diminishing returns to scale.

Note

On rare occasions, you can spread your workload over multiple volumes to improve the performance, but this mostly applies to cases where a single application is consuming all the capabilities of a controller. For example, if you have a one large database that needs to push 500K IOPS, and you want to minimize every microsecond of latency, then you need more than one volume. If you are in this situation, you should consider all aspects of the configuration, not just the number of volumes in use.

Based on the data presented above, we can draw the following inferences:

- Spreading your work across multiple LUNs improves performance significantly.
- A single LUN can achieve approximately 35K IOPS. With two LUNs, the performance doubles to approximately 70K IOPS.
- Benefits start diminishing as you reach eight LUNs. Using four LUNs is acceptable, but eight is slightly better.
- A single volume with eight LUNs delivers an easy 100K random IOPS at good latency, which is more I/O than is required by 99% of the common databases.

Note

Figure 10 and Figure 11 show that the numbers in the conclusions previously discussed were from a specific test. The conclusions are valid, and they illustrate the Fujitsu best practice recommendations. However, the specific numbers listed are included to illustrate the concepts and recommendations. They shouldn't be taken as guarantees or guidelines for what a given volume, LUN, or application can achieve regarding performance.

Volume Configuration

When provisioning volumes in a cluster, many considerations regarding deduplication, space reservations, and storage efficiency are the same. One major difference is that volumes on ONTAP storage clusters are oriented to SVM instead of to individual nodes, and a side effect is that they can be mapped into an SVM-wide global namespace for the purpose of exporting file systems by using NFS or CIFS protocols. However, the presence or absence of a given volume in the global namespace has no effect on data that is served by using connections such as FC or iSCSI.

Best Practices
Volumes that contain LUNs do not need to be junctioned to the global namespace to serve data by using block protocols; they only require an igroup-to-LUN mapping.

Host Connectivity

Hosts that access data served by an ONTAP storage cluster using a block protocol are expected to make use of the ALUA extension to the SCSI protocol to determine which paths are direct and which are indirect to any particular LUN. The ALUA standard refers to direct paths as active/optimized and to indirect paths as active/nonoptimized. All ALUA information is requested and delivered in band using the same connection such as iSCSI or FC that is used for data.

The status of a given path is discoverable by a host that sends a path status inquiry down each of the paths it has discovered for a given LUN. This path status inquiry can be triggered when the storage system sends extra data along with the result of a SCSI request to inform a host that paths' statuses have been updated and that their priorities should be rediscovered.

ALUA is a well-known and widely deployed standard and is a requirement for access to block data served by ONTAP. Any operating systems tested and qualified to work with ONTAP block access protocols support ALUA.

Path Selection

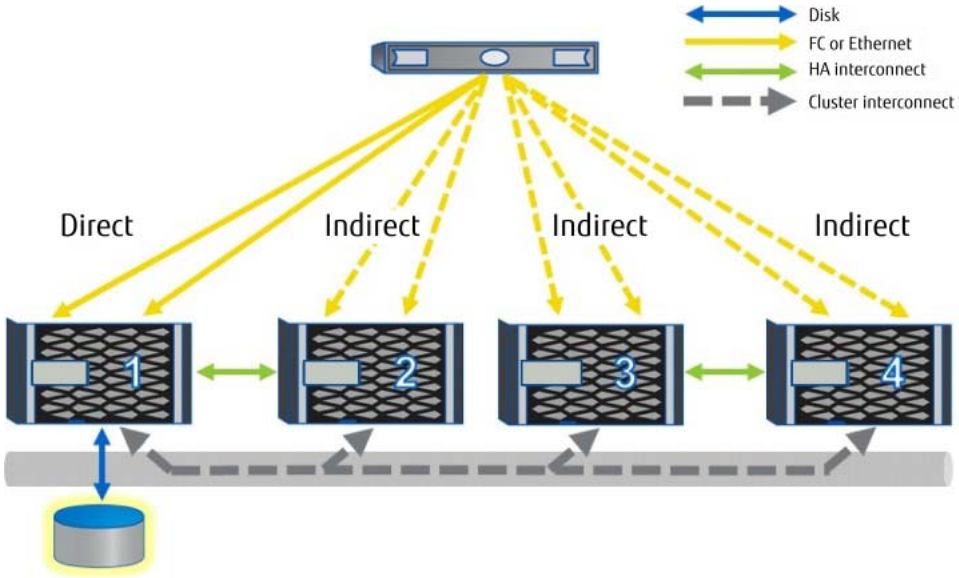
Even though every LIF owned by an SVM accepts writes and read requests for its LUNs, only one of the cluster nodes actually owns the disks backing that LUN at any given moment. This effectively divides available paths to a LUN into two types: direct and indirect paths.

A direct path for a LUN is a path where an SVM's LIFs and the LUN being accessed reside on the same node. To go from a physical target port to disk, it is not necessary to traverse the cluster network.

Figure 12 shows a host accessing data on a LUN owned by the node labeled 1 on direct paths. Any paths to this node are direct paths, because the LUN is on its storage aggregates. It is common in any SAN protocol setup to have multiple direct paths to a LUN. For purposes of redundancy and data access resiliency, a second path is commonly over separate Ethernet networks or FC fabrics, with additional paths per network or fabric possible for throughput purposes.

The use of ALUA allows hosts to direct traffic over any available direct paths before relying on indirect paths, and so any use of indirect paths in a non-failure scenario is rare.

Figure 12 Overview of paths in ONTAP



Indirect paths are data paths where an SVM's LIFs and the LUN being accessed reside on different nodes. Data must traverse the cluster network in order to go from a physical target port to disk. Because the cluster network is fast and highly available, this does not add a great deal of latency to the round trip, but it is not the maximally efficient data path. In a well-configured SAN environment, a host's use of indirect paths is minimal.

Because every host communicates only with SVMs that use physical resources anywhere in the cluster, in practice this means that all connections to a cluster are managed by MPIO software running on the host that is accessing LUNs, with the result that only direct paths are used during normal operation.

Best Practices
All SVMs should be assigned LIFs on each cluster node and each FC fabric or Ethernet network. For instance, if a four-node cluster is connected to two independent FC fabrics, A and B, using its 3a and 4a FC target ports, an SVM that serves data by using FC should have eight LIFs, on node1:3a, node1:4a, node2:3a, node2:4a, node3:3a, node3:4a, node4:3a, and node4:4a. Clusters with more than four nodes should limit the number of paths used to access any given LUN for ease of manageability and in deference to operating system path count limitations. For a more in-depth discussion, see "Path Management and Selective LUN Mapping" (page 35).

For administrators who are used to using an ONTAP storage cluster with NAS protocols such as NFS and CIFS, there is a distinction to be made between LIFs that serve these protocols and LIFs that serve block iSCSI or FC. NAS LIFs

can be freely moved from node to node, or they can belong to a failover group that specifies to which node and port they move during an HA or port failover. SAN LIFs, by comparison, represent the endpoint of a number of paths, all established simultaneously between SCSI initiator and SCSI target, and the host's MPIO software manages which paths actually receive I/O. As a result, unlike NAS LIFs, SAN LIFs do not fail over. The failover mechanism for SAN is provisioning multiple paths and using multipathing (MPIO) software on hosts to manage the multiple paths presented to them.

Because of this difference in behavior, Ethernet LIFs that serve data by using the iSCSI protocol cannot also serve data by using a NAS protocol.

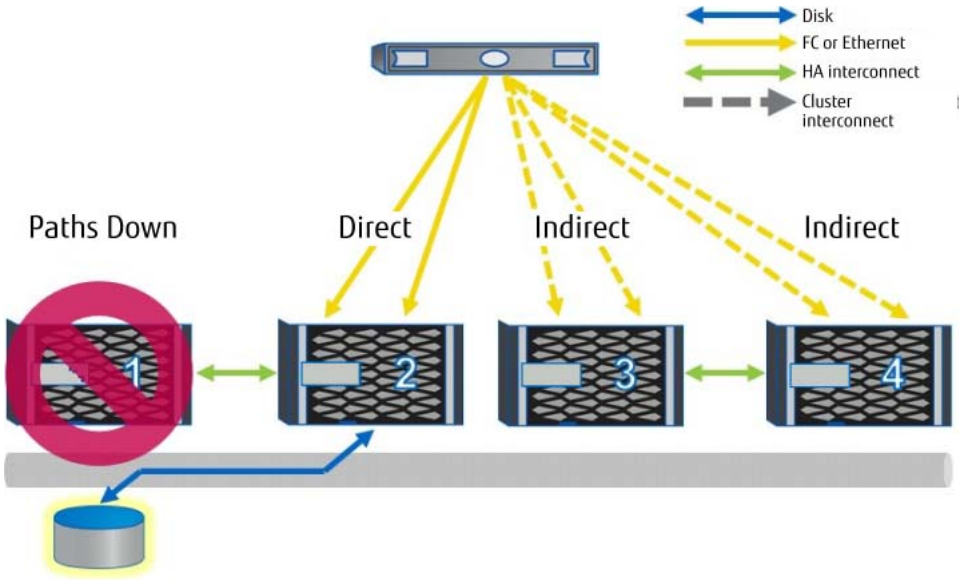
Path Selection Changes

There are three events that could change the path selected by a host to access data on a cluster:

- HA Failover

In an HA failover event, LIFs on the down node are seen as offline, and LIFs on the HA partner that has taken over for the down node are now direct paths. This state changes automatically due to ALUA path inquiry, and no administrative changes are necessary. Figure 13 shows the paths during HA failover.

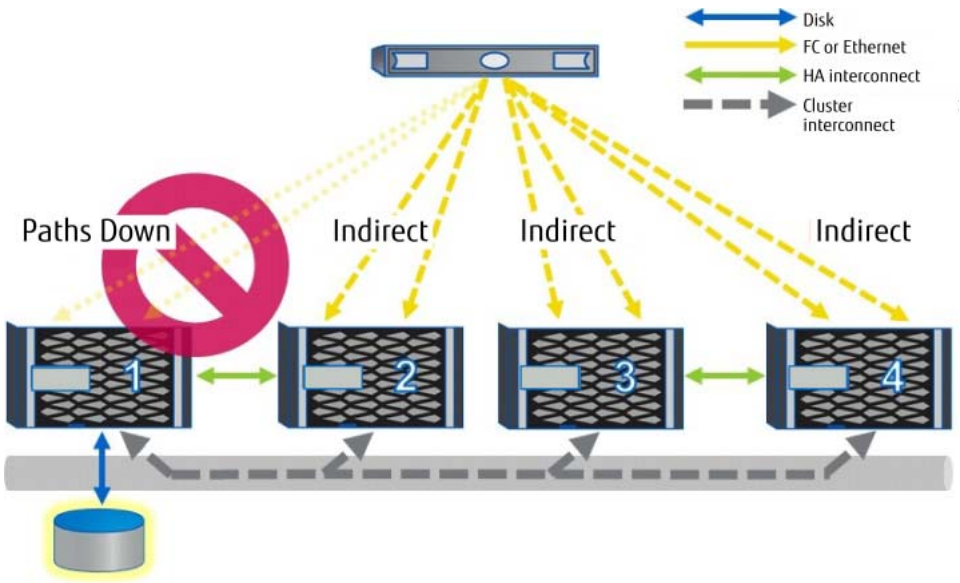
Figure 13 Paths during HA failover



Port or Switch Failure

In a port or switch failure, no more direct paths are available. Path priority remains the same, and MPIO software running on the host selects alternate indirect paths until a direct path becomes available again. The ALUA path states do not change. Figure 14 shows the paths during port or switch failure.

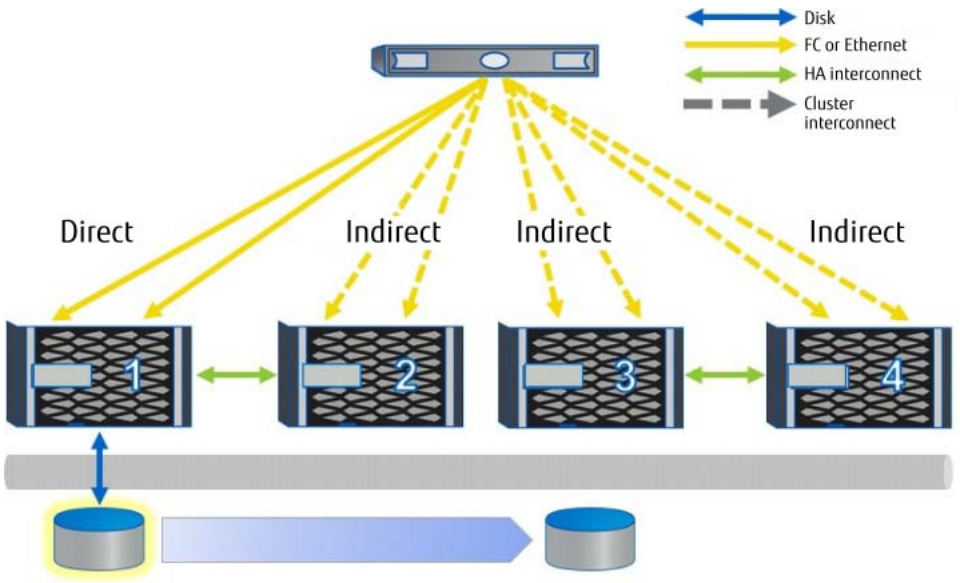
Figure 14 Paths during port or switch failure



Volume or LUN Mobility

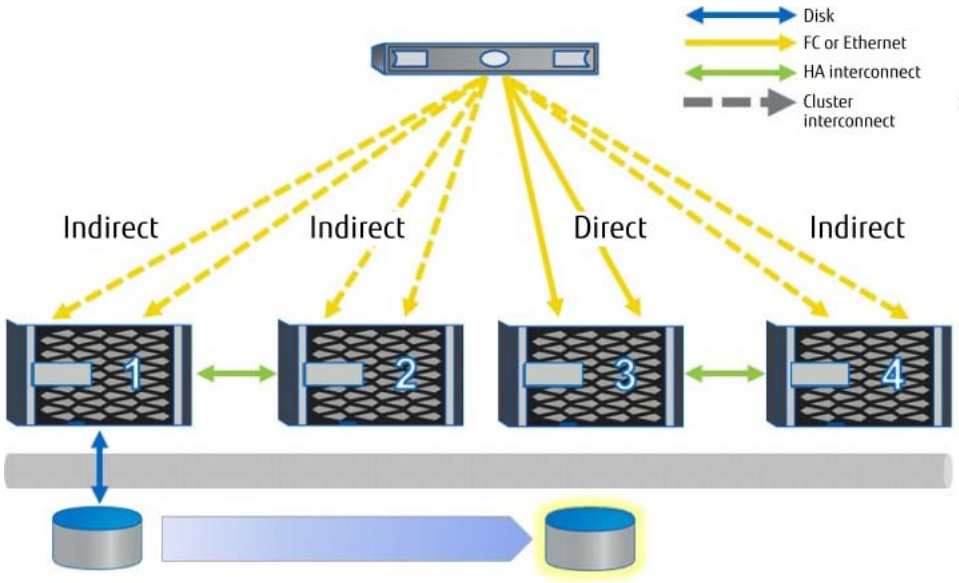
A volume is moved transparently between nodes by using `volume move` functionality, or a LUN is moved transparently between nodes using `lun move`. Figure 15 shows the paths during volume or LUN mobility.

Figure 15 Paths during volume or LUN mobility



For a volume move, when the cutover occurs and the volume's new node begins to handle read and write requests, the path status is updated so that the new node has direct paths and the old node has indirect paths. All paths remain available at all times. Figure 16 shows the paths after volume or LUN mobility.

Figure 16 Paths after volume or LUN mobility



For a LUN move, the cutover occurs before all of the data has been transferred to the destination, and read requests are passed through the cluster network to the source node to be fulfilled. For more details about the behavior of LUN move functionality, see "DataMotion for LUNs" (page 39).

FC and NPIV

An ONTAP node uses NPIV to permit every logical interface to log in to an FC fabric with its own WWPN, rather than using a single worldwide node name (WWNN) and associated WWPNs based on the address of the HA pair's physical FC target adapters. This permits a host connected to the same FC fabric switch to communicate with the same SCSI target regardless of which physical node owns which LIF. The virtual port presents the SCSI target service and sends and receives data.

Best Practices
NPIV-compatible FC switches are required for FC LIFs to operate correctly. Before creating FC LIFs, make sure that any fabric switches attached to an ONTAP system have NPIV enabled.

When using Cisco NX-OS, the status of NPIV can be verified by running the `show npiv status` command.

```
N5K-A# show npiv status
NPIV is enabled
```

When using Brocade FabOS, the `portcfgshow` command shows NPIV capability and status.

```
BRCD_8K:admin> portcfgshow
Ports of Slot 0    0   1   2   3     4   5   6   7     8   9  10  11    12  13  14  15
-----------------+---+---+---+---+-------+---+---+---+-------+---+---+---+-------+---+---+---
Speed             AN  AN  AN  AN    AN  AN  AN  AN    10  10  10  10    10 10 10  10
Fill Word          0   0   0   0     0   0   0   0     -   -   -   -     -   -   -   -
AL_PA Offset 13   ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Trunk Port        ON  ON  ON  ON    ON  ON  ON  ON     -   -   -   -     -   -   -   -
Long Distance     ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
VC Link Init      ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Locked L_Port     ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Locked G_Port     ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Disabled E_Port   ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Locked E_Port     ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
ISL R_RDY Mode    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
RSCN Suppressed   ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Persistent Disable..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
LOS TOV enable    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
NPIV capability   ON  ON  ON  ON    ON  ON  ON  ON    ON  ON  ON  ON    ON  ON  ON  ON
NPIV PP Limit    126 126 126 126   126 126 126 126   126 126 126 126   126 126 126 126
QOS E_Port        AE  AE  AE  AE    AE  AE  AE  AE    ..  ..  ..  ..    ..  ..  ..  ..
EX Port           ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Mirror Port       ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Rate Limit        ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Fport Buffers     ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
Port Auto Disable ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
CSCTL mode        ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
```

From the storage administration console, it is not possible to inquire about NPIV status on an attached switch directly, but examining the local FC topology can show whether fabric switch ports have NPIV enabled. In the following example, NPIV must be enabled, because port 2/1 has multiple attached WWPNs, some of which are virtual ports.

```
cluster::> node run -node node01 fcp topology show
   Switch Name: N5K-A
 Switch Vendor: Cisco Systems, Inc.
Switch Release: 5.0(2)N2(1a)
 Switch Domain: 200
    Switch WWN: 20:66:00:0d:ec:b4:94:01

Port    Port WWPN                State   Type     Attached WWPN           Port ID
-------------------------------------------------------------------------------
2/1     20:01:00:0d:ec:b4:94:3f  Online  F-Port   50:0a:09:83:8d:4d:bf:f1 0xc80033
                                                  20:1c:00:a0:98:16:54:8c 0xc80052*
                                                  20:0e:00:a0:98:16:54:8c 0xc80034*
                                                  20:10:00:a0:98:16:54:8c 0xc8003f
2/3     20:03:00:0d:ec:b4:94:3f  Online  F-Port   50:0a:09:83:8d:3d:c0:1c 0xc8002c
```

Best Practices
Physical WWPNs (beginning with 50:0a:09:8x) do not present a SCSI target service and should not be included in any zone configurations on the FC fabric, though they show as logged in to the fabric. These WWPNs are listed by using the `fcp adapter show -fields fc-wwpn` command or using the FC/FCoE Adapters pane under NETWORK > FC Ports in System Manager, as shown in Figure 17. Instead, use only virtual WWPNs (WWPNs starting with 20:) visible in the output of the `network interface show` command and in the System Manager NETWORK > Ethernet Ports pane, as shown in Figure 18.

Figure 17 FC adapters in System Manager

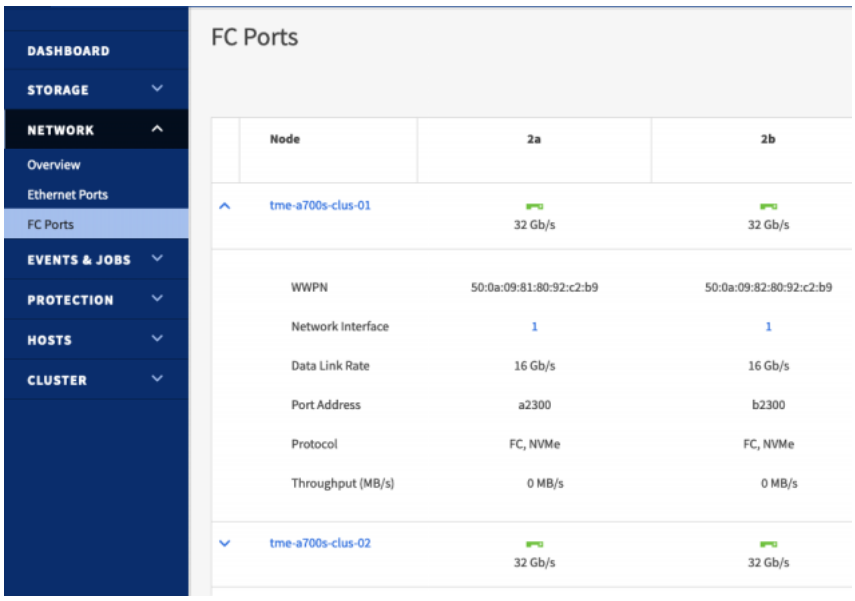
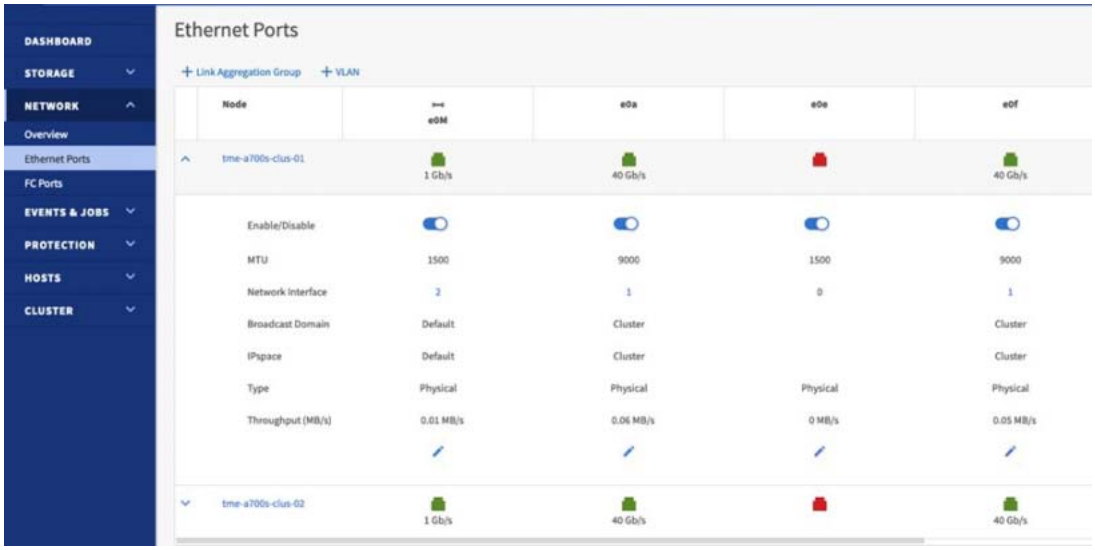


Figure 18 Network interfaces in System Manager



Path Management and Selective LUN Mapping

Clusters with more than two nodes are likely to have more paths than has commonly been the case in the past. Clusters attached to more than one fabric, or with storage controllers attached more than once per fabric, can quickly multiply the number of potential paths available.
This presents the following potential problems to the storage administrator:

- Having a large number of target ports can be good for redundancy, but it can become operationally burdensome. In an FC environment, it requires larger, more complex zones and zonesets; a larger table of WWPNs belonging to cluster SVMs of which to keep track; or, in the case of an iSCSI environment, a large number of sessions to be established for every host that requires a LUN.

- Many operating systems have an upper limit to the number of paths it is feasible for them to access. Especially for hosts that have many paths and many LUNs, this can lead to LUN enumeration or path status problems.
- Some demanding, high-throughput workloads can benefit from having their traffic segregated from less critical traffic to reduce contention, but ALUA path statuses provide no mechanism to prioritize one direct path over another.
- The ONTAP storage OS has an upper tested limit to the total number of established paths (known as an initiator-target nexus, or ITN). For further details about the limit for any Fujitsu storage controller, see the ETERNUS AX/HX Series SAN Configuration Guide in the Fujitsu manual site.

You should consider limiting the total number of paths presented. However, to make sure of both a direct path to data and availability/redundancy in the case of an HA failover or path failure, at a minimum, both the node that contains the volume with the data being accessed and its HA partner must present paths.

There are two methods for limiting paths presented by a LUN by using storage OS capabilities, as opposed to limiting paths only using FC zoning or iSCSI session management: selective LUN mapping, which is enabled by default, and port sets.

Selective LUN Mapping

SLM is an addition to the LUN mapping table that already exists in an ONTAP cluster, which consists of every logical linking of LUN path, igroup, and LUN ID. This table is necessary to get a full description of every LUN mapping, because LUNs might be mapped to multiple igroups (especially in host-based clustering scenarios), and because igroups might have multiple LUNs mapped.

In storage clusters running ONTAP, in addition to these properties, every mapping also contains a list of reporting nodes that show that LUN as present from the storage controllers listed to the igroup specified in the same mapping, as shown here:

```
san-cluster::> lun mapping show -fields reporting-nodes
vserver          path                igroup  lun-id   reporting-nodes
---------------  ------------------  ------- -------  -----------------
SAN_Default_SVM  /vol/host1/lun0     linux1        0  node-01,node-02
SAN_Default_SVM  /vol/host2/lun0     linux2        0  node-01,node-02
SAN_Default_SVM  /vol/host2/lun1     linux2        1  node-03,node-04
```

Note

- By default, any created LUN mappings have the default selective LUN mapping policy applied: presenting the LUN from the node that contains the volume in which the LUN resides and its HA partner.
- LUN mappings created in an earlier version of the ONTAP storage OS have a wildcard entry in the list of reporting nodes to reflect the earlier default behavior of presenting the LUN from all nodes in a storage cluster.

However, a LUN mapping can also contain any or all other nodes in the cluster, as long as they are grouped in HA pairs, or it might be a blank or a wild card, in which case the LUN is reported as present by every node in the cluster. In this way, storage administrators can choose which storage controllers present paths in a highly granular fashion.

Port Sets

Port sets allow administrators to mask an interface group so that the LUNs that are mapped to it are available on a subset of the total number of available target ports. This functionality is available in clustered ONTAP.

Note

A LIF that is currently a member of a port set cannot be modified until it is removed from the port set. It can be added to the port set after modification, but care should be taken to leave enough LIFs in the port set to satisfy host requirements for a path to data.

To make sure of both a direct path to data and availability/redundancy in the case of an HA failover or non-disruptive operation event, the only paths required are to the node that contains the volume with the data being accessed and its HA partner.

Management Interfaces

Because LIFs belonging to SVMs that serve data by using block protocols cannot also be used for administration purposes and because the logical unit of management on an ONTAP storage cluster is the SVM, every SVM must have a management interface in addition to interfaces that are serving data using block protocols.

Best Practices

A management interface on an SVM serving block data should have the following properties:

- A LIF type of `data`
- No data protocols assigned (`-data-protocols none`)
- A firewall policy that permits management access (`-firewall-policy mgmt`)
- A failover group and policy that keep the LIF accessible to hosts that might need to contact it for data management purposes, such as creating or managing Snapshot copies (For more information about failover groups, see "Configuring failover groups and policies for LIFs" in the ETERNUS AX/HX Series Network Management Reference in the Fujitsu manual site.)

Additionally, an SVM-level administrative account should be available. The default account created during SVM creation is the vsadmin account, but it must first be assigned a password with the `security login password` command and then unlocked by using the `security login unlock` command.

When administering a cluster using System Manager, an SVM management LIF can be created during SVM creation, or it can be designated a management LIF during normal LIF creation. See Figure 19, Figure 20, and Figure 21.

Figure 19 Creating a management LIF during SVM creation

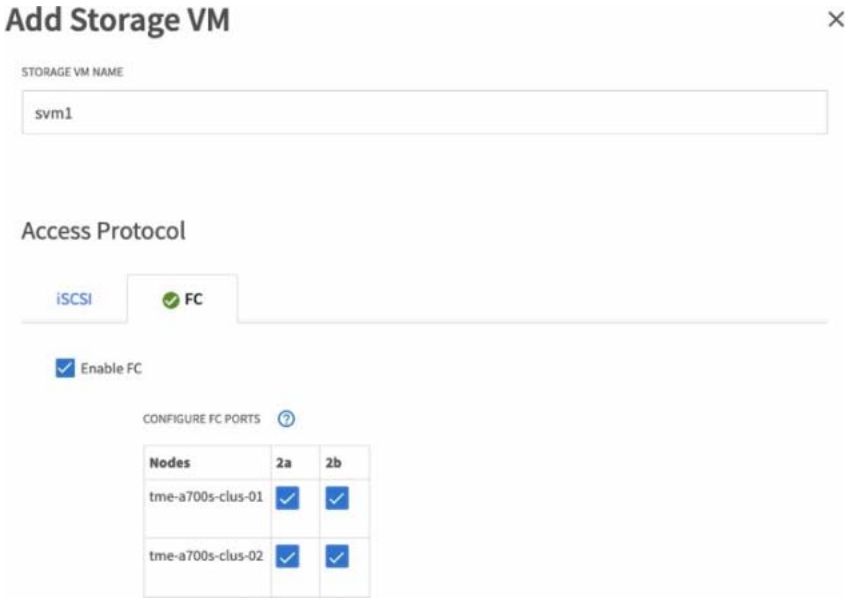
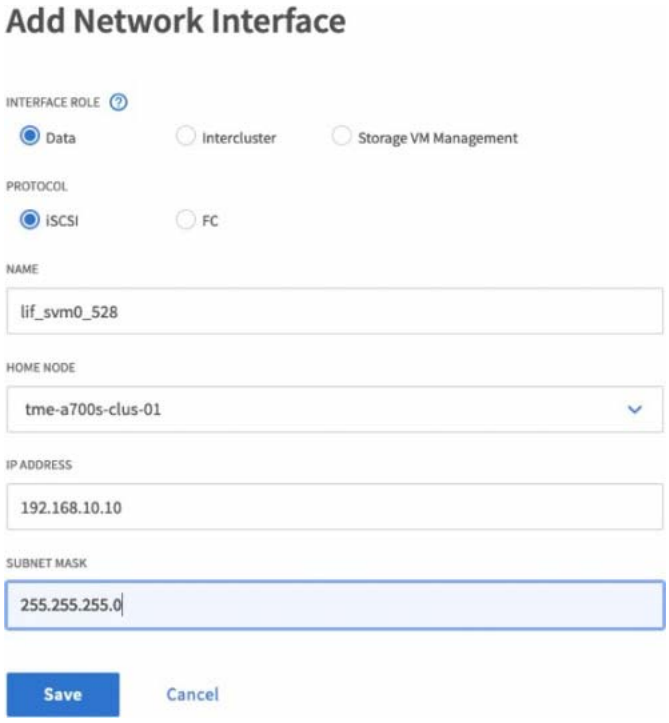
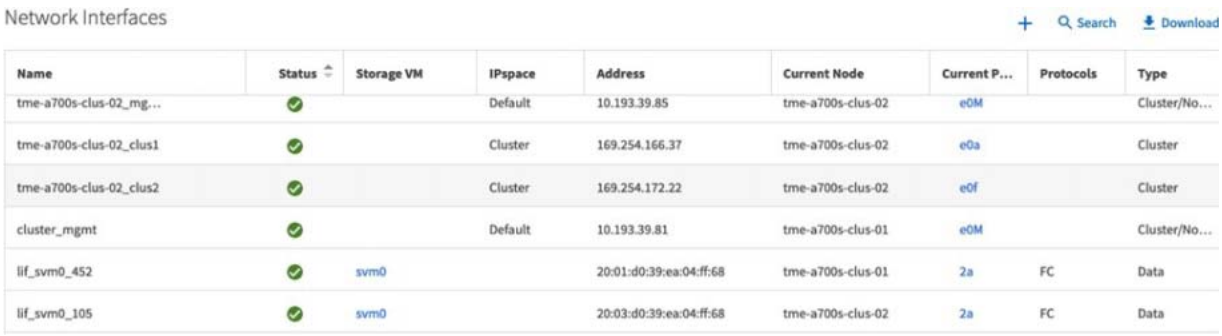


Figure 20 Creating a management LIF for an existing SVM



The management LIF details should look like the details in Figure 21.

Figure 21 Management LIF details



DataMotion for LUNs

LUNs can be moved and copied between volumes, aggregates, storage controllers, and HA pairs on a per-LUN rather than a per-volume basis, using the `lun move` and `lun copy` commands, or using an API call.

LUNs moved or copied using this method become available almost instantly. After the destination LUN is created and its metadata and attributes arrive, the LUN is "promoted" so that it can receive I/O requests from hosts. Meanwhile, data from the source is copied in the background across the cluster interconnect. Incoming read requests for data that has not yet arrived at the destination trigger the destination to reach back to the source before fulfilling the read requests. Incoming write requests are written directly into the destination.

LUN Move and LUN Copy Comparison

There are some differences between using DataMotion for LUNs to move a LUN and using it to copy a LUN.

LUNs can be copied between volumes in the same SVM or to volumes in other SVMs (when performed by a cluster administrator). LUN moves are only possible from volume to volume within the same SVM, because it can't be assumed that the destination SVM has the same configuration. It has an entirely separate FC WWNN or iSCSI target name. Because the destination of a LUN copy does not have one or more LUN mappings to go with it, this is not a problem for inter-SVM copies.

The source of a LUN move must be in the active file system; the source of a LUN copy can be inside a Snapshot copy. Snapshot copies are immutable and cannot have data moved out of them.

By default, a LUN copy is promoted early, whereas a LUN move is promoted late:

- Early promotion means that a LUN can receive I/O, but Snapshot copies can't be taken.
- Late promotion means that a LUN can receive I/O and that Snapshot copies can be taken.

Storage Efficiency Considerations

LUNs that have been moved or copied using DataMotion for LUNs do not arrive compressed or deduplicated at their destination.

Best Practices
If a destination volume has not previously contained deduplicated or compressed data, turning on deduplication or compression adds the arriving LUN's blocks to the list of blocks to be processed during the next storage efficiency run, and they do not need to be discovered through a block scan.

Data can only be shared using deduplication or cloning within a volume; any data in a copied LUN is a duplicate of data in its source volume, and any data belonging to a LUN that is locked in a Snapshot copy on the source volume remains on disk until that Snapshot copy expires or is deleted, even if the LUN has been moved.

Volumes with inline compression turned on do not compress LUNs arriving through a DataMotion for LUNs operation.

Data Protection Considerations

Data protection considerations apply primarily to LUNs that have been moved and not copied, because a copy implies that the source data still exists in its source volume.

LUNs that have been moved do not bring with them any associated Snapshot copies that might still exist in the source volume. If the LUN data in the source volume's Snapshot copies must also be moved, LUN copy can be used to copy LUNs from its Snapshot copies. After deduplication, they share any available duplicate blocks with the LUN that has been moved into that volume.

A further consideration for a LUN that has been moved is that it does not necessarily participate any longer in data protection relationships associated with its source volume. Therefore, a follow-up action such as creating a new SnapMirror relationship might be necessary. If the destination already participates in a such a relationship, it might be necessary to take actions caused by more space being consumed by the data replication destination.

Best Practices
When using LUN move in conjunction with software external to the ONTAP storage cluster to manage Snapshot copies containing LUNs, make sure that the software is aware of the capabilities of DataMotion for LUNs and can (for example) restore a LUN from Snapshot copies in volumes in which it might no longer exist. If this is not possible, LUN move might have an effect on data protection workflows.

Scalability and Throughput Considerations

LUN move or copy operations can be throttled on a per-operation basis using the `-max-throughput argument`. Throttles can be applied either when the operation is started or to an already-existing operation using the `lun copy modify` or `lun move modify` command.

The maximum number of move or copy operations that can operate concurrently is up to 50. Further operations are queued. This limit applies to the destination side of the move or copy operation.

Best Practices
A LUN copy or move operation can be paused and resumed at any time after data begins copying in the background. Pausing the move or copy only prevents data from being moved in the background, but does not prevent requests for data that hasn't yet arrived from being forwarded to the source LUN for fulfillment.

Data Management and Workflow Considerations

There are a few other interactions with other ONTAP features to take into account when using DataMotion for LUNs:

- LUNs used as the source for a LUN move or copy cannot be removed while the operation is under way.
- LUNs used as the source for a LUN move or copy cannot be replaced using SnapRestore while the operation is under way.

If a LUN used as the source for a LUN move or copy is in a volume that is also being moved using a volume move operation, the LUN move or copy pauses during the moving volume's cutover period.

Best Practices

Some existing workflows can take advantage of DataMotion for LUNs to shorten the number of required steps:

- Previously, to duplicate a volume containing LUNs, the entire volume needed to be cloned. Now any empty or already-occupied volume can be filled with LUN copies from another volume's Snapshot copies, even if that volume is in a separate SVM. Effectively, the subvolume LUN cloning capability previously available within a volume can now be extended to other volumes.
- Previously, to change the existing layout and ratio of LUNs and volumes, it was necessary to clone volumes and then remove unneeded LUNs or to use a host-side copy using volume management to fill a new LUN with an old LUN's data. Now, if storage efficiency can be better served by consolidating LUNs in fewer volumes, or if a single LUN in a volume containing others needs to relocate to satisfy performance or storage tiering needs, LUNs can be moved nondisruptively between volumes on the fly.

DataMotion and Selective LUN Mapping: Discovering and Discarding Paths

When altering the LUN mapping on the storage cluster to create new paths or remove existing ones, the hosts attached to that LUN must perform a SCSI bus rescan. Therefore, when moving LUNs between HA pairs, the procedure should be as follows:

Procedure

1. Change the LUN mapping to add the new reporting nodes using the `lun mapping add-reporting-nodes` command.
2. Perform a SCSI bus rescan on the hosts accessing the LUN, discovering the new paths.
3. Move the LUN nondisruptively; ALUA signals a path status change to the host, and the host begins moving I/O down the new direct paths.
4. Change the LUN mapping to remove the old reporting nodes using the `lun mapping remove-reporting-nodes` command.
5. Perform a SCSI bus rescan on the hosts accessing the LUN, discarding the old paths.

More than one LUN can have new paths discovered or old ones removed during a rescan.

Caution

Do NOT remove reporting nodes until the LUN move is complete and any host remediation steps, for example, SCSI bus rescans, are completed. If reporting nodes are removed prior to adding new reporting nodes, completing the LUN move, and all host remediation steps are completed, you could lose access to the LUN that was moved.

Path Management Best Practices

You should use ONTAP features to limit the number of available paths at the storage management level.

Best Practices
• For storage controllers that have a single target LIF on each connected FC fabric or Ethernet network, the default number of paths presented by a LUN mapping is two direct paths from the storage controller that contains the volume and LUN being accessed and two indirect paths from its HA partner, for a total of four paths. • Selective LUN mapping by default limits a LUN's paths to the storage controller that owns it and its HA partner, but extra nodes might be part of a mapping on either a temporary or permanent basis. • In clusters that have more than one target LIF per connected FC fabric or Ethernet network, you can use the extra paths to provide more bandwidth or queue depth on a per-LUN basis, or port sets can be used to channel traffic on a per-igroup basis to specific LIFs. • For LUNs that require more paths than a default LUN mapping provides, eight paths are almost always sufficient and is a path count supported by all host SAN implementations. For LUNs that require even more paths, the ETERNUS AX/HX Series SAN Configuration Guide in the Fujitsu manual site lists the tested maximum number of paths for each supported host OS. • LUN mobility events such as `vol move` or `lun move` that involve moving a LUN from one HA pair in the cluster to another should include a step to confirm that the LUN is being presented using the destination storage controllers before the mobility event is initiated. The `lun mapping add-reporting-nodes` command can be used to add the new paths. After the move is complete, use the `lun mapping remove-reporting-nodes` command to remove the original, no longer direct path. • Changing the paths presented for a LUN also means that a host SCSI bus rescan should be performed in order to discover new paths and discard stale ones. For best practices from a host perspective on path changes and for the procedure to be used when a LUN mapping must change to accommodate its moving to an HA pair that currently does not present paths, see "DataMotion and Selective LUN Mapping: Discovering and Discarding Paths" (page 41). • Because a change on the host accessing the LUN is necessary for a LUN mapping change, consider expanding the list of nodes in LUN mapping situations where administrative steps taken on the host are undesirable or when LUN mobility between HA pairs is frequent.

9. Scalable SAN Key Value Propositions and Features

This section highlights some of Fujitsu's principal design goals. These goals included providing a unified architecture at scale that enables nondisruptive operations for data mobility, performance optimization, capacity planning, and even system-level hardware replacement. Although this is not an exhaustive list of key features now available, it does show how scalable SAN features and ONTAP are set apart from the rest of the storage market.

SVM as Unified Target and Unit of Management

The storage cluster extends a single WWNN on an SVM basis to every member of a cluster, so that every node presents the same target and permits multiple targets to coexist on the same physical hardware.

The same concept also applies to storage management. Because all data is served from volumes associated with an SVM and from an iSCSI or FC target configured as part of an SVM, a cluster is administered on a per-SVM basis, rather than the time-consuming process of administering storage a single node at a time.

This focus on management at the SVM level means that it is possible to implement a secure multitenancy model of storage management.

Scalability at the Node and Cluster Levels

ONTAP offers scale at both the node level and cluster level and has increased the scalability at both since SAN protocols were introduced in ONTAP. For the latest full details about SAN configuration limits, see the ETERNUS AX/HX Series SAN Configuration Guide in the Fujitsu manual site. For a summary, see Table 2.

Table 2 Scalability in ONTAP

Version of ONTAP	9.7 and later
Nodes per cluster	12
LUNs per node	8,192
LUNs per cluster	98,304
iSCSI sessions/node	8,192
FC I_T_Ns/node	8,192

Intracluster LUN and LIF Migration

Previous versions of ONTAP allowed volumes to be moved nondisruptively from any node to any other node in the cluster. It's also possible to copy and move LUNs between volumes and storage controllers on a per-LUN rather than a per-volume basis. LUN copy can be used to shorten cloning operations by making LUNs instantly available.

During normal operations, there is no need for LIFs or volumes to move from one cluster node to another, but in some circumstances nondisruptive migration of either volumes or LIFs from one node to another might be desirable.

Migrating LUNs and volumes from one node to another requires only that the destination node be able to provide a direct path for the host (see "Path Selection" (page 30)).

Migrating a LIF from one node and port to another can be made less administratively burdensome by modifying rather than deleting and recreating it; its IP address or WWPN remains the same, so no fabric zoning or host

changes are needed. SAN LIFs can be modified only when the LIF (but not the port) in question is offline. SAN LIFs can be set administratively offline by using the `network interface modify -status-admin down` command.

Best Practices
Do not exceed the cluster size limit when making changes to cluster membership. For information about the cluster size limit when using block protocols, see the ETERNUS AX/HX Series SAN Configuration Guide in the Fujitsu manual site.

Foreign LUN Import (FLI)

You can import LUNs from third-party arrays and the ETERNUS AB/HB series storage controllers using FC.

This functionality is included in the ONTAP software and does not require a license to use or any additional equipment; it only requires having some of a storage controller's FC or UTA2 ports set to initiator mode during the import process. If using UTA2 ports, those ports would need to be set to their FCP personalities, because FC is the only transport FLI supports.

A storage controller performing a FLI examines a LUN presented from an FC target to create a LUN of identical size and geometry inside an existing volume on its own storage and then creates a block-by-block copy of all the source LUN's data, with offsets if necessary to maintain proper block alignment. Because LUNs created with ONTAP are protocol agnostic, LUNs imported using FC might be presented to hosts using iSCSI the same way any native LUN could be.

This import procedure can be performed in online or offline mode. An online FLI import means that the LUN is offline only if it takes to create an import relationship between the source and destination LUN and for the host to mount the storage at its new location. I/O to that LUN can then continue as usual, with ONTAP multiplexing incoming data to both source and destination until the import is complete and the relationship is broken. During an offline FLI import, both source and destination LUNs are inaccessible to hosts until the import has completed and the import relationship has been broken.

In the ETERNUS AX series, imports can be performed using the ETERNUS AF/DX series as a source. See the ETERNUS AX/HX Series SAN Migration Using Foreign LUN Import in the Fujitsu manual site for details about this version of the FLI procedure.

FUJITSU Storage
ETERNUS AX series All-Flash Arrays,
ETERNUS HX series Hybrid Arrays
Best Practices for ONTAP SAN Configurations

P3AG-6332-01ENZ0

Date of issuance: November 2021
Issuance responsibility: FUJITSU LIMITED


FUJITSU