Lenovo

Best Practices for Modern SAN

ONTAP 9

Abstract

This technical report gives an overview of block protocols in the Lenovo® ONTAP 9 data management software along with best practice recommendations.

Second edition (October 2023)

TABLE OF CONTENTS

1 Overview

This document presents an overview of clustered SAN implementations from the point of view of SAN-attached hosts. It covers new features and describes using prescribed Lenovo AFA SAN configurations to optimize performance. Additionally, it describes best practices for leveraging the high-availability and data mobility features of the ONTAP data management software.

1.1 Audience

This document is intended for system and storage architects who design iSCSI and FC solutions with Lenovo storage solutions running ONTAP 9.4 or later. It assumes that the reader:

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

1.2 Caveats

This document is not meant to be a general introduction to ONTAP administration. An introduction is covered by the ONTAP 9 System Administration Reference and the ONTAP 9 SAN Administration Guide. SAN-related limits for ONTAP clusters that use SAN protocols can be found in the ONTAP 9 SAN Configuration Guide.

For the regularly updated and complete matrix of tested and supported SAN configurations, refer to the Lenovo Storage Interoperation Center (LSIC) to validate that the exact product and feature versions described in this document are supported for your specific environment. The LSIC defines the product components and versions that have been tested together and qualified by Lenovo to work together. Specific results depend on each customer's installation in accordance with published specifications.

2 Summary of Best Practices

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

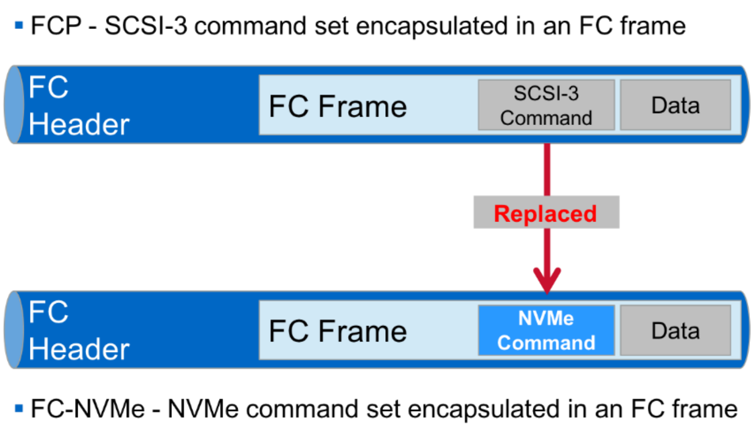
- Create the FCP or iSCSI service at the same time as creating an SVM.
- Combine LUNs that are related, have similar performance requirements, and management requirements into the same volume to reduce administrative effort, operate as a consistency group, and to maximize storage efficiency.
- Optimize performance by increasing the number of volumes and increase the number of LUNs. The ideal layout in most cases will be approximately 8 volumes and 8-16 LUNs.
- Volumes containing LUNs do not need to be junctioned to a namespace in order to serve data using FCP or iSCSI.
- 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.
- Fibre Channel fabrics attached to a clustered Data ONTAP storage cluster must have N_Port ID virtualization (NPIV) enabled.
- 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.
- 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.
- 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.

- Follow a standard procedure on hosts when changing LUN mappings so that they discover new paths and discard paths that have been removed.
- SVMs serving data with FCP or iSCSI need an SVM management interface.
- When using LUN move on a LUN that is deduplicated or compressed, make sure that the destination volume also has these policies enabled.
- If a 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.
- Make use of the LUN move's pause and throttle features for more granular control over LUN replication.
- Use LUN move to shorten existing data mobility and replication workflows.
- Do not exceed the cluster size limit for clusters serving SAN data, as specified in the SAN Configuration Guide.
- Install the Host Utilities Kit on hosts accessing LUNs.
- Use Storage Manager Application Aware Data Management to provision new workloads to best practices, set protection policies, and target specific performance tiers.
- Use 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.
- Set guarantee In Order Delivery on all Fibre Channel switches in your fabrics.

3 ONTAP 9.5 New Features

ONTAP 9.5 introduced its first support of an NVMe over Fabrics (NVMe-oF) protocol— NVMe over Fibre Channel (NVMe/FC). NVMe/FC encapsulates the NVMe command set inside an FC frame, replacing the existing SCSI-3 command descriptor block, as shown in **Reference error for Figure 1**.

Figure 1) FC vs. 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. For more information about NVMe/FC review, see:

- NVMe Configuration over Fibre Channel for Lenovo ONTAP 9.5 or Greater

The principal enhancements added to ONTAP block protocols with ONTAP 9.5 include:

- SnapMirror Synchronous (SM-S)
- Asymmetric Namespace Access (ANA) support added to the NVMe/FC target stack
- One host OS added for NVMe/FC support

3.1 SnapMirror Synchronous

ONTAP 9.5 introduced synchronous replication with the introduction of SM-S. MetroCluster provides synchronous replication, however, it requires all data to be mirrored. The new synchronous replication feature that was released with ONTAP 9.5 provided customers the ability to choose which workloads needed synchronous replication instead of the previous all-or-nothing approach. 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 point 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.
- 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. It is supported with all shipping AFA and Hybrid platforms with at least 16GB of memory and that support ONTAP 9.5 and later.

3.2 ONTAP 9.5 Asymmetric Namespace Access Support

ONTAP 9.5 introduced ANA 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 MPIO or Device Mapper Multipathing (DM-Multipath) provide path management and failover capabilities for NVMe/FC.

NVMe Configuration over Fibre Channel for Lenovo ONTAP 9.5 or Greater

3.3 New NVMe/FC Qualified Hosts

ONTAP 9.5 added the following two hosts for support of NVMe/FC:

- SUSE Enterprise Linux 15 (Kernel 4.12.14-150-27.1).

Note: SUSE Enterprise Linux 12 SP3 is not supported for ANA and is not recommended.

4 ONTAP 9.6 New Features

ONTAP 9.6 SAN enhancements include:

- ONTAP SAN and cluster resiliency and hardening
- NVMe enhancements:
  - NVMe volume move
  - Growing NVMe interoperability
  - 512-byte block-size support
  - NVMe adds VMware copy and write (Compare and Write/Atomic test and Set [CAW/ATS) support

4.1 512-Byte Block Size Support

ONTAP 9.6 added a 512-byte block size option to NVMe namespaces in addition to the 4k block size ONTAP natively supports. This feature was added to 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 now available for both the `vserver nvme namespace create` command and the matching API.

4.2 NVMe Adds VMware Compare and Write Support

ONTAP 9.6 added VMware Compare and Write/Atomic Test and Set (CAW/ATS) fused operation support to ONTAP to support features such as VMware Storage vMotion. VMware vStorage APIs - Array Integration (VAAI). 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.

4.3 NVMe Adds Nondisruptive Volume Move

NVMe-oF adds volume move support, which allows storage administrators to nondisruptively move volumes that contain NVMe namespaces from one aggregate to another. In previous releases, volume move operations for volumes that contained one or more mapped NVMe namespaces would fail. As a workaround, users would unmap the namespace from the subsystem prior to performing the move. This workaround is no longer necessary.

4.4 NVMe Adds QoS Maximums Support

ONTAP 9.6 adds support for volume-level QoS policies (storage virtual machine [SVM]-level policies are implicitly supported). Ceilings or QoS maximums are available, whereas floors and namespace-granular QoS are not supported.

NVMe QoS support adds a new event management service (EMS) event: `nvmf.qos.mismatched.policy`. The EMS was created to warn 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.

5 ONTAP 9.7 New Features

ONTAP 9.7 presents the following SAN enhancements:

- Increased maximum number of volumes per node from 1,000 to 2,500.
- SM-S added NVMe-oF support.
- A new version of Storage Manager is debuted with ONTAP 9.7.

5.1 New Storage Manager Version

This new feature is part of an effort to rewrite Storage Manager with a principle design goal of simplifying operations. The new version gains a cleaner, modern look. Storage Manager can now display the REST API calls for every action performed. This functionality will assist 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.

6 ONTAP 9.8 New Features

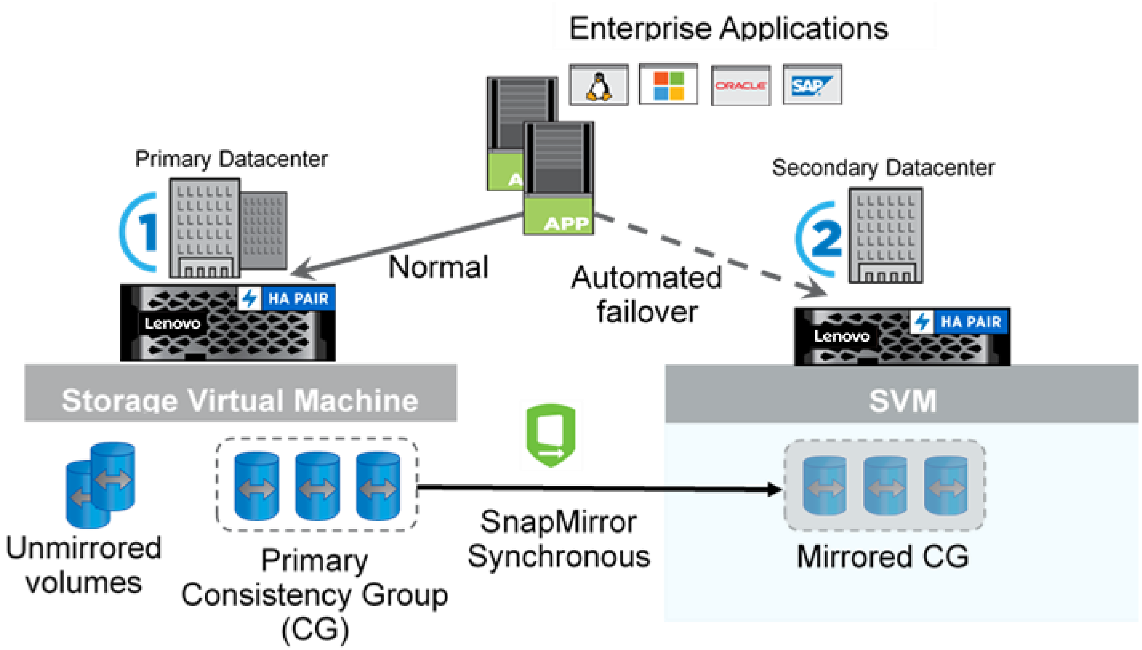
ONTAP 9.8 adds several new features, some of which are available on unified platforms. The expectation is that these features will likely be added to unified ONTAP clusters in upcoming releases of ONTAP. ONTAP's new SAN features include:

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

6.1 SnapMirror Business Continuity

ONTAP 9.8 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 2 shows the SMBC topology. For more information, see SnapMirror Business Continuity (SM-BC) for ONTAP.

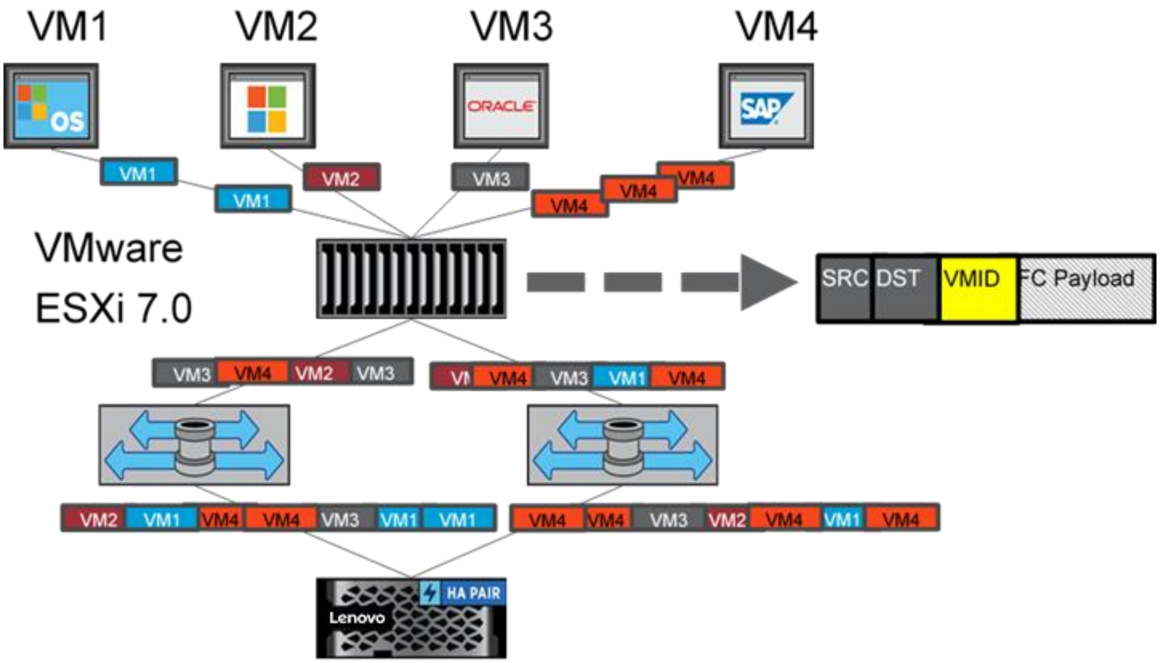
Figure 2) SMBC topology.



6.2 Virtual Machine ID

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 Gen 6 and 7 switches.

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.

6.3 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 could be released more quickly 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.

7 ONTAP 9.9.1 New Features

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

- Single LUN performance improvements
- Nested igroups
- An FLI field qualification script
- VMware vSphere Virtual Volumes (vVols) support for NVMe/FC
- NVMe-oF remote I/O support
- ThinkSystem Storage Manager Ansible playbook generation

7.1 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 controlllers. We anticipate that the bulk of the benefits to single LUN performance is likely to be most valuable in two areas:

- Virtualization where a datastore is being backed by an LUN.
- POCs that are incorrectly set up to run I/O to a single LUN and then use those performance numbers to extrapolate performance. Several competitors attempt to define competitive POCs this way because they could take advantage of this previous weakness.

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

7.2 Nested igroups

ONTAP 9.9.1 adds nested igroups to simplfy 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 si 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).

7.3 NVMe/FC vVols Support

ONTAP 9.9.1 added NVMe/FC vVol support. This enhancement allows VMware administration teams to manage and automate storage using vVols after vSphere supports that option within vVols/vCenter.

7.4 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.

Figure4 illustrates the NVMe-oF without remote I/O support.

Figure 4) NVMe-oF without remote I/O support.

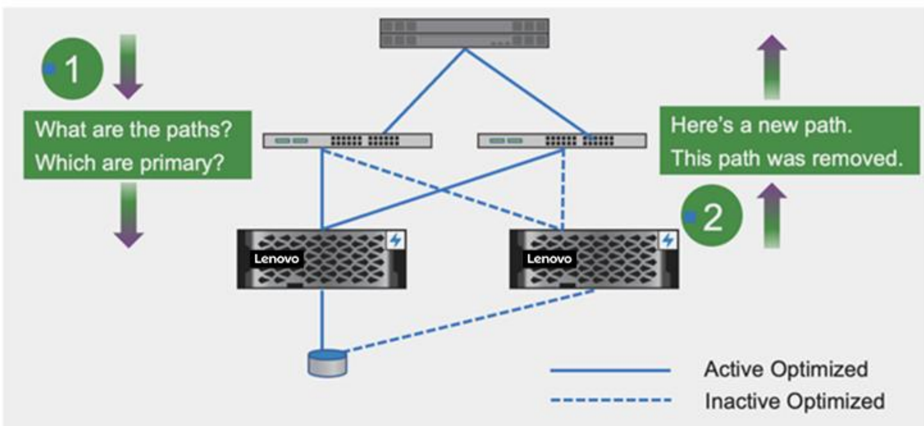
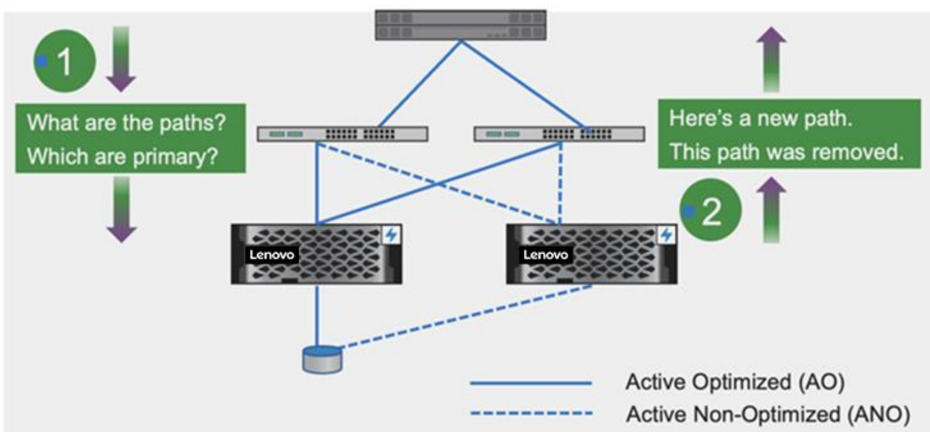


Figure 5) NVMe-oF with remote I/O support.



Comparing Figure4 and Figure5, 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.

7.5 Ansible Playbook Support in Unified Manager

Unified Manager for adds Ansible Playbook autocreation for all Storage Manager workflows. This feature can be very helpful for teams that are trying to grow their DevOps practice and processes.

Figure6, Figure7, and Figure8 are examples of generating an Ansible YAML playbook from the Storage Manager Add LUNs workflow.

Figure 6) Unified Manager Add LUNs workflow.

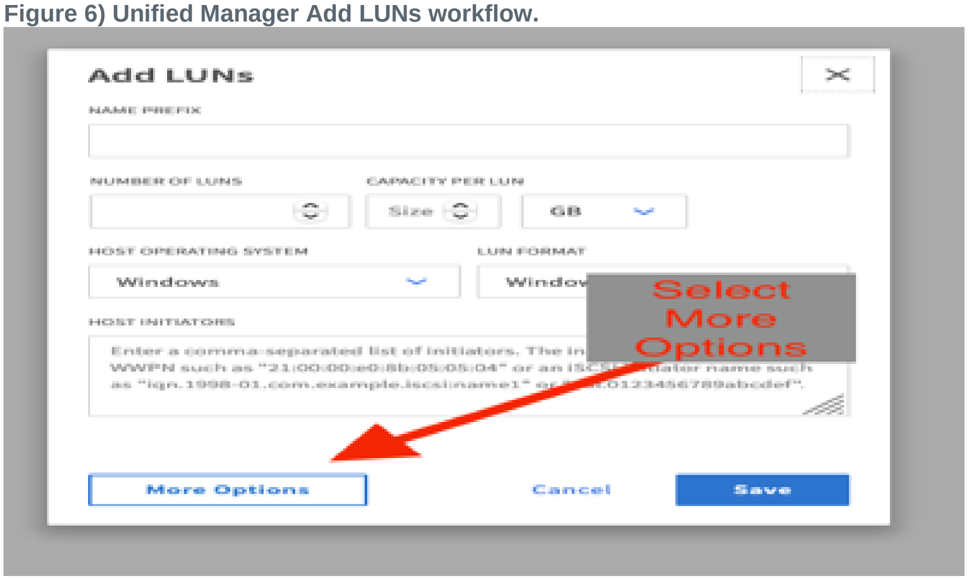


Figure 7) Unified Manager Add LUNs: Save to Ansible Playbook.

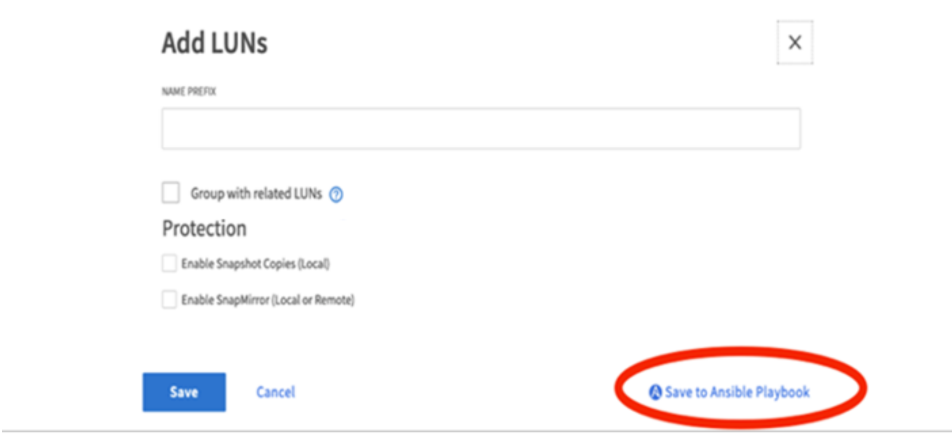
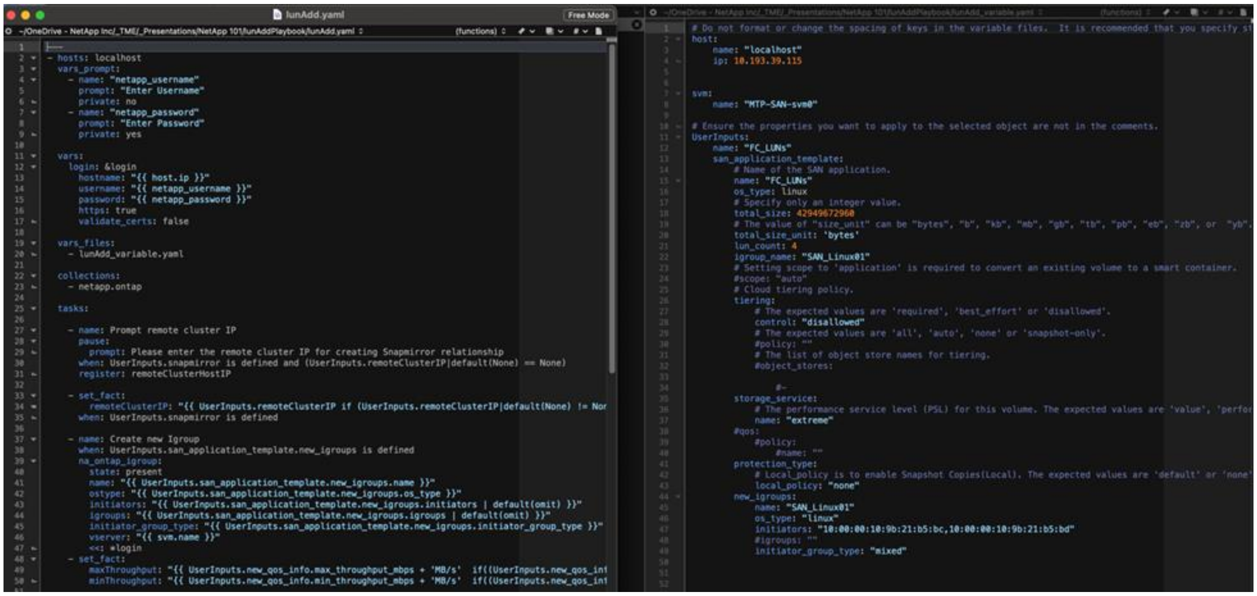


Figure 8) Unified Manager -generated Add LUN Ansible YAML files.



8 ONTAP 9.10.1 New Features

8.1 NVMe Protocol Support for TCP

ONTAP 9.10.1 adds a second NVMe over Fabrics (NVMe-oF) transport for NVMe. NVMe/TCP is added to NVMe/FC. Now ONTAP NVMe-oF support offers protocols that can use either Fibre Channel or Ethernet/TCP/IP respectively.

9 ONTAP 9.11.1 New Features

9.1 Non-destructive Conversion from LUN to NVMe Namespaces and Vice Versa

ONTAP 9.11.1 also adds an on-box very fast in-place conversion utility that performs bidirectional in-place conversions between NVMe namespaces and SCSI LUNs. The utility is very efficient because it does not modify any data, only the metadata describing the LUN or namespace.

10 ONTAP 9.12.1 New Features

10.1 NVMe Limits Increase

NVMe limits are continuing to increase with the goal of at least parity with SCSI block limits (iSCSI and FCP). The limits that increase in ONTAP 9.12.1 are as follows:

- Up to 8,000 subsystems on a single SVM or cluster
- NVMe can now scale to up to 12-node clusters, which is parity with SCSI block protocols
- NVMe/FC supports up to 256 controllers per node
- NVMe/TCP supports up to 2,000 controllers per node

10.2 NVMe/TCP Support for Secure Authentication

ONTAP 9.12.1 introduces the DH-HMAC-CHAP authentication protocol; this enables secure unidirectional/bidirectional authentication between an NVMe/TCP host and a controller.

10.3 MetroCluster IP support for NVMe

MetroCluster IP support for NVMe is added for 4-node MetroCluster IP.

11 ONTAP 9.12.1 P2 new features

11.1 Larger Maximum Sizes for LUNs, Files, and Volumes

The new larger LUN, file and volume maximum sizes are available with ONTAP 9.12.1P2 for all ONTAP platforms. The new maximum sizes are:

- Maximum LUN size: 128TiB
- Maximum file size: 128TiB
- Maximum volume size: 300TiB

After upgrading to ONTAP 9.12.1P2 or later, you can create new LUNs, files, and volumes up to the new maximum sizes. You can also grow existing LUNs, files, and volumes to the new maximum sizes.

12 ONTAP and SAN Protocols

12.1 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 that consists of two 10 Gigabit Ethernet switches (switched clusters) or direct attached Ethernet connections for switchless clusters.

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.

12.2 Considerations for Optimizing SAN Performance

ONTAP is optimized to use as many processor cores as possible 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 non-distributable threads can limit the maximum performance that could be achieved if workloads are properly split across multiple objects. By using more objects, we 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. The word volume can be used to mean many things, depending on the storage vendor being considered. 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 ease administration, can be used as a consistency group, and increase storage efficiencies, we also need to consider how many volumes should be provisioned to optimize performance.

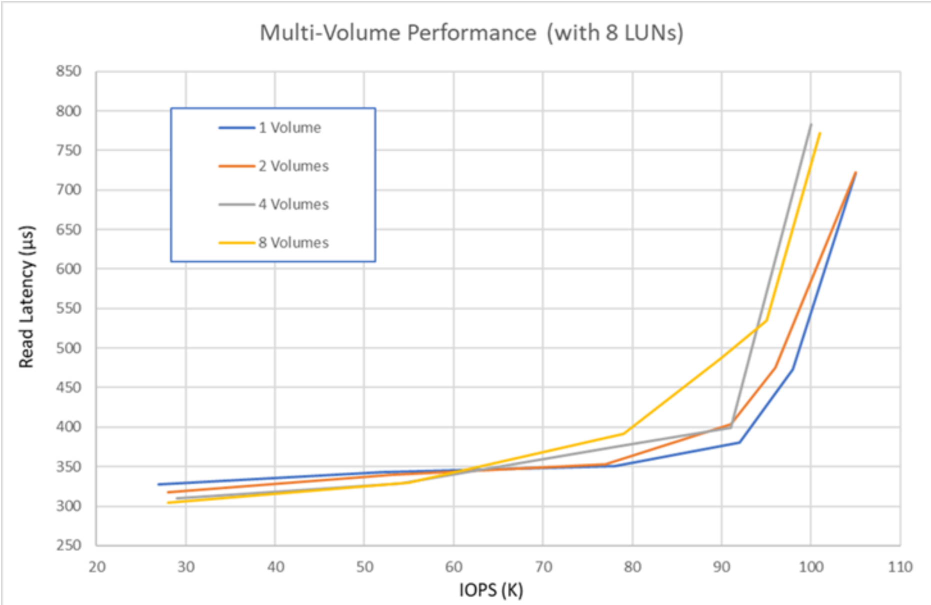
As previously mentioned, some processing threads can't be distributed across multiple processor cores. Not being able to spread this work across multiple cores can be managed by spreading the workload across multiple volumes. By doing this, you can break up the amount of non-distributable work that has to be done by a single core. Each volume has its own non-distributable threads, so the more volumes we 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, performance can be optimized by increasing the number of volumes up to a point and after that more volumes don't increase performance, but they do increase complexity and reduce storage efficiency.

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

Note: Figure9 and Figure10 show the effects of adding more volumes and LUNs are displayed for illustrative purposes. They are not meant to guarantee or benchmark specific workloads or provide estimates of performance.

Figure 9) The effect of spreading work across more volumes.



Best Practice
In most cases, Lenovo recommends using between 8–16 volumes to maximize performance. This is assuming that these are the only volumes on a given controller. If other volumes are present, then you can generally consider using up to 8 additional LUNs to maximize performance.

Best Practice

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.

12.3 LUNs

LUNs, either iSCSI or FC, have some threads that can't be distributed across multiple processor cores. As a result, this means that Lenovo recommends using more smaller LUNs, as opposed to fewer larger

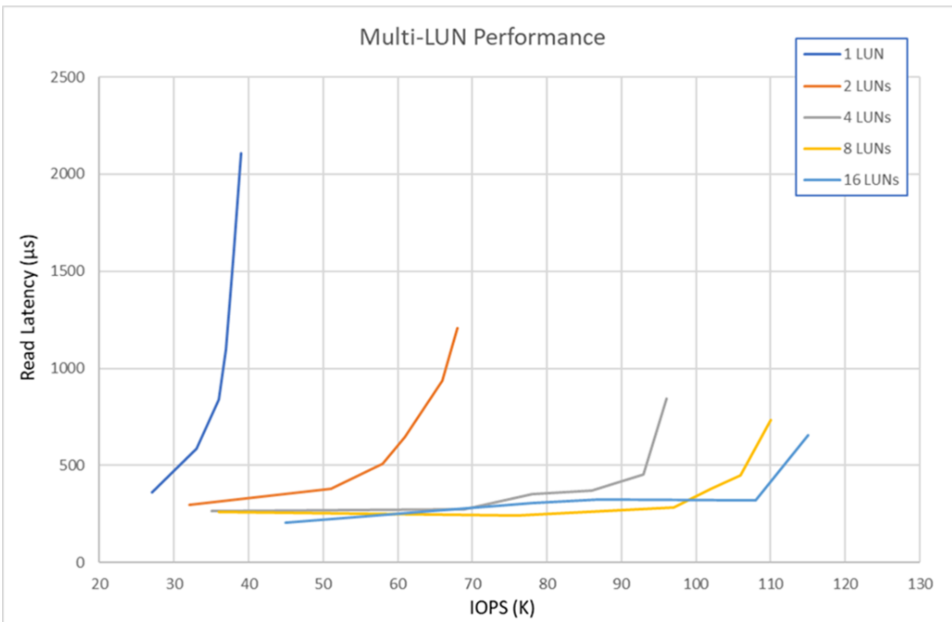
LUNs, if performance is the primary consideration. Like volumes, workload performance can be optimized by spreading work across multiple LUNs, this allows more processor cores to be used to concurrently process workload I/O.

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 OS. 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 Practice
Lenovo recommends using more smaller LUNs versus fewer bigger ones. Ideally, between 8–16 LUNs.

Figure 10) 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 rely on 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. Figure9 and Figure10 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, 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 be working with a solution architect to 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 support around 35K IOPS. Two LUNs almost doubles the limit.
- Benefits start diminishing as you reach eight LUNs. 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 99% of all databases require.

  Note: Figure9 and Figure10 show that the numbers in the conclusions discussed above were from a specific test. The conclusions are valid, and they illustrate the Lenovo 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.

12.4 Scalable SAN

When an SVM is first created and a block protocol (FC or iSCSI) is enabled, the SVM gets either an FC worldwide name (WWN) or an iSCSI qualified name (IQN), respectively. This identifier is used regardless of which physical node is being addressed by a host, with ONTAP making sure that SCSI target ports on all of the cluster nodes work together to present a virtual, distributed SCSI target to hosts that are accessing block storage.

In practice, this means that no matter which physical node a host is communicating, it is communicating with the same SCSI target. This method of access presents new opportunities for data resiliency and mobility, and it also has implications for best practices when accessing data using block protocols on a cluster.

Best Practice
When creating iSCSI or FC LIFs for the first time for an existing SVM, make sure that the FC and/or iSCSI service for that SVM has been created and is turned on by using the `fcp show` or `iscsi show` command or by navigating to the Storage Virtual Machines → Configuration → Protocols pane in ThinkSystem Storage Manager. This step is not necessary if the SVM was originally set up to serve these protocols by using an automated process such as the Storage Manager SVM Setup wizard.

12.5 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 containers 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 FC or iSCSI.

Best Practice
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.

12.6 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 iSCSI or FC connection 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.

12.7 Path Selection

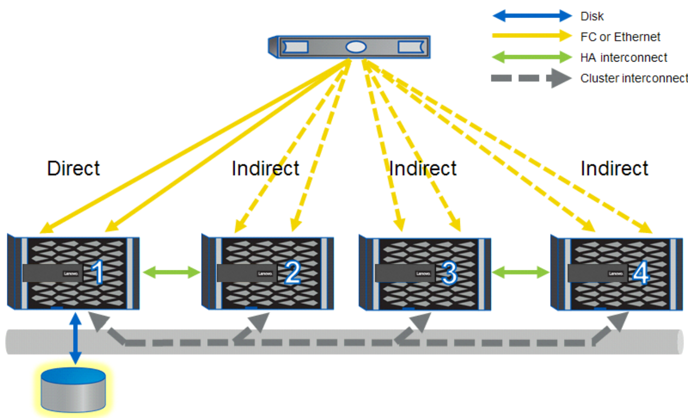
Even though every LIF owned by an SVM accepts writes and read requests for its LUNs, only one of the cluster nodes actually owns the disks backing that LUN at any given moment. This effectively divides available paths to a LUN into two types: direct and indirect paths.

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.

Figure11 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 11) Overview of paths in ONTAP.



Note: In this illustration, each numbered chassis represents a node for explanation/illustration purposes.

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 multipath I/O (MPIO) software running on the host that is accessing LUNs, with the result that only direct paths are used during normal operation.

Best Practice
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 section titled "Path Management and Selective LUN Mapping".

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.

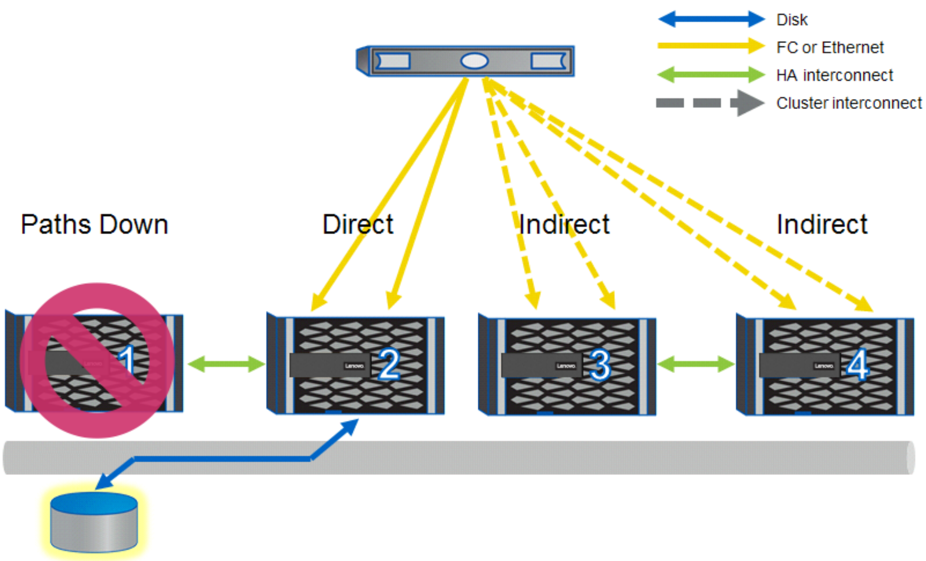
Because of this difference in behavior, Ethernet LIFs that serve data by using the iSCSI protocol cannot also serve data by using a NAS protocol.

12.8 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.

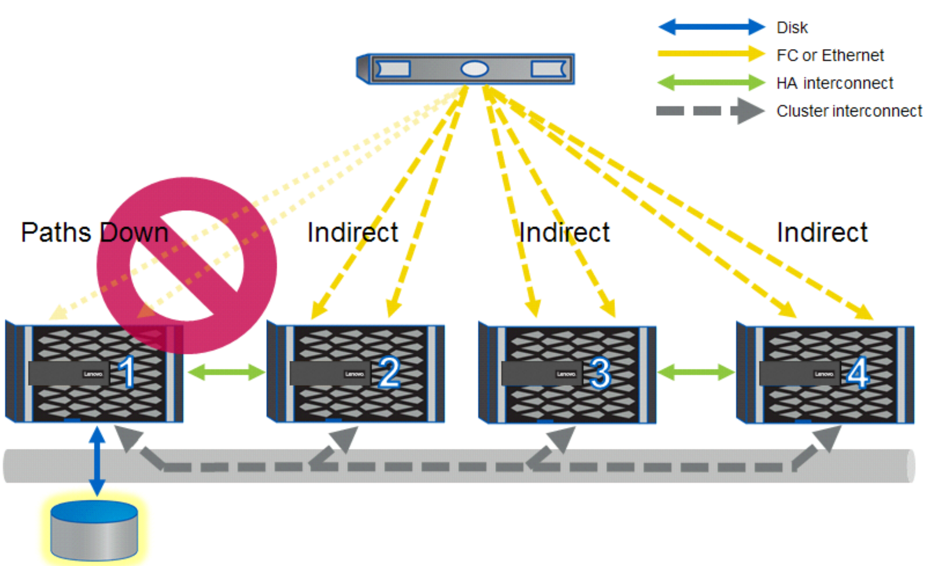
Figure 12) Paths during HA failover.



Note: In this illustration, each numbered chassis represents a node for explanation/illustration purposes.

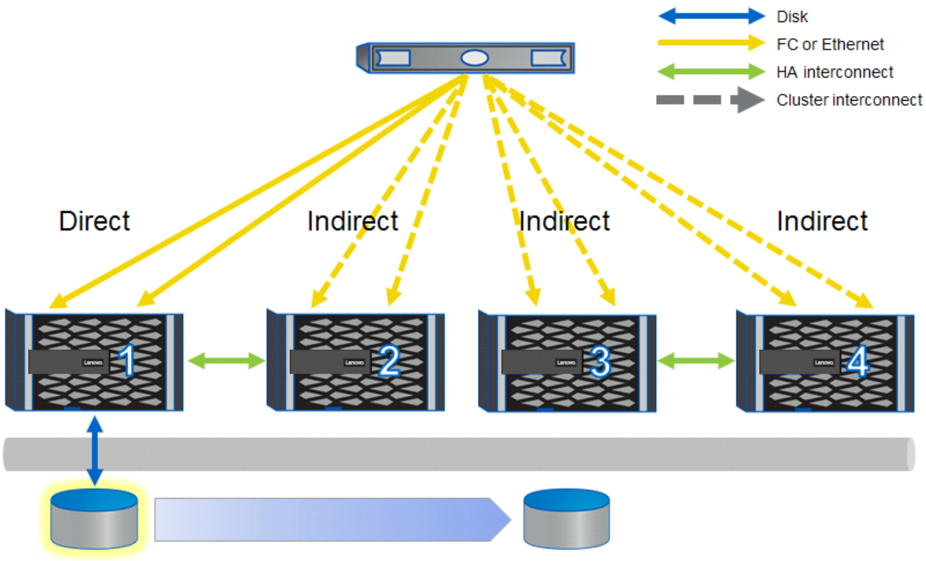
Port or switch failure. In a port or switch failure, no more direct paths are available. Path priority remains the same, and MPIO software running on the host selects alternate indirect paths until a direct path becomes available again. The ALUA path states do not change.

Figure 13) Paths during port or switch failure.



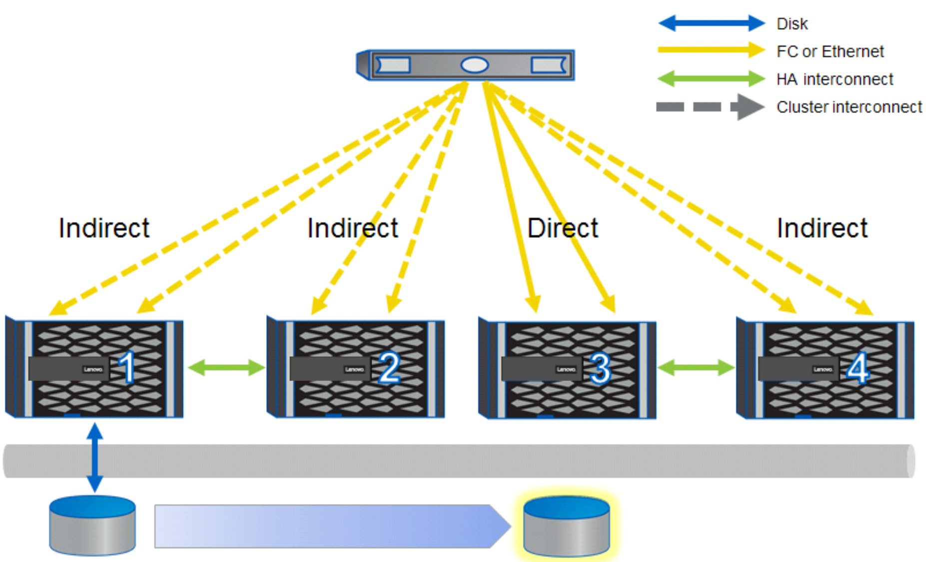
Volume or LUN mobility. A volume is moved transparently between nodes by using `volume move` functionality, or a LUN is moved transparently between nodes using `lun move`.

Figure 14) 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 15) Paths after volume or LUN mobility.



For a LUN move, the cutover happens 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 section titled "LUN Management".

12.9 FC and NPIV

An ONTAP node uses N_Port ID virtualization (NPIV) to permit every logical interface to log in to an FC fabric with its own worldwide port name (WWPN). This permits a host connected to the same FC fabric 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 Practice

NPIV is required for FC LIFs to operate correctly. Before creating FC LIFs, make sure that any fabrics attached to an ONTAP system have NPIV 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         ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..    ..  ..  ..  ..
```

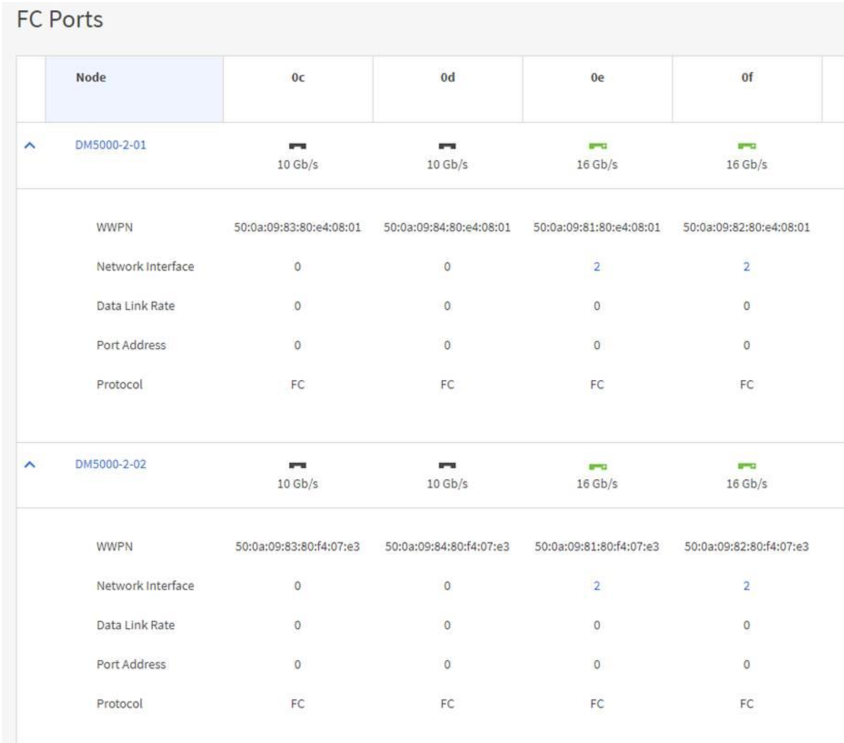
From the storage administration console, it is not possible to inquire about NPIV status on an attached switch directly, but examining the local FC topology can show whether fabric switch ports have NPIV enabled.

Best Practice

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 Adapters pane under Network → FC Adapters in Storage Manager. In the new Storage Manager interface (sysmgr/v4), use the FC Ports pane under Network → FC Ports.

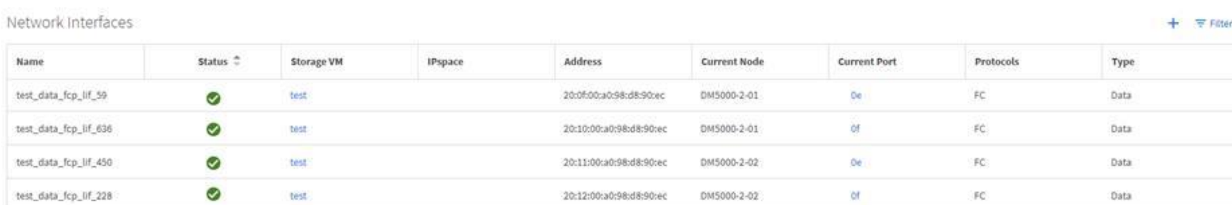
Instead, use only virtual WWPNs (WWPNs starting with 20:) visible in the output of the `network interface show` command and under Network → Network Interfaces in Storage Manager. In the new Storage Manager interface (sysmgr/v4), use Network Overview pane under Network → Overview → Network Interfaces, as shown in Figure 27.

Figure 16) FC adapters in Storage Manager.


FC Ports

Node	0c	0d	0e	0f
DM5000-2-01	10 Gb/s	10 Gb/s	16 Gb/s	16 Gb/s
WWPN	50:0a:09:83:80:e4:08:01	50:0a:09:84:80:e4:08:01	50:0a:09:81:80:e4:08:01	50:0a:09:82:80:e4:08:01
Network Interface	0	0	2	2
Data Link Rate	0	0	0	0
Port Address	0	0	0	0
Protocol	FC	FC	FC	FC
DM5000-2-02	10 Gb/s	10 Gb/s	16 Gb/s	16 Gb/s
WWPN	50:0a:09:83:80:f4:07:e3	50:0a:09:84:80:f4:07:e3	50:0a:09:81:80:f4:07:e3	50:0a:09:82:80:f4:07:e3
Network Interface	0	0	2	2
Data Link Rate	0	0	0	0
Port Address	0	0	0	0
Protocol	FC	FC	FC	FC

Figure 27) Network interfaces in Storage Manager.


Network Interfaces

Name	Status	Storage VM	IPspace	Address	Current Node	Current Port	Protocols	Type
test_data_fcp_lif_59		test		20:0f:00:a0:98:d8:90:ec	DM5000-2-01	0e	FC	Data
test_data_fcp_lif_636		test		20:10:00:a0:98:d8:90:ec	DM5000-2-01	0f	FC	Data
test_data_fcp_lif_450		test		20:11:00:a0:98:d8:90:ec	DM5000-2-02	0e	FC	Data
test_data_fcp_lif_228		test		20:12:00:a0:98:d8:90:ec	DM5000-2-02	0f	FC	Data

12.10 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 Lenovo storage controller, see the SAN Configuration Guide.

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 portsets.

12.11 Selective LUN Mapping

Selective LUN mapping (SLM) is an addition to the LUN mapping table already existing in a Data 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 may be mapped to multiple igroups (especially in host-based clustering scenarios), and because igroups may have multiple LUNs mapped.

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 from the output after running `lun mapping show -reporting-nodes <node_name>`.

Note: By default, any LUN mappings created 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.

However, a LUN mapping may also contain any or all other nodes in the cluster, as long as they are grouped in HA pairs, or it may 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 may choose which storage controllers present paths in a highly granular fashion.

12.12 Portsets

Portsets permit 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. They are used for the purpose of limiting the number of paths presented in a scenario where storage controllers and SVMs have more than one target LIF available per FC fabric or Ethernet network. In such cases, for example, it may be considered desirable to limit traffic for a set of hosts or for an application to a dedicated subset of the total number of target ports.

Note: A LIF that is currently a member of a portset cannot be modified until it is removed from the portset. It can be added to the portset after modification, but care should be taken to leave enough LIFs in the portset 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.

12.13 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 ONTAP Network Management Guide.)

Additionally, an SVM-level administrative account should be available. The default account created during SVM creation is the vsadmin account and the password is assigned during account creation, but it must first be assigned a password with the security login password -username vsadmin -vserver <svm> command and then unlocked by using the security login unlock -username vsadmin -vserver <svm> command. In the new Storage Manager interface (sysmgr/v4), the option to create an administrator account is selectable under “Storage VM Administration” during Storage VM creation. The default username is vsadmin, a password is assigned during creation and will be unlocked. For more details, see “Delegating Administration to SVM Administrators” in the ONTAP System Administration Guide.

When administrating a cluster using Storage Manager, an SVM management LIF may be created during normal LIF creation by selecting the interface role as “serves data” and checking “Enable Management Access”. In the new Storage Manager interface (sysmgr/v4), the management LIF can be created during SVM creation by checking “Manage administrator account” and “Add a network interface for storage VM management”.” See Figure8, Figure 39.

Figure 18) Creating a management LIF during SVM creation (/sysmgr/v4).

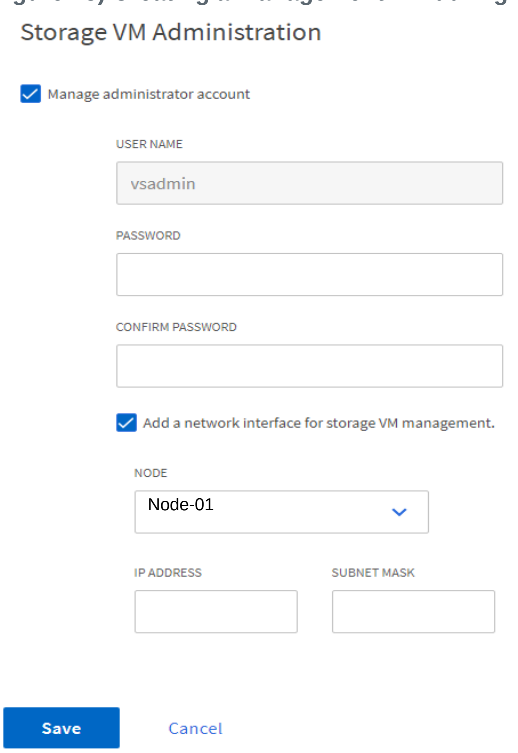
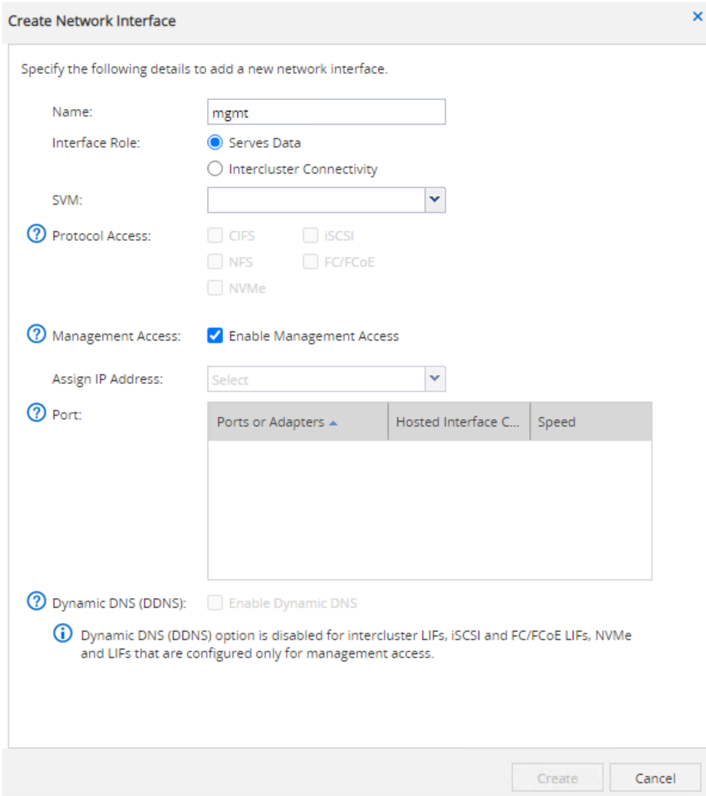


Figure 39) Creating a management LIF for an existing SVM.



12.14 LUN Management

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 LUN Management 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 do not arrive compressed or deduplicated at their destination.

Best Practice
If a destination volume previously has not 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 LUN move 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 may be necessary. If the destination already participates in such a relationship, it may be necessary to take actions caused by more space being consumed by the data replication destination.

Best Practice
When using LUN move in conjunction with software external to the storage cluster to manage Snapshot copies containing LUNs, make sure that the software is aware of the capabilities of LUN management and can (for example) restore a LUN from Snapshot copies in volumes in which it may no longer exist. If this is not possible, LUN move may 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 Practice
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 with LUN management:

- 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.

LUN Management 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:

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 non-disruptively; 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.

12.15 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 may 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 portsets 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 SAN Configuration Guide 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. See section LUN Management and Selective LUN Mapping: Discovering and Discarding Paths 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. • 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.

13 Scalable SAN Key Value Propositions and Features

This section highlights principal design goals. These goals included providing a unified architecture at scale that enables non-disruptive 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.

13.1 SVM as Unified Target and Unit of Management

Storage controllers running Data ONTAP, when a member of an HA configuration, already present a single WWNN to an attached FC fabric. The storage cluster extends this 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.

13.2 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. For the latest full details about SAN configuration limits, see the SAN Configuration Guide. For a summary, see the table below.

Version of ONTAP	9.8
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

13.3 Cluster-Wide Consistency Groups

Snapshot consistency groups are a way for Snapshot copies on multiple storage controllers to be taken simultaneously, allowing a host with LUNs served from multiple volumes within an SVM to synchronize Snapshot copy creation, which allows for consistent Snapshot copies across multiple LUNs even when those LUNs reside on multiple cluster nodes.

Rather than directing a Snapshot copy to be taken on multiple storage controllers at once, a host can take a copy across multiple cluster nodes and volumes simultaneously with a single command. Consistency groups work on a per-SVM basis, so any volumes owned by an SVM that is receiving the command are candidates for a Snapshot copy.

13.4 Intracluster LUN and LIF Mobility

With ONTAP, volumes are allowed to be moved non-disruptively from any node to any other node participating in the cluster. However, 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 non-disruptive 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 section Path Selection)

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 Practice
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 ONTAP 9 SAN Configuration Guide.

13.5 Foreign LUN Import

LUNs can be imported from third-party arrays and first-party DE Series storage controllers using FC.

This functionality is included 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 foreign LUN import 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 may 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.

For an overall FLI migration strategy, see SAN Migration Using Foreign LUN Import.

14 Host Integration

14.1 Lenovo Host Utilities Kit

Installation of the Host Utilities Kit sets timeout and other operating system–specific values to their recommended defaults and includes utilities for examining LUNs provided by Lenovo storage.

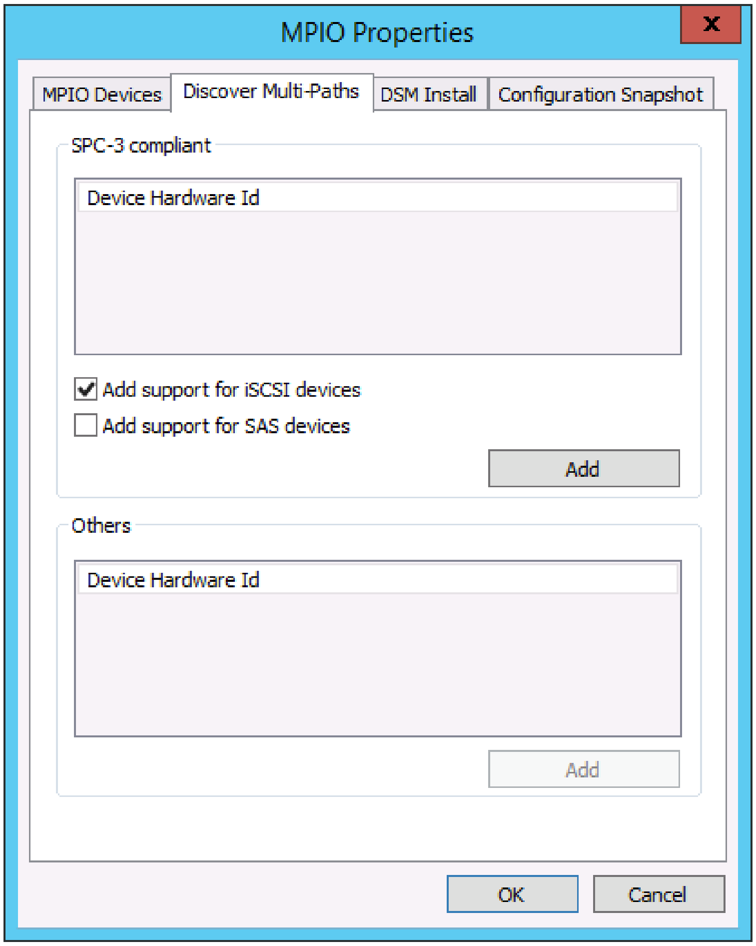
14.2 Microsoft Windows

Microsoft Windows and Native MPIO

To operate as intended, accessing ONTAP storage clusters requires that hosts use MPIO and ALUA. In the case of Microsoft Windows 2008 and Windows 2012, these are natively supported whenever the multipath I/O feature is installed.

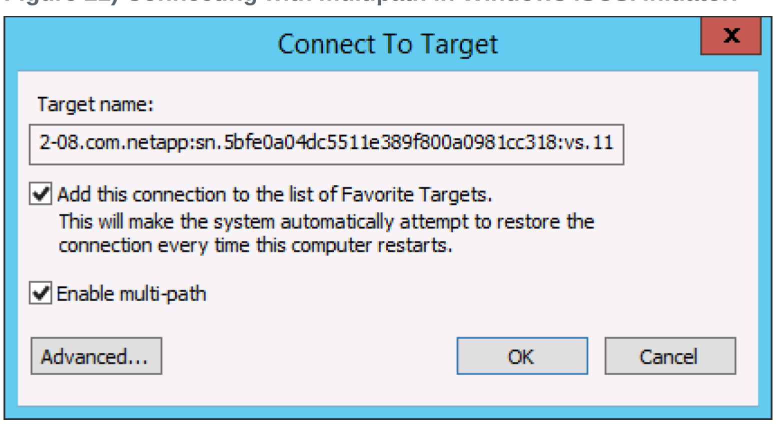
When using the iSCSI protocol, it's necessary to tell Windows to apply multipathing support to iSCSI devices in the MPIO properties administrative application: From the Discover Multi-Paths tab, select the Add Support for iSCSI Devices option, and click Add.

Figure 20) MPIO properties in Windows 2012.



It's also necessary to create multiple sessions from the host initiators to the target iSCSI LIFs on the storage cluster. This can be accomplished using the native iSCSI initiator: Select the "Enable multi-path" option when logging on to a target.

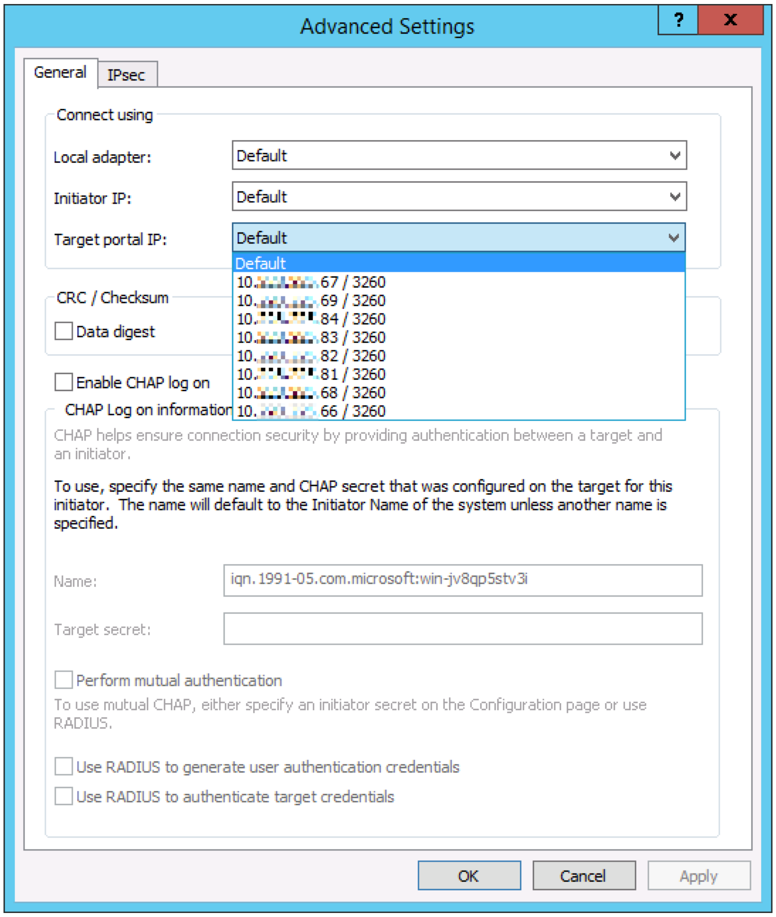
Figure 21) Connecting with multipath in Windows iSCSI initiator.



To manually create additional sessions, highlight the corresponding target in the Targets pane of the iSCSI initiator and click Log on. Make sure that the session is automatically restored after the next reboot and that the new session is identified as a multipath session by selecting both checkboxes.

Click Advanced. From the Target Portal IP drop-down menu select the IP address of the logical interface that is the target of the new iSCSI session.

Figure 22) Multiple target ports in Windows iSCSI initiator.



Sessions can also be managed by using the SnapDrive® iSCSI management pane. This is the preferred method, because SnapDrive remembers which target logical interfaces already have an established session and preselects an unused target portal.

Host Utilities Kit

The Lenovo Host Utilities Kit can also be installed. The appropriate values are changed in the Windows registry to optimize performance and provide correct operation during failover events.

14.3 Linux

Host Utilities Kit

The Lenovo Host Utilities Kit contains utilities that are useful for viewing LUN configuration at the SVM level. Using the Host Utilities Kit can provide extended information about the SVM to which an attached LUN belongs, in addition to its volume and path name in the SVM context.

```
# sanlun lun show all
controller(7mode)/               device       host                   lun
vserver(Cmode)  lun-pathname     filename     adapter    protocol    size     mode
--------------------------------------------------------------------------------
vs              /vol/vol1/linux1 /dev/sdcx    host1      FCP         25g      C
vs              /vol/vol2/linux2 /dev/sdcw    host1      FCP         25g      C
vs              /vol/vol3/linux3 /dev/sdck    host1      FCP         25g      C
```

Additionally, the Host Utilities Kit can be used to display which of an SVM's logical interfaces are providing the direct and indirect paths for a given LUN, labeled here as primary for direct paths and secondary for indirect paths.

```
# sanlun lun show -p
                    ONTAP Path: vs:/vol/vol1/linux1
                           LUN: 0
                      LUN Size: 25g
                          Mode: C
                   Host Device: 3600a09803246664c422b2d51674f7470
              Multipath Policy: round-robin 0
            Multipath Provider: Native
--------- ---------- ------- ------------ -------------------------------------
host      vserver
path      path       /dev/   host         vserver
state     type       node    adapter      LIF
--------- ---------- ------- ------------ -------------------------------------
up        primary    sdfo    host0        fcoe_lif_1
up        primary    sdfk    host1        fcoe_lif_2
up        secondary  sdga    host0        fcoe_lif_3
up        secondary  sdge    host1        fcoe_lif_4
up        secondary  sdgm    host1        fcoe_lif_5
up        secondary  sdgj    host0        fcoe_lif_6
up        secondary  sdfw    host0        fcoe_lif_7
up        secondary  sdgq    host1        fcoe_lif_8
```

SnapDrive

As with Microsoft Windows, the unit of management when using SnapDrive and ONTAP storage clusters is at the individual SVM rather than at the node or cluster level. The `SnapDrive config set` command must be used in conjunction with a management logical interface and an SVM administrative user, as described in section Management Interfaces, in order to administer LUNs attached by using iSCSI or FC from an attached host.

```
# snapdrive config set vsadmin vs
Password for vsadmin:
Retype password:

# snapdrive config list
username     appliance name    appliance type
-------------------------------------------
vsadmin      vs                StorageSystem
```

Where to Find Additional Resources

To learn more about the information that is described in this document, review the following documents and/or websites:

- SAN Migration Using Foreign LUN Import
  https://download.lenovo.com/storage/san_migration_using_foreign_lun_import_v2.pdf
- ONTAP 9 SAN Configuration Guide
  http://thinksystem.lenovofiles.com/help/topic/san_configuration_guide/overview.html

- ONTAP 9 SAN Administration Guide
  http://thinksystem.lenovofiles.com/help/topic/san_administration_guide/overview.html
- ONTAP 9 System Administration Reference
  https://thinksystem.lenovofiles.com/storage/help/topic/system_administration_guide/overview.html?cp=2_9_1_6
- ONTAP 9 Network Management Guide
  https://thinksystem.lenovofiles.com/storage/help/topic/ontap_networking/networking.pdf

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

Contact Support

You can contact Support to obtain help for your issue.

You can receive hardware service through a Lenovo Authorized Service Provider. To locate a service provider authorized by Lenovo to provide warranty service, go to https://datacentersupport.lenovo.com/serviceprovider and use filter searching for different countries. For Lenovo support telephone numbers, see https://datacentersupport.lenovo.com/supportphonelist for your region support details.

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area.

Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document is not an offer and does not provide a license under any patents or patent applications. You can send inquiries in writing to the following:

Lenovo (United States), Inc.

8001 Development Drive

Morrisville, NC 27560

U.S.A.

Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT,

MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Trademarks

