Technical Report

ONTAP Select on VMware

Product Architecture and Best Practices

Tudor Pascu, NetApp
November 2019 | TR-4517

Attention: Beginning with ONTAP Select 9.7, the content in this technical report, as well as the core ONTAP Select 9.7 documentation, is maintained in the new ONTAP Select content repository.

If you are new to ONTAP Select or if you are working with ONTAP Select 9.7 or later, you can find the most current ONTAP Select documentation at https://docs.netapp.com/us-en/ontap-select.

NetApp®

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Attention: Beginning with ONTAP Select 9.7, the content in this technical report, as well as the core ONTAP Select 9.7 documentation, is maintained in the new ONTAP Select content repository.

If you are new to ONTAP Select or if you are working with ONTAP Select 9.7 or later, you can find the most current ONTAP Select documentation at https://docs.netapp.com/us-en/ontap-select.

1 Introduction

NetApp® ONTAP® Select is the NetApp solution for the software-defined storage (SDS) market. ONTAP Select brings enterprise-class storage management features to the software-defined data center and extends the Data Fabric solution to extreme edge use cases including IoT and tactical servers.

This document describes the best practices that should be followed when building an ONTAP Select cluster, from hardware selection to deployment and configuration. Also, it aims to answer the following questions:

- How is ONTAP Select different from engineered FAS storage platforms?
- Why were certain design choices made when creating the ONTAP Select architecture?
- What are the performance implications of the various configuration options?

1.1 Software-Defined Infrastructure

The implementation and delivery of IT services through software enable administrators to rapidly provision resources with a speed and agility that was previously impossible.

Modern data centers are moving toward software-defined infrastructures as a mechanism to provide IT services with greater agility and efficiency. Separating out IT value from the underlying physical infrastructure allows IT services to react quickly to changing needs by dynamically shifting infrastructure resources to where they are needed most.

Software-defined infrastructures are built on these three tenets:

- Flexibility
- Scalability
- Programmability

Software-Defined Storage

The shift toward software-defined infrastructures might be having its greatest impact in an area that has traditionally been one of the least affected by the virtualization movement: storage. Software-only solutions that separate out storage management services from the physical hardware are becoming more commonplace. This is especially evident within private cloud environments. Enterprise-class, service-oriented architectures are designed from the ground up to be software defined. Many of these environments are built on commodity hardware: white-box servers with locally attached storage that have software controlling the placement and management of user data.

This is also seen within the emergence of hyper converged infrastructures (HCIs), a building-block style of IT design with bundled compute, storage, and networking services. The rapid adoption of hyper converged solutions over the past several years has revealed a desire for simplicity and flexibility. However, companies have started replacing enterprise-class storage arrays with a more customized, make your own model by building storage management solutions on top of home-grown components. Therefore, a new set of problems has emerged.

In a commodity world in which data is fragmented across silos of direct-attached storage (DAS), data mobility and management have become complex problems. This is where NetApp can help.

1.2 Running ONTAP as Software

There is a compelling value proposition in allowing customers to determine the physical characteristics of their underlying hardware while still consuming ONTAP and all its storage management services. Decoupling ONTAP from the underlying hardware allows us to provide enterprise-class file and replication services within an SDS environment.

Still, one question remains: why do we require a hypervisor?

Running ONTAP as software on top of another software application allows us to leverage much of the qualification work done by the hypervisor. This capability is critical for helping us rapidly expand our list of supported platforms. Also, positioning ONTAP as a virtual machine (VM) allows customers to plug into existing management and orchestration frameworks, which allows rapid provisioning and end-to-end automation from deployment to sunsetting.

This is the goal of ONTAP Select.

1.3 ONTAP Select Versus ONTAP Edge

This section describes differences between ONTAP Select and ONTAP Edge. Although many of the differences are covered in detail in the section "Architecture Overview," Table 1 highlights some of the major differences between the two products.

Table 1) ONTAP Select versus ONTAP Edge.

Description	ONTAP Select	ONTAP Edge
Node count	Single-node, two-node HA, four-node, six-node, and eight-node HA	Single node
VM CPU/memory	4 vCPUs / 16GB (small instance) 8 vCPUs / 64GB (medium instance) 16 vCPUs / 128GB (large instance) starting with ONTAP Select 9.6 and ONTAP Deploy 2.12	2 vCPUs/8GB
Hypervisor	Check the NetApp Interoperability Matrix Tool (IMT) for the latest supported versions	vSphere 5.1, 5.5
High availability (HA)	Yes	No
iSCSI/CIFS/NFS	Yes	Yes
NetApp SnapMirror® and NetApp SnapVault® technologies	Yes	Yes
Full suite of storage efficiency policies	Yes	No
NetApp Volume Encryption	Yes	No
Data retention and compliance (NetApp SnapLock® Enterprise)	Yes	No
Capacity limit	Up to 400TB per ONTAP Select node starting with ONTAP Select 9.5 and ONTAP Deploy 2.10	Up to 10TB, 25TB, or 50TB

Hardware platform support	Wider support for major vendor offerings that meet minimum criteria	Select families within qualified server vendors

1.4 ONTAP Select Small Versus ONTAP Select Medium Versus ONTAP Select Large

ONTAP Select can be deployed in three sizes: a small VM, a medium VM, and a large VM. A Premium XL license can be used with all the VM sizes. A Premium license can be used with either a small instance or a medium instance, while the Standard license can only be used with a small instance. The difference between the various VM sizes consists of the amount of resources reserved for each instance of ONTAP Select. For example, the large VM consumes 16 cores and 128GB of RAM, the medium VM consumes eight CPU cores and 64GB of RAM, while the small VM consumes four cores and 16GB of RAM. More information can be found in the section "VM Properties."

The number of cores and the amount of memory per ONTAP Select VM cannot be further modified. In addition, a Premium or a Premium XL license is required when using solid-state drives (SSDs) for DAS configurations (hardware RAID controller or ONTAP software RAID) or for any NetApp MetroCluster™ SDS constructs.

In a four-node cluster, it is possible to have a two-node medium HA system and a two-node small HA system. Within an HA pair, however, the ONTAP Select VM type should be identical.

Note that, starting with ONTAP Select 9.6 and ONTAP Deploy 2.12, it is possible to convert from a Standard license to a Premium or Premium XL license, but not the other way around. See section 7.4 for additional information.

Note that, in the current release of the product, the ONTAP Select large VM (requiring a Premium XL license) is only supported for software RAID installation.

1.5 ONTAP Select Evaluation Software Versus Running ONTAP Select in Evaluation Mode

The ONTAP Select version available on the web portal (Downloads/Software) is a full version of the product that can be run in evaluation mode. This means that the client can test the full solution, including ONTAP Deploy, which is the ONTAP Select setup product. ONTAP Deploy checks and enforces all minimum requirements for ONTAP Select, which is useful for both documenting the procedure and vetting your environment for suitability.

However, at times the test environment does not match the production environment or does not meet the minimum requirements enforced by ONTAP Deploy. For a quick test of only ONTAP Select, we are providing an Open Virtualization Format (OVF) download of only ONTAP Select (Downloads/Product Evaluation). When using this OVF download, the ONTAP Deploy utility is not used. Instead, you directly install a time-limited single-node ONTAP Select cluster, just like the single-node cluster created using the Deploy tool in evaluation mode. The main benefit of the OVF setup is that it lowers the requirements for testing ONTAP Select.

Note that, once an evaluation trial has expired, the evaluation software cannot be extended. Starting with ONTAP Select 9.4, the expired trial functionality is severely limited as follows:

- **Single-node cluster.** No new aggregates can be created, and, after the first reboot, the aggregates do not come online. Data is inaccessible.
- **Nodes in an HA pair.** No new aggregates can be created, and, after the first reboot, only the remote aggregates are available. Remote aggregates are not normally hosted by the node that is available.

Starting with ONTAP Deploy 2.11 and ONTAP Select 9.5, evaluation clusters created with ONTAP Deploy can be converted to production by applying the appropriate capacity tier license, as long as the evaluation period has not yet expired. Moving from an evaluation to a capacity pool license is not yet supported. Furthermore, evaluation builds are no longer capped at 2TB. When converting from an evaluation to a production instance, the capacity tier license must cover the entire amount of storage used during evaluation.

1.6 ONTAP Select Platform and Feature Support

The abstraction layer provided by the hypervisor allows ONTAP Select to run on a wide variety of commodity platforms from virtually all the major server vendors, providing they meet minimum hardware criteria. These specifications are detailed in the following sections.

Hardware Requirements

ONTAP Select Standard VM requires that the hosting physical server meets the following minimum requirements:

- Intel Xeon E5-26xx v3 (Haswell) CPU or greater
- Intel CPU Skylake Server processors (see this link) are supported starting with ONTAP Select 9.3 and ONTAP Deploy 2.7.2
- 6 x vCPUs (4 for ONTAP Select; 2 for the hypervisor)
- 24GB RAM (16GB for ONTAP Select; 8GB for the OS)
- Starting with ONTAP Select 9.3, some configurations with a single 10Gb port are now qualified and supported. For prior ONTAP Select versions, the minimum requirements are still as follows:
  - Minimum of 2 x 1Gb network interface card (NIC) ports for single-node clusters
  - Minimum of 4 x 1Gb NIC ports for two-node clusters
  - 2 x 10GbE NIC ports (four recommended) for four-node clusters

Note: The ONTAP Select medium VM reserves 8 cores and 64GB of RAM, and the ONTAP Select large VM reserves 16 cores and 128GB of RAM. Therefore, the minimum server requirements should be adjusted accordingly.

For locally attached storage (DAS), the following requirements also apply:

The requirements for deploying on DAS with a hardware RAID controller are as follows:

- Hardware RAID controller with 512MB writeback (battery backed-up) cache and 12Gbps of throughput
- A total of up to 60 drives or 400TB per node can be supported starting with the minimum versions of ONTAP Deploy 2.7 and ONTAP Select 9.3. To support a large drive count, an external shelf or drive enclosure can be used. It is important to make sure that the hardware RAID controller can support that number of drives and total capacity.
- For prior versions of ONTAP Select, the limits on the number of drives are as follows:
  - 8 to 24 internal disks (SAS, NL-SAS, or SATA)
  - 4 to 24 SSDs (Premium license required)

The requirements for deploying on DAS when using ONTAP software RAID are as follows:

- ONTAP Select 9.5 and Deploy 2.10 or newer
- A total of up to 60 drives or 400TB per node can be supported. To support a large drive count, an external shelf or drive enclosure can be used.
- The minimum number of drives required for using ONTAP software RAID depends on the configuration. See Table 2.

Table 2) ONTAP Software RAID minimum number of drives.

Cluster Size	RAID Type	Minimum Drives Required	Layout (Disk Types
Single node	RAID-4	4*	1 Service**
			1 Parity***
			2 Data
	NetApp RAID-DP® technology	6*	1 Service
			2 Parity
			3 Data
	NetApp RAID-TEC™ technology	8*	1 Service
			3 Parity
			4 Data
Multimode (each node)	RAID-4	7*	1 Service
			2x1 Parity
			2x2 Data
	RAID-DP	11*	1 Service
			2x2 Parity
			2x3 Data
	RAID-TEC	15*	1 Service
			2x3 Parity
			2x4 Data

*A spare disk is optional but recommended. To include a Spare disk, add one above. The Spare disk does not count towards the license.

**A service (or system) disk does not count towards the license. This service (system) disk must be virtualized. In other words, a VMFS datastore must exist on this drive.

***A parity disk is not counted towards the license.

For shared storage (virtual SAN [VSAN] and some HCI appliances or external arrays), using a RAID controller is no longer a requirement. However, the following restrictions and best practices should be considered when selecting the type of datastore used for hosting ONTAP Select:

- Support for VSAN and external arrays requires the following minimum versions: ONTAP Select 9.1 and Deploy 2.3.
- Support for VMware HA, vMotion, and Distributed Resource Scheduler (DRS) requires the following minimum versions: ONTAP Select 9.2 and Deploy 2.4.
- Multinode clusters on shared storage are supported starting with ONTAP Deploy 2.8 and ONTAP Select 9.4. For prior releases, only single-node clusters are supported with VSAN or external array-type datastores.
- The VSAN configuration or the external array must be supported by VMware as evidenced by the configuration present on the VMware hardware compatibility list (HCL).

ONTAP Select Feature Support

ONTAP Select offers full support for most ONTAP functionality except for those features that have hardware-specific dependencies.

Supported functionality includes the following:

- NFS, CIFS, and iSCSI
- SnapMirror and SnapVault
- NetApp FlexClone® technology
- NetApp SnapRestore® technology
- NetApp Volume Encryption
- SnapLock Enterprise (separate license)
- FabricPool (separate license)
- NetApp FlexCache® technology (separate license at zero cost)
- NetApp SyncMirror® technology (separate license at zero cost)
- NetApp Data Availability Services (separate license)
- MetroCluster SDS (formerly called an ONTAP Select two-node stretched cluster; ONTAP Select Premium license)

In addition, support for the NetApp OnCommand® management suite is included. This suite includes most tooling used to manage NetApp FAS arrays, such as OnCommand Unified Manager, OnCommand Insight, OnCommand Workflow Automation, and NetApp SnapCenter®. Using SnapCenter, NetApp SnapManager®, or NetApp SnapDrive® with ONTAP Select requires server-based licenses.

Consult the [IMT](#) for a complete list of supported management applications.

The following ONTAP features are not supported by ONTAP Select:

- Interface groups (ifgroups)
- Service Processor
- Hardware-centric features such as the traditional FAS/AFF MetroCluster architecture that requires dedicated hardware infrastructure between sites, Fibre Channel (FC/FCoE), and full disk encryption (FDE)
- NetApp Storage Encryption drives

ONTAP Select Storage Efficiency Support

ONTAP Select provides storage efficiency options that are similar to the storage efficiency options present on FAS and AFF arrays.

ONTAP Select virtual NAS (vNAS) deployments using all-flash VSAN or generic flash arrays should follow the best practices for ONTAP Select with non-SSD DAS storage.

In ONTAP Select 9.5, an AFF-like personality is automatically enabled on new installations as long as the following conditions are met: DAS storage with SSD drives and a Premium license.

With an AFF-like personality, the following inline SE features are automatically enabled during installation:

- Inline zero pattern detection
- Volume inline deduplication
- Volume background deduplication
- Adaptive inline compression
- Inline data compaction
- Aggregate inline deduplication

- Aggregate background deduplication

To verify that ONTAP Select has enabled all the default storage efficiency policies, run the following command on a newly created volume:

```
<system name>::> set diag
Warning: These diagnostic commands are for use by NetApp personnel only.
Do you want to continue? {y|n}: y
twonode95IP15::*> sis config
Vserver:                                      SVM1
Volume:                                       _export1_NFS_volume
Schedule:                                     -
Policy:                                       auto
Compression:                                  true
Inline Compression:                           true
Compression Type:                             adaptive
Application IO Size:                          8K
Compression Algorithm:                        lzopro
Inline Dedupe:                                true
Data Compaction:                              true
Cross Volume Inline Deduplication:            true
Cross Volume Background Deduplication:        true
```

Note: For ONTAP Select upgrades from 9.4 to 9.5, you must install ONTAP Select 9.4 on DAS SSD storage with a Premium license. In addition, the Enable Storage Efficiencies check box must be checked during initial cluster installation with ONTAP Deploy. Enabling an AFF-like personality post-ONTAP upgrade when prior conditions have not been met requires the manual creation of a boot argument and a node reboot. Contact technical support for further details.

Table 3 summarizes the various storage efficiency options available, enabled by default, or not enabled by default but recommended, depending on the ONTAP Select version and media type.

Table 3) ONTAP Select storage efficiency configurations.

ONTAP Select /Features	9.6 / 9.5 Premium or Premium XL[4] (DAS SSD)	9.4[1] / 9.3[2] Premium (DAS SSD)	9.6 / 9.5 / 9.4[1] / 9.3[2] All Licenses (DAS HDD)	9.6 All Licenses (vNAS)	9.5 / 9.4[1] / 9.3[2] Premium or Standard (vNAS)[3]
Inline zero detection	Yes (default)	Yes Enabled by user on a per-volume basis	Yes Enabled by user on a per-volume basis	Yes Enabled by user on a per-volume basis	Not supported
Volume inline deduplication	Yes (default)	Yes (recommended) Enabled by user on a per-volume basis	Not available	Not supported	Not supported
32K inline compression (secondary compression)	Yes Enabled by user on a per volume basis.	Yes Enabled by user on a per-volume basis	Yes Enabled by user on a per-volume basis	Not supported	Not supported
8K inline compression (adaptive compression)	Yes (default)	Yes (recommended) Enabled by user on a per-volume basis	Yes Enabled by user on a per volume basis	Not supported	Not supported

ONTAP Select /Features	9.6 / 9.5 Premium or Premium XL[4] (DAS SSD)	9.4[1] / 9.3[2] Premium (DAS SSD)	9.6 / 9.5 / 9.4[1] / 9.3[2] All Licenses (DAS HDD)	9.6 All Licenses (vNAS)	9.5 / 9.4[1] / 9.3[2] Premium or Standard (vNAS)[3]
Background compression	Not supported	Not supported	Yes Enabled by user on a per volume basis	Yes Enabled by user on a per-volume basis	Not supported
Compression scanner	Yes	Yes Enabled by user on a per-volume basis	Yes	Yes Enabled by user on a per-volume basis	Not supported
Inline data compaction	Yes (default)	Yes (recommended) Enabled by user on a per-volume basis	Yes Enabled by user on a per volume basis	Not supported	Not supported
Compaction scanner	Yes	Yes Enabled by user on a per-volume basis	Yes	Not supported	Not supported
Aggregate inline deduplication	Yes (default)	Yes (recommended) Enabled by user on a per volume basis with space guarantee = none)	N/A	Not supported	Not supported
Volume background deduplication	Yes (default);	Yes (recommended)	Yes Enabled by user on a per volume basis	Yes Enabled by user on a per-volume basis	Not supported
Aggregate background deduplication	Yes(default)	Yes (recommended) Enabled by user on a per volume basis with space guarantee = none)	N/A	Not supported	Not supported

[1]ONTAP Select 9.4 on DAS SSDs (requires Premium license) allows existing data in an aggregate to be deduped using aggregate-level background cross volume scanners. This one-time operation is performed manually for volumes created before 9.4.

[2]ONTAP Select 9.3 on DAS SSDs (requires Premium license) supports aggregate-level background deduplication; however, this feature must be enabled after creating the aggregate.

[3]ONTAP Select 9.5 vNAS by default does not support any storage efficiency policies. Review the vNAS section for details on Single Instance Data Logging (SIDL).

[4]ONTAP Select 9.6 supports a new license (Premium XL) and a new VM size (large). However, the large VM is only supported for DAS configurations using software RAID. Hardware RAID and vNAS configurations are not supported with the large ONTAP Select VM in the current release.

Notes on Upgrade Behavior for DAS SSD Configurations

After upgrading to ONTAP Select 9.5, wait for the `system node upgrade-revert show` command to indicate that the upgrade has completed before verifying the storage efficiency values for existing volumes.

On a system upgraded to ONTAP Select 9.5, a new volume created on an existing aggregate or a newly created aggregate has the same behavior as a volume created on a fresh deployment on ONTAP Select 9.5. Existing volumes that undergo the ONTAP Select code upgrade have most of the same storage efficiency policies as a newly created volume on ONTAP Select 9.5 with some variations:

Scenario 1. If no storage efficiency policies were enabled on a volume prior to the upgrade, then:

- Volumes with `space guarantee = volume` do not have inline data-compaction, aggregate inline deduplication, and aggregate background deduplication enabled. These options can be enabled post-upgrade.
- Volumes with `space guarantee = none` do not have background compression enabled. This option can be enabled post upgrade.
- Storage efficiency policy on the existing volumes is set to auto after upgrade.

Scenario 2. If some storage efficiencies are already enabled on a volume prior to the upgrade, then:

- Volumes with `space guarantee = volume` do not see any difference after upgrade.
- Volumes with `space guarantee = none` have aggregate background deduplication turned on.
- Volumes with `storage policy inline-only` have their policy set to auto.
- Volumes with user defined storage efficiency policies have no change in policy, with the exception of volumes with `space guarantee = none`. These volumes have aggregate background deduplication enabled.

Notes on Upgrade Behavior for DAS HDD Configuration

Storage efficiency features enabled prior to the upgrade are retained after the upgrade to ONTAP Select 9.5. If no storage efficiencies were enabled prior to the upgrade, no storage efficiencies are enabled post-upgrade.

2 Architecture Overview

ONTAP Select is ONTAP deployed as a VM. It provides storage management services on a virtualized commodity server.

ONTAP Select can be deployed two ways:

- **Non-HA (single node).** The single-node version of ONTAP Select is well suited for storage infrastructures that provide their own storage resiliency. These include VSAN datastores or external arrays that offer data protection at the array layer and work along with VMware HA. A single-node Select cluster can also be used for remote and branch offices in which the data is protected by replication to a core location.
- **HA (multinode).** The multinode version of ONTAP Select uses two, four, six, or eight ONTAP Select nodes and adds support for HA and Data ONTAP nondisruptive operations, all within a shared-nothing environment.

When choosing a solution, resiliency requirements, environment restrictions, and cost factors should be considered. Although both versions run ONTAP and support many of the same core features, the multinode solution provides HA and supports nondisruptive operations, a core value proposition for ONTAP.

Note: The single-node and multinode implementations of ONTAP Select are deployment options, not separate products. Although the multinode solution requires the purchase of additional node licenses, both share the same product model, FDvM300.

This section provides a detailed analysis of the various aspects of the system architecture for both the single-node and multinode solutions while highlighting important differences between the two variants.

2.1 VM Properties

The ONTAP Select VM has a fixed set of properties, which are described in Table 4. Increasing or decreasing the amount of resources allocated to the VM is not supported. Additionally, the ONTAP Select instance hard reserves the CPU and memory resources, meaning that the physical resources backed by the VM are unavailable to any other VMs hosted on the server.

Table 4) ONTAP Select VM properties.

Description	Single Node	Multinode (per Node)
CPU/memory	4 vCPUs / 16GB RAM or 8 vCPUs / 64GB RAM or 16 vCPUs / 128GB RAM[1]	4 vCPUs / 16GB RAM or 8 vCPUs / 64GB RAM or 16 vCPUs / 128GB RAM[1]
Virtual network interfaces	3 (2 for ONTAP Select versions before 9.3)	7 (6 for ONTAP Select versions before 9.3)
SCSI controllers	4	4
System boot disk	10GB	10GB
System core dump disk	120GB for small and medium VMs 240GB for large VMs	120GB for small and medium VMs 240GB for large VMs
Mailbox disk	N/A	556MB
Cluster root disk[2]	68GB	68GB x 2
NVRAM partition	4GB (ONTAP Select 9.5 on ESX 6.5 U2 and higher only)	4GB (ONTAP Select 9.5 on ESX 6.5 U2 and higher only)

[1] ONTAP Select large (version 9.6 and later).

[2] The root aggregate no longer counts against the capacity license starting with ONTAP Select 9.4.

Note: The NVRAM partition was separated as its own disk starting with ONTAP Select 9.5 installed on ESX 6.5 U2 and later. ONTAP Select 9.5 installed on older versions of ESX, or ONTAP Select 9.5 installations that were upgraded from older versions and all prior versions of ONTAP Select, collocated the NVRAM partition on the boot disk.

Starting with ONTAP Select 9.2, the ONTAP console is accessible through the VM video console tab in the vSphere client.

Note: The serial ports were removed from the ONTAP Select 9.2 VM, which allows ONTAP Select 9.2 to support and install on any vSphere license. Before ONTAP Select 9.2, only the vSphere Enterprise/Enterprise+ licenses were supported.

Table 5 lists the differences between the ONTAP Select releases 9.3 through 9.6. There are no differences in the properties of ONTAP Select 9.3 and 9.4.

Table 5) ONTAP Select release comparison.

Description	ONTAP Select 9.3/9.4	ONTAP Select 9.5	ONTAP Select 9.6
ONTAP Select license	Standard or Premium	Standard or Premium	Standard or Premium or Premium XL[4]
CPU/memory	4 vCPUs / 16GB or 8 vCPUs / 64GB	4 vCPUs / 16GB or 8 vCPUs / 64GB	4 vCPUs / 16GB or 8 vCPUs / 64GB or 16 vCPU / 128GB
Disk type	SAS, NL-SAS, SATA, or SSD[3]	SAS, NL-SAS, SATA, or SSD[3]	SAS, NL-SAS, SATA, or SSD[3]
Minimum number of disks (with hardware RAID controller)	8 SAS, NL-SAS, or SATA, or 4 SSD[3]	8 SAS, NL-SAS, or SATA, or 4 SSD[3]	8 SAS, NL-SAS, or SATA, or 4 SSD[3]
Minimum number of disks (with ONTAP software RAID)	N/A	4 SSD[3] drives (single node with RAID 4 and no parity) 7 SSD[3] drives (multi-node with RAID 4 and no parity)	4 SSD[3] drives (single node with RAID 4 and no parity) 7 SSD[3] drives (multi-node with RAID 4 and no parity)
Maximum number of disks	60	60	60
vSphere license requirements	All vSphere licenses are supported[2]	All vSphere licenses are supported[2]	All vSphere licenses are supported[2]
VMware HA	vNAS only	vNAS only	vNAS only
VMware storage vMotion	Yes[1]	Yes[1]	Yes[1]
Cluster size	Single node Two node Four node Six node Eight node	Single node Two node Four node Six node Eight node	Single node Two node Four node Six node Eight node
Maximum capacity per node	400TB	400TB	400TB

[1]Requires ONTAP Deploy 2.7 and ONTAP Select 9.3

[2]The ESXi free license is not supported.

[3]The Premium license is required for all SSDs.

[4]The large ONTAP Select VM (requiring the Premium XL license) is only supported with software RAID configurations in the current release.

When using locally attached storage (DAS), certain restrictions apply to the ONTAP Select VM, specifically:

- Only one ONTAP Select VM can reside on a single server.

- vSphere fault tolerance (FT) is not supported.

2.2 Hardware RAID Services for Local Attached Storage

If a hardware RAID controller is present, then ONTAP Select can take advantage of it and its local cache to achieve both a write performance boost and the added benefit of protection against physical drive failures. It does this by moving RAID services to the hardware controller. As a result, RAID protection for all nodes within the ONTAP Select cluster is provided by the locally attached RAID controller and not through ONTAP software RAID.

Note: ONTAP Select data aggregates are configured to use RAID 0 because the physical RAID controller is providing RAID striping to the underlying drives. No other RAID levels are supported.

RAID Controller Configuration for Local Attached Storage

All locally attached disks that provide ONTAP Select with backing storage must sit behind a RAID controller. Most commodity servers come with multiple RAID controller options across multiple price points, each with varying levels of functionality. The intent is to support as many of these options as possible, providing they meet certain minimum requirements placed on the controller.

The RAID controller that manages the ONTAP Select disks must meet the following requirements:

- The hardware RAID controller must have a battery backup unit (BBU) or flash-backed write cache (FBWC) and support 12Gbps of throughput.
- The RAID controller must support a mode that can withstand at least one or two disk failures (RAID 5 and RAID 6).
- The drive cache must be set to disabled.
- The write policy must be configured for writeback mode with a fallback to write through upon BBU or flash failure.
- The I/O policy for reads must be set to cached.

All locally attached disks that provide ONTAP Select with backing storage must be placed into RAID groups running RAID 5 or RAID 6. For SAS drives and SSDs, using RAID groups of up to 24 drives allows ONTAP to reap the benefits of spreading incoming read requests across a higher number of disks. Doing so provides a significant gain in performance. With SAS/SSD configurations, performance testing was performed against single-LUN versus multi-LUN configurations. No significant differences were found, so, for simplicity's sake, NetApp recommends creating the fewest number of LUNs necessary to support your configuration needs.

NL-SAS and SATA drives require a different set of best practices. For performance reasons, the minimum number of disks is still eight, but the RAID group size should not be larger than 12 drives. NetApp also recommends using one spare per RAID group; however, global spares for all RAID groups can also be used. For example, you can use two spares for every three RAID groups, with each RAID group consisting of eight to 12 drives.

Note: The maximum extent and datastore size for older ESX releases is 64TB, which can affect the number of LUNs necessary to support the total raw capacity provided by these large capacity drives.

RAID Mode

Many RAID controllers support up to three modes of operation, each representing a significant difference in the data path taken by write requests. These three modes are as follows:

- **Writethrough.** All incoming I/O requests are written to the RAID controller cache and then immediately flushed to disk before acknowledging the request back to the host.

- **Writearound.** All incoming I/O requests are written directly to disk, circumventing the RAID controller cache.
- **Writeback.** All incoming I/O requests are written directly to the controller cache and immediately acknowledged back to the host. Data blocks are flushed to disk asynchronously using the controller.

Writeback mode offers the shortest data path, with I/O acknowledgment occurring immediately after the blocks enter cache. This mode provides the lowest latency and highest throughput for mixed read/write workloads. However, without the presence of a BBU or nonvolatile flash technology, users run the risk of losing data if the system incurs a power failure when operating in this mode.

ONTAP Select requires the presence of a battery backup or flash unit; therefore, we can be confident that cached blocks are flushed to disk in the event of this type of failure. For this reason, it is a requirement that the RAID controller be configured in writeback mode.

Best Practice
The server RAID controller should be configured to operate in writeback mode. If write workload performance issues are seen, check the controller settings and make sure that writethrough or writearound is not enabled.

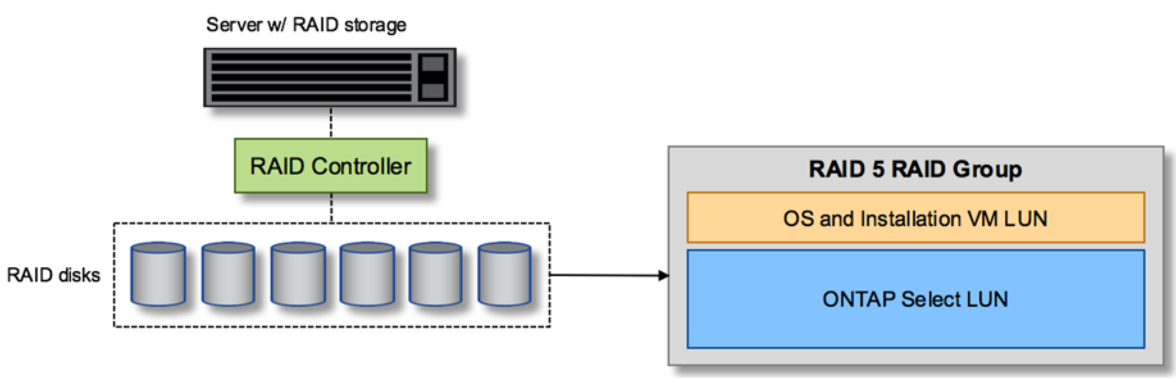
Local Disks Shared Between ONTAP Select and OS

The most common server configuration is one in which all locally attached spindles sit behind a single RAID controller. You should provision a minimum of two LUNs: one for the hypervisor and one for the ONTAP Select VM.

For example, consider an HP DL380 g8 with six internal drives and a single Smart Array P420i RAID controller. All internal drives are managed by this RAID controller, and no other storage is present on the system.

Figure 1 shows this style of configuration. In this example, no other storage is present on the system; therefore, the hypervisor must share storage with the ONTAP Select node.

Figure 1) Server LUN configuration with only RAID-managed spindles.



Provisioning the OS LUNs from the same RAID group as ONTAP Select allows the hypervisor OS (and any client VM that is also provisioned from that storage) to benefit from RAID protection. This configuration prevents a single-drive failure from bringing down the entire system.

Best Practice
If the physical server contains a single RAID controller managing all locally attached disks, NetApp recommends creating a separate LUN for the server OS and one or more LUNs for ONTAP Select. In the event of boot disk corruption, this best practice allows the administrator to recreate the OS LUN without affecting ONTAP Select.

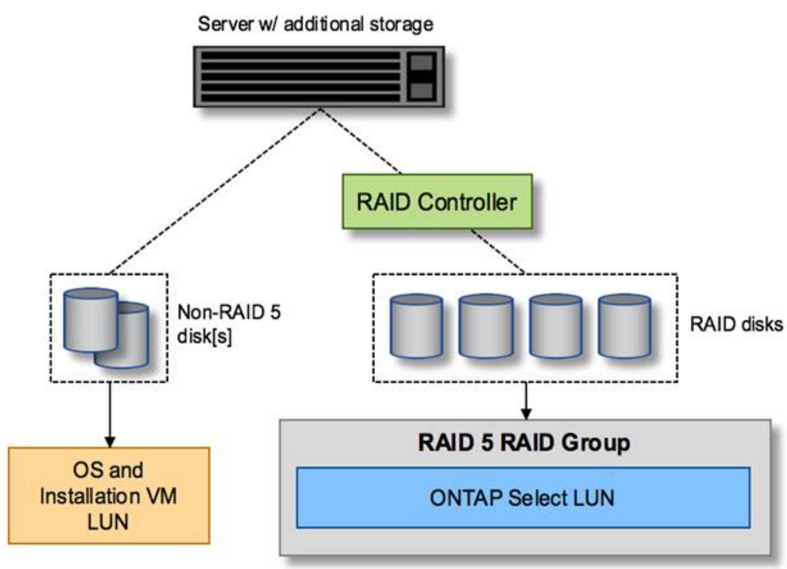
Local Disks Split Between ONTAP Select and OS

The other possible configuration provided by server vendors involves configuring the system with multiple RAID or disk controllers. In this configuration, a set of disks is managed by one disk controller, which might or might not offer RAID services. A second set of disks is managed by a hardware RAID controller that is able to offer RAID 5/6 services.

With this style of configuration, the set of spindles that sits behind the RAID controller that can provide RAID 5/6 services should be used exclusively by the ONTAP Select VM. Depending on the total storage capacity under management, you should configure the disk spindles into one or more RAID groups and one or more LUNs. These LUNs would then be used to create one or more datastores, with all datastores being protected by the RAID controller.

The first set of disks is reserved for the hypervisor OS and any client VM that is not using ONTAP storage, as shown in Figure 2.

Figure 2) Server LUN configuration on mixed RAID/non-RAID system.



Multiple LUNs

There are two cases for which single-RAID group/single-LUN configurations must change. When using NL-SAS or SATA drives, the RAID group size must not exceed 12 drives. In addition, a single LUN can become larger than the underlying hypervisor storage limits either individual file system extent maximum size or total storage pool maximum size. Then the underlying physical storage must be broken up into multiple LUNs to enable successful file system creation.

Best Practice
ONTAP Select receives no performance benefits by increasing the number of LUNs within a RAID group. Multiple LUNs should only be used to follow best practices for SATA/NL-SAS configurations or to bypass hypervisor file system limitations.

VMware vSphere Virtual Machine File System Limits

The maximum size of a datastore on some versions of ESX is 64TB.

If a server has more than 64TB of storage attached, multiple LUNs might need to be provisioned, each smaller than 64TB. Creating multiple RAID groups to improve the RAID rebuild time for SATA/NL-SAS drives also results in multiple LUNs being provisioned.

When multiple LUNs are required, a major point of consideration is making sure that these LUNs have similar and consistent performance. This is especially important if all the LUNs are to be used in a single ONTAP aggregate. Alternatively, if a subset of one or more LUNs has a distinctly different performance profile, we strongly recommend isolating these LUNs in a separate ONTAP aggregate.

Multiple file system extents can be used to create a single datastore up to the maximum size of the datastore. To restrict the amount of capacity that requires an ONTAP Select license, make sure to specify a capacity cap during the cluster installation. This functionality allows ONTAP Select to use (and therefore require a license for) only a subset of the space in a datastore.

Alternatively, one can start by creating a single datastore on a single LUN. When additional space requiring a larger ONTAP Select capacity license is needed, then that space can be added to the same datastore as an extent, up to the maximum size of the datastore. After the maximum size is reached, new datastores can be created and added to ONTAP Select. Both types of capacity extension operations are supported and can be achieved by using the ONTAP Deploy storage-add functionality. Each ONTAP Select node can be configured to support up to 400TB of storage. Provisioning capacity from multiple datastores requires a two-step process.

The initial cluster create can be used to create an ONTAP Select cluster consuming part or all of the space in the initial datastore. A second step is to perform one or more capacity addition operations using additional datastores until the desired total capacity is reached. This functionality is detailed in the section “Increasing the ONTAP Select Capacity Using ONTAP Deploy.”

Note: VMFS overhead is nonzero (see VMware KB 1001618), and attempting to use the entire space reported as free by a datastore has resulted in spurious errors during cluster create operations.

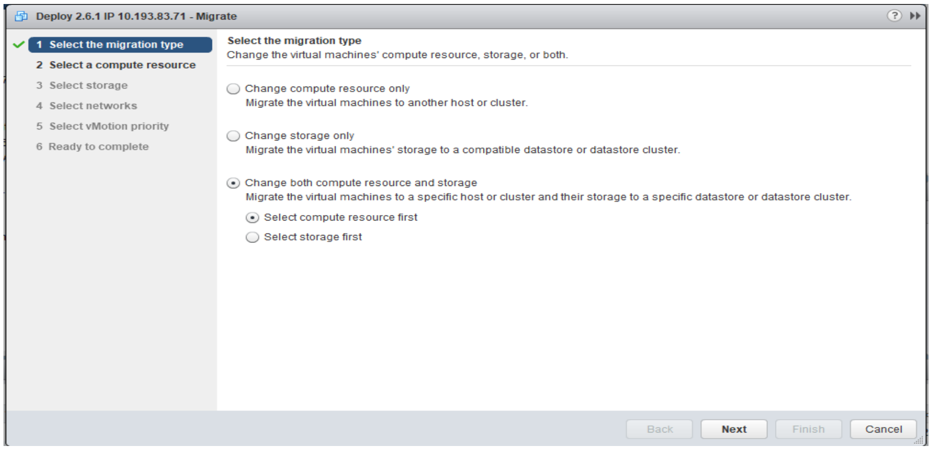
Starting with ONTAP Deploy 2.7, a 2% buffer is left unused in each datastore. This space does not require a capacity license because it is not used by ONTAP Select. ONTAP Deploy automatically calculates the exact number of gigabytes for the buffer, as long as a capacity cap is not specified. If a capacity cap is specified, that size is enforced first. If the capacity cap size falls within the buffer size, the cluster create fails with an error message specifying the correct maximum size parameter that can be used as a capacity cap:

```
“InvalidPoolCapacitySize: Invalid capacity specified for storage pool “ontap-select-storage-pool”, Specified value: 34334204 GB. Available (after leaving 2% overhead space): 30948”
```

Starting with ONTAP Select 9.3 and ONTAP Deploy 2.7, VMFS 6 is supported for both new installations and as the target of a Storage vMotion operation of an existing ONTAP Deploy or ONTAP Select VM.

VMware does not support in-place upgrades from VMFS 5 to VMFS 6. Therefore, Storage vMotion is the only mechanism that allows any VM to transition from a VMFS 5 datastore to a VMFS 6 datastore. However, support for Storage vMotion with ONTAP Select and ONTAP Deploy was expanded to cover other scenarios besides the specific purpose of transitioning from VMFS 5 to VMFS 6.

For ONTAP Select VMs, support for Storage vMotion includes both single-node and multinode clusters and includes both storage-only and compute and storage migrations.



At the end of a Storage vMotion operation, ONTAP Deploy should be used to trigger a cluster refresh operation. The purpose of this operation is to update the ONTAP Deploy database of the ONTAP Select node's new location.

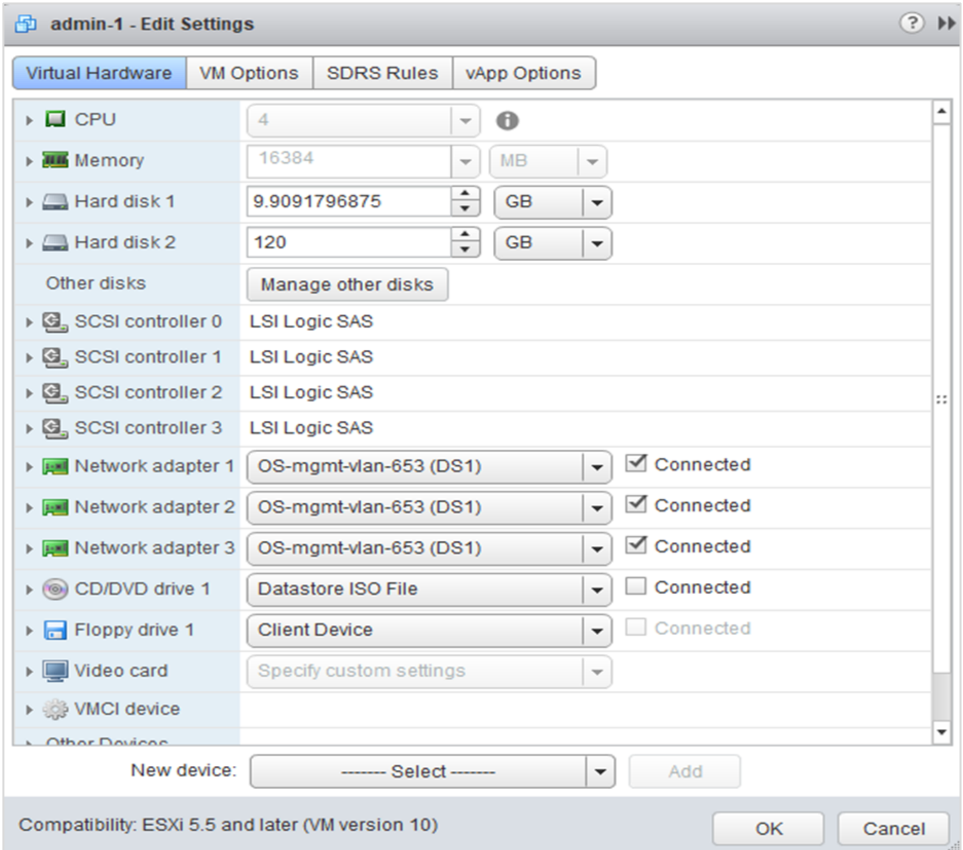
Note: Although support for Storage vMotion provides a lot of flexibility, it is important that the new host can appropriately support the ONTAP Select node. If a RAID controller and DAS storage were used on the original host, a similar setup should exist on the new host. Severe performance issues can result if the ONTAP Select VM is rehosted on an unsuitable environment.

Best Practice
Available capacity on a new host is not the only factor when deciding whether to use VMware Storage vMotion with an ONTAP Select node. The underlying storage type, host configuration, and network capabilities should be able to sustain the same workload as the original host.

When using Storage vMotion, complete the following procedure:

1. Shut down the ONTAP Select VM. If this node is part of an HA pair, perform a storage failover first.
2. Clear the CD/DVD drive option.

   Note: This step does not apply if you installed ONTAP Select without using ONTAP Deploy.



3. After the Storage vMotion operation completes, power on the ONTAP Select VM.

   Note: If this node is part of an HA pair, you can perform a manual giveback.

4. Issue a cluster refresh operation with ONTAP Deploy and make sure that it is successful.
5. Back up the ONTAP Deploy database.

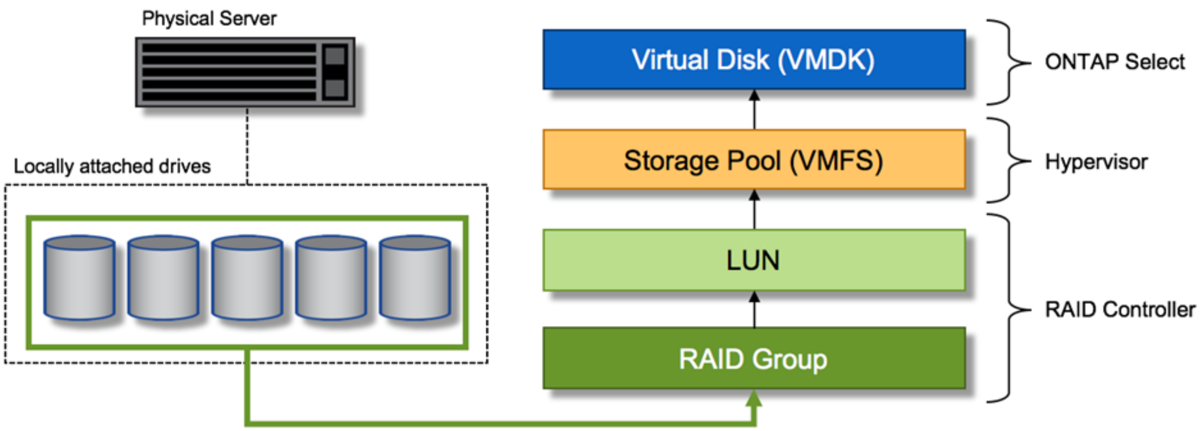
ONTAP Select Virtual Disks

At its core, ONTAP Select presents ONTAP with a set of virtual disks provisioned from one or more storage pools. ONTAP is presented with a set of virtual disks that it treats as physical, and the remaining portion of the storage stack is abstracted by the hypervisor. Figure 3 shows this relationship in more detail, highlighting the relationship between the physical RAID controller, the hypervisor, and the ONTAP Select VM.

Notes:

- RAID group and LUN configuration occur from within the server's RAID controller software. This configuration is not required when using VSAN or external arrays.
- Storage pool configuration occurs from within the hypervisor.
- Virtual disks are created and owned by individual VMs; in this example, by ONTAP Select.

Figure 3) Virtual disk to physical disk mapping.



Virtual Disk Provisioning

To provide for a more streamlined user experience, the ONTAP Select management tool, ONTAP Deploy, automatically provisions virtual disks from the associated storage pool and attaches them to the ONTAP Select VM. This operation occurs automatically during both initial setup and during storage-add operations. If the ONTAP Select node is part of an HA pair, the virtual disks are automatically assigned to a local and mirror storage pool.

ONTAP Select breaks up the underlying attached storage into equal-sized virtual disks, each not exceeding 16TB. If the ONTAP Select node is part of an HA pair, a minimum of two virtual disks are created on each cluster node and assigned to the local and mirror plex to be used within a mirrored aggregate.

For example, an ONTAP Select can assigned a datastore or LUN that is 31TB (the space remaining after the VM is deployed and the system and root disks are provisioned). Then four ~7.75TB virtual disks are created and assigned to the appropriate ONTAP local and mirror plex.

Note: Adding capacity to an ONTAP Select VM likely results in VMDKs of different sizes. For details, see the section “Increasing the ONTAP Select Capacity Using ONTAP Deploy.” Unlike FAS systems, different sized VMDKs can exist in the same aggregate. ONTAP Select uses a RAID 0 stripe across these VMDKs, which results in the ability to fully use all the space in each VMDK regardless of its size.

Virtualized NVRAM

NetApp FAS systems are traditionally fitted with a physical NVRAM PCI card, a high-performing card containing nonvolatile flash memory. This card provides a significant boost in write performance by granting ONTAP with the ability to immediately acknowledge incoming writes back to the client. It can also schedule the movement of modified data blocks back to the slower storage media in a process known as destaging.

Commodity systems are not typically fitted with this type of equipment. Therefore, the functionality of this NVRAM card has been virtualized and placed into a partition on the ONTAP Select system boot disk. It is for this reason that placement of the system virtual disk of the instance is extremely important. This is also why the product requires the presence of a physical RAID controller with a resilient cache for local attached storage configurations.

Starting with new installations of ONTAP Select 9.5 on ESXi 6.5 U2 and newer, NVRAM is placed on its own VMDK. Prior versions of ONTAP Select, instances that are upgraded to version 9.5, or instances of

ONTAP Select 9.5 on older versions of ESXi use the 9.9GB boot disk for their NVRAM partitions. Splitting the NVRAM in its own VMDK allows the ONTAP Select VM to use the vNVMe driver to communicate with its NVRAM VMDK. It also requires that the ONTAP Select VM uses hardware version 13, which is compatible with ESX 6.5 and newer.

Data Path Explained: NVRAM and RAID Controller

The interaction between the virtualized NVRAM system partition and the RAID controller can be best highlighted by walking through the data path taken by a write request as it enters the system.

Incoming write requests to the ONTAP Select VM are targeted at the VM's NVRAM partition. At the virtualization layer, this partition exists within an ONTAP Select system disk, a VMDK attached to the ONTAP Select VM. At the physical layer, these requests are cached in the local RAID controller, like all block changes targeted at the underlying spindles. From here, the write is acknowledged back to the host.

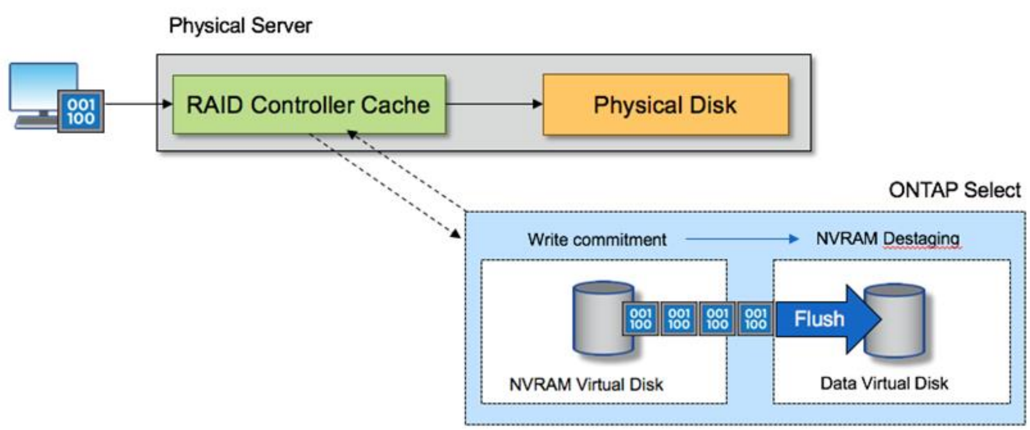
At this point, physically, the block resides in the RAID controller cache, waiting to be flushed to disk. Logically, the block resides in NVRAM waiting for destaging to the appropriate user data disks.

Because changed blocks are automatically stored within the RAID controller's local cache, incoming writes to the NVRAM partition are automatically cached and periodically flushed to physical storage media. This should not be confused with the periodic flushing of NVRAM contents back to ONTAP data disks. These two events are unrelated and occur at different times and frequencies.

Figure 4 shows the I/O path an incoming write takes. It highlights the difference between the physical layer (represented by the RAID controller cache and disks) and the virtual layer (represented by the VM's NVRAM and data virtual disks).

Note: Although blocks changed on the NVRAM VMDK are cached in the local RAID controller cache, the cache is not aware of the VM construct or its virtual disks. It stores all changed blocks on the system, of which NVRAM is only a part. This includes write requests bound for the hypervisor, if it is provisioned from the same backing spindles.

Figure 4) Incoming writes to ONTAP Select VM.



Best Practice
The RAID controller cache is used to store all incoming block changes, not just those targeted toward the NVRAM partition. Therefore, when choosing a RAID controller, select one with the largest cache available. A larger cache allows less frequent disk flushing and an increase in performance for the ONTAP Select VM, the hypervisor, and any compute VMs collocated on the server.

Note that, starting with new installations of ONTAP Select 9.5 on ESX version 6.5 U2 or later, the NVRAM partition is separated on its own VMDK. That VMDK is attached using the vNVME driver available in ESX versions of 6.5 or later. This change is most significant for ONTAP Select installations with software RAID, which do not benefit from the RAID controller cache.

2.3 VSAN and External Array Configurations for ONTAP Select 9.4 and Later

ONTAP Select clusters are supported on VSAN, some HCI products, NetApp HCI technology, and external array types of datastores. This deployment model is generally referred to as virtual NAS or vNAS. In these configurations, datastore resiliency is assumed to be provided by the underlying infrastructure. The minimum requirement is that the underlying configuration is supported by VMware and, therefore, should be listed on the respective VMware HCLs.

vNAS Architectures

The vNAS nomenclature is used for all setups that do not use DAS. For multinode ONTAP Select clusters, this includes architectures for which the two ONTAP Select nodes in the same HA pair share a single datastore (including vSAN datastores). The nodes can also be installed on separate datastores from the same shared external array. This allows for array-side storage efficiencies to reduce the overall footprint of the entire ONTAP Select HA pair. The architecture of ONTAP Select vNAS solutions is very similar to that of ONTAP Select on DAS with a local RAID controller. That is to say that each ONTAP Select node continues to have a copy of its HA partner's data. ONTAP storage efficiency policies are node scoped. Therefore, array side storage efficiencies are preferable because they can potentially be applied across data sets from both ONTAP Select nodes.

It is also possible that each ONTAP Select node in an HA pair uses a separate external array. This is a common choice when using ONTAP Select Metrocluster SDS with external storage.

Note: Metrocluster SDS support for vNAS requires ONTAP Select 9.5 and ONTAP Deploy 2.10.1.

When using separate external arrays for each ONTAP Select node, it is very important that the two arrays provide similar performance characteristics to the ONTAP Select VM.

vNAS Architectures Versus local DAS with Hardware RAID Controllers

The vNAS architecture is logically most similar to the architecture of a server with DAS and a RAID controller. In both cases, ONTAP Select consumes datastore space. That datastore space is carved into VMDKs, and these VMDKs form the traditional ONTAP data aggregates. ONTAP Deploy makes sure that the VMDKs are properly sized and assigned to the correct plex (in the case of HA pairs) during cluster - create and storage-add operations.

There are two major differences between vNAS and DAS with a RAID controller. The most immediate difference is that vNAS does not require a RAID controller. vNAS assumes that the underlying external array provides the data persistence and resiliency that a DAS with a RAID controller setup would provide. The second and more subtle difference has to do with NVRAM performance.

vNAS NVRAM

The ONTAP Select NVRAM is a VMDK. In other words, ONTAP Select emulates a byte addressable space (traditional NVRAM) on top of a block addressable device (VMDK). However, the performance of the NVRAM is absolutely critical to the overall performance of the ONTAP Select node.

For DAS setups with a hardware RAID controller, the hardware RAID controller cache acts as the de facto NVRAM cache, because all writes to the NVRAM VMDK are first hosted in the RAID controller cache.

For VNAS architectures, ONTAP Deploy automatically configures ONTAP Select nodes with a boot argument called Single Instance Data Logging (SIDL). When this boot argument is present, ONTAP Select bypasses the NVRAM and writes the data payload directly to the data aggregate. The NVRAM is only used to record the address of the blocks changed by the WRITE operation. The benefit of this feature is that it avoids a double write: one write to NVRAM and a second write when the NVRAM is destaged. This feature is only enabled for vNAS because local writes to the RAID controller cache have a negligible additional latency.

The SIDL feature is not compatible with all ONTAP Select storage efficiency features. See Table 3 for an overview of all storage efficiency policies available with ONTAP Select.

The SIDL feature can be disabled at the aggregate level using the following command:

```
storage aggregate modify -aggregate aggr-name -single-instance-data-logging off
```

Note that write performance is affected if the SIDL feature is turned off. It is possible to re-enable the SIDL feature after all the storage efficiency policies on all the volumes in that aggregate are disabled:

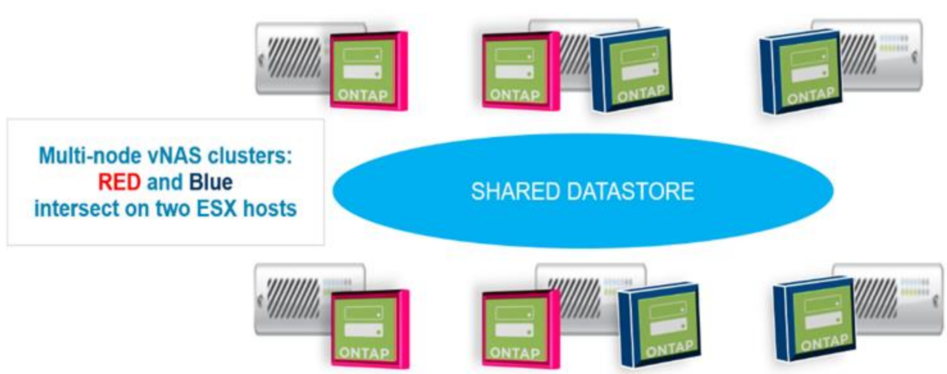
```
volume efficiency stop -all true -vserver * -volume * (all volumes in the affected aggregate)
```

Collocating ONTAP Select Nodes When Using vNAS

ONTAP Select 9.4 and ONTAP Deploy 2.8 include support for multinode ONTAP Select clusters on shared storage. ONTAP Deploy 2.8 enables the configuration of multiple ONTAP Select nodes on the same ESX host as long as these nodes are not part of the same cluster. Note that this configuration is only valid for VNAS environments (shared datastores). Multiple ONTAP Select instances per host are not supported when using DAS storage because these instances compete for the same hardware RAID controller.

ONTAP Deploy 2.8 makes sure that the initial deployment of the multinode VNAS cluster does not place multiple ONTAP Select instances from the same cluster on the same host. Figure 5 shows for an example of a correct deployment of two four-node clusters that intersect on two hosts.

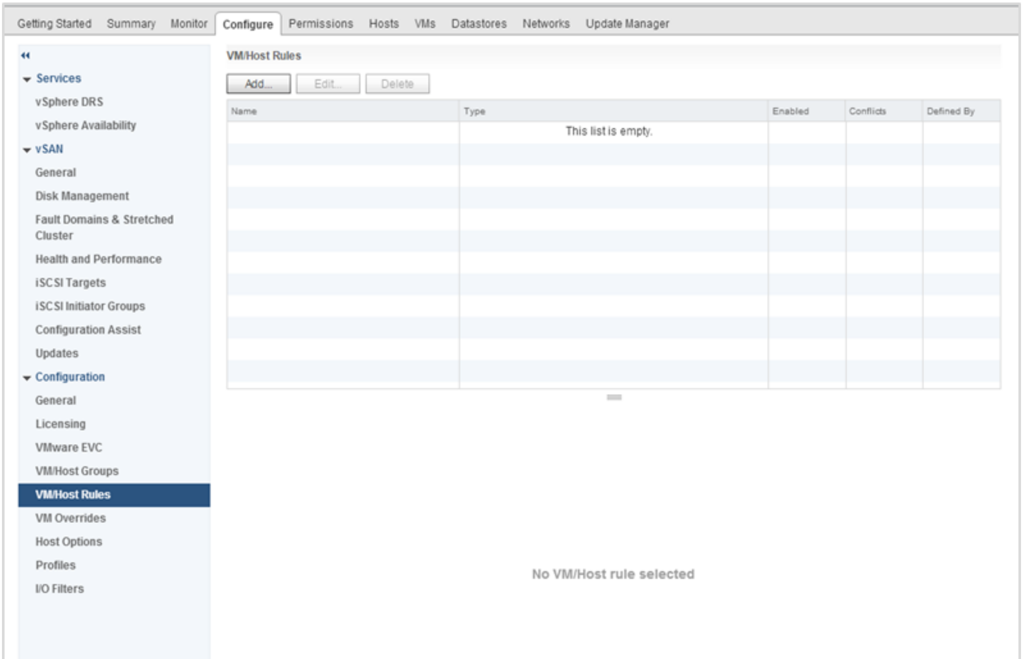
Figure 5) Initial deployment of multinode VNAS clusters.

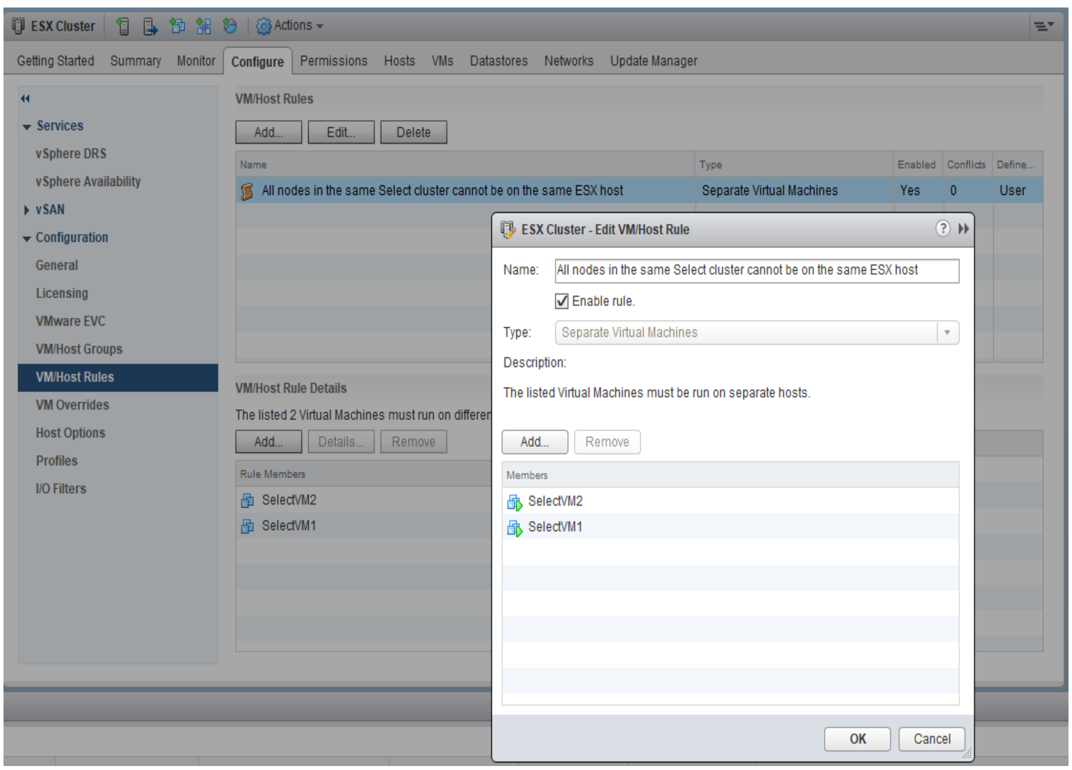


After deployment, the ONTAP Select nodes can be migrated between hosts. This could result in nonoptimal and unsupported configurations for which two or more ONTAP Select nodes from the same cluster share the same underlying host. NetApp recommends the manual creation of VM anti-affinity rules so that VMware automatically maintains physical separation between the nodes of the same cluster, not just the nodes from the same HA pair.

Note: Anti-affinity rules require that DRS is enabled on the ESX cluster.

See the following example on how to create an anti-affinity rule for the ONTAP Select VMs. If the ONTAP Select cluster contains more than one HA pair, all nodes in the cluster must be included in this rule.





Two or more ONTAP Select nodes from the same ONTAP Select cluster could potentially be found on the same ESX host for one of the following reasons:

- DRS is not present due to VMware vSphere license limitations or if DRS is not enabled.
- The DRS anti-affinity rule is bypassed because a VMware HA operation or administrator-initiated VM migration takes precedence.

Note that ONTAP Deploy does not proactively monitor the ONTAP Select VM locations. However, a cluster refresh operation reflects this unsupported configuration in the ONTAP Deploy logs:

2.4 Software RAID Services for Local Attached Storage

Independent of the hardware RAID configurations, ONTAP Select also provides a software RAID option. Software RAID is a RAID abstraction layer implemented within the ONTAP software stack. It provides the same functionality as the RAID layer within a traditional ONTAP platform such as FAS. The RAID layer performs drive parity calculations and provides protection against individual drive failures within an ONTAP Select node. A hardware RAID controller might not be available or might be undesirable in certain environments, such as when ONTAP Select is deployed on a small form-factor commodity hardware. Software RAID expands the available deployment options to include such environments. To enable software RAID in your environment, here are some points to remember:

- This feature is available starting with ONTAP Select 9.5 and Deploy 2.10, with the ESX hypervisor.
- It is available with a Premium or Premium XL license.
- It only supports SSD drives for ONTAP root and data disks.
- It requires a separate system disk for the ONTAP Select VM boot partition.
  - Choose a separate disk, either an SSD or an NVMe drive, to create a datastore for the system disks (NVRAM, Boot/CF card, Coredump, and Mediator in a multi-node setup).

Note: The terms service disk and system disk are used interchangeably. Service disks are the VMDKs that are used within the ONTAP Select VM to service various items such as clustering, booting, and so on. Service disks are physically located on a single physical disk (collectively called the service/system physical disk) as seen from the host. That physical disk must contain a DAS datastore. ONTAP Deploy creates these service disks for the ONTAP Select VM during cluster deployment.

Note: With the current release, it is not possible to further separate the ONTAP Select system disks across multiple datastores or across multiple physical drives.

Note: Hardware RAID is not deprecated.

Software RAID Configuration for Local Attached Storage

When using software RAID, the absence of a hardware RAID controller is ideal, but, if a system does have an existing RAID controller, it must adhere to the following requirements:

- The hardware RAID controller must be disabled such that disks can be presented directly to the system (a JBOD). This change can usually be made in the RAID controller BIOS
- Or the hardware RAID controller should be in the SAS HBA mode. For example, some BIOS configurations allow an “AHCI” mode in addition to RAID, which could be chosen to enable the JBOD mode. This enables a passthrough, so that the physical drives can be seen as is on the host.

Depending on maximum number of drives supported by the controller, an additional controller may be required. With the SAS HBA mode, ensure that the IO controller (SAS HBA) is supported with a minimum of 6Gb/s speed. However, NetApp recommends a 12Gbps speed.

No other hardware RAID controller modes or configurations is supported. For example, some controllers allow a RAID 0 support that can artificially enable disks to pass-through but the implications can be undesirable. The supported size of physical disks (SSD only) is between 200GB – 16TB

Note: Administrators need to keep track of which drives are in use by the ONTAP Select VM and prevent inadvertent use of those drives on the host.

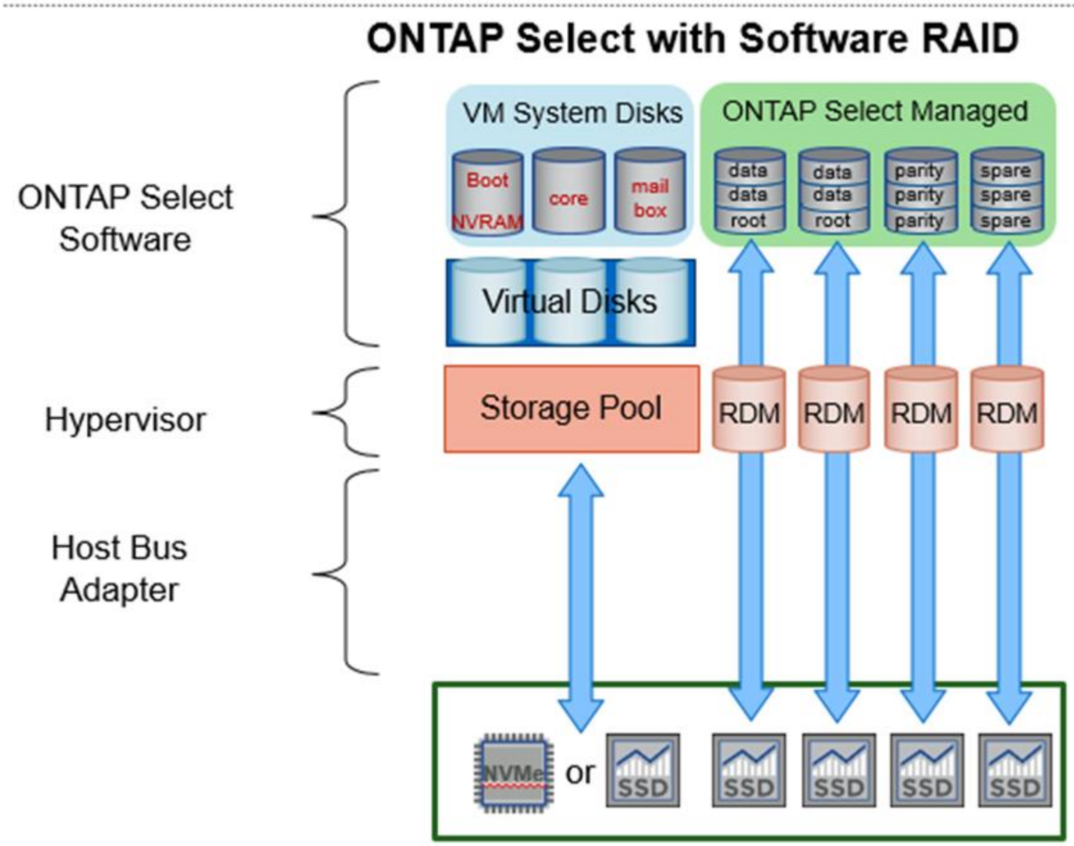
ONTAP Select Virtual and Physical Disks

For configurations with hardware RAID controllers, physical disk redundancy is provided by the RAID controller. ONTAP Select is presented with one or more VMDKs from which the ONTAP admin can configure data aggregates. These VMDKs are striped in a RAID 0 format because using ONTAP software RAID is redundant, inefficient, and ineffective due to resiliency provided at the hardware level. Furthermore, the VMDKs used for system disks are in the same datastore as the VMDKs used to store user data.

When using software RAID, ONTAP Deploy presents ONTAP Select with a set of virtual disks (VMDKs) and physical disks (Raw Device Mappings [RDMs]).

Figure 6 shows this relationship in more detail, highlighting the difference between the virtualized disks used for the ONTAP Select VM internals and the physical disks used to store user data.

Figure 6) ONTAP Select software RAID: use of virtualized disks and RDMs.



The system disks (VMDKs) reside in the same datastore and on the same physical disk. The virtual NVRAM disk requires a fast and durable media. Therefore, only NVMe and SSD-type datastores are supported.

Note: With the current release, it is not possible to further separate the ONTAP Select system disks across multiple datastores or multiple physical drives.

The data disks are presented to the ONTAP Select VM as raw disks through RDMs. RDMs contain metadata for managing and redirecting disk access to the physical device, which allows the host to pass

SCSI commands from the VM directly to the physical disk drives. Each raw disk exposed is divided into three parts: a small root partition (stripe) and two equal-sized partitions to create two data disks seen within the ONTAP Select VM. Partitions use the Root Data Data (RD2) schemes as shown in Figure 7 for a single node cluster and in Figure 8 for a node in an HA pair.

P denotes a parity drive. DP denotes a dual parity drive and S denotes a spare drive.

Figure 7) RDD disk partitioning for single-node clusters.

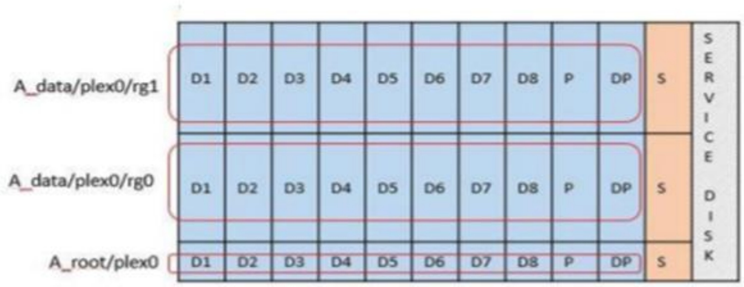
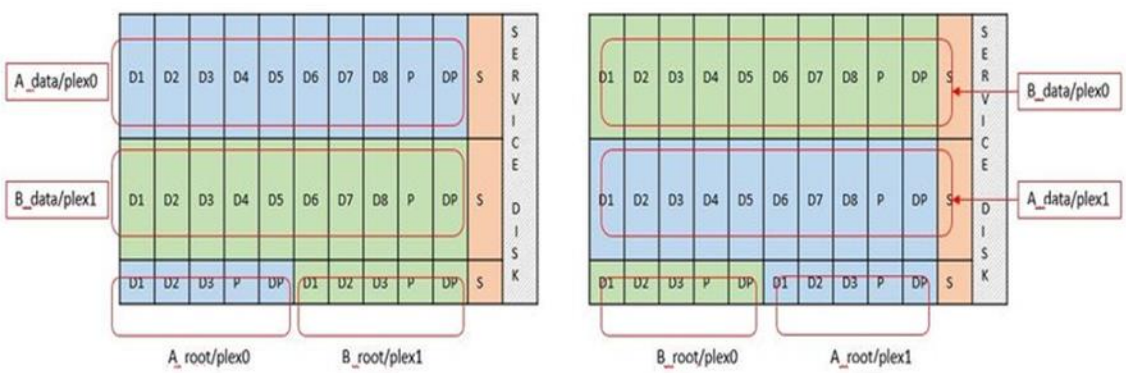


Figure 8) RDD disk partitioning for multinode clusters (HA pairs).



ONTAP software RAID supports the following RAID types: RAID 4, RAID-DP, and RAID-TEC. These are the same RAID constructs used by FAS and AFF platforms. However, ONTAP Select HA uses a shared-nothing architecture that replicates each node's configuration to the other node. That means each node has a root partition and a copy of the its peer's root partition. Each raw disk has a single root partition, which means that ONTAP Select has a minimum number of disks required to support software RAID. The minimum number of disks varies depending on whether the ONTAP Select node is part of an HA pair as well as the RAID type. See Table 2 for the specific number of drives required by each configuration.

Physical and Virtual Disk Provisioning

To provide a more streamlined user experience, ONTAP Deploy automatically provisions the system (virtual) disks from the specified datastore (physical system disk) and attaches them to the ONTAP Select VM. This operation occurs automatically during the initial setup so that the ONTAP Select VM can boot. The RDMs are partitioned and the root aggregate is automatically built. If the ONTAP Select node is part of an HA pair, the data partitions are automatically assigned to a local storage pool and a mirror storage pool. This assignment occurs automatically during both cluster-creation operations and storage-add operations.

Because the data disks on the ONTAP Select VM are associated with the underlying physical disks, there are performance implications for creating configurations with a larger number of physical disks.

Best Practice
NetApp recommends eight to 12 drives as the optimal RAID-group size. The maximum number of drives per RAID group is 24.

Note: The root aggregate’s RAID group type depends on the number of disks available. ONTAP Deploy picks the appropriate RAID group type. If it has sufficient disks allocated to the node, it uses RAID-DP, otherwise it creates a RAID-4 root aggregate.

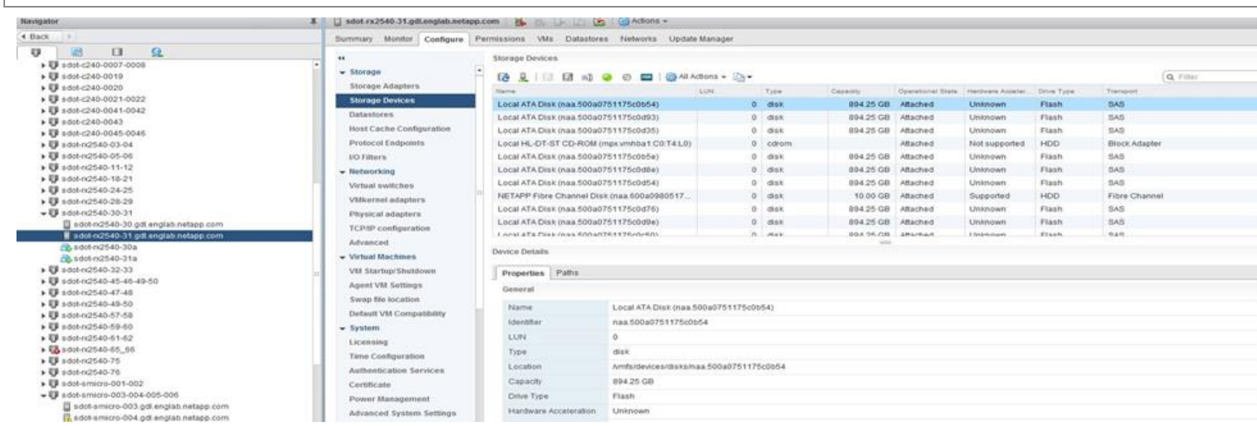
When adding capacity to an ONTAP Select VM using software RAID, the administrator must consider the physical drive size and the number of drives required. For details, see the section, “Increasing capacity for ONTAP Select with Software RAID”.

Similar to FAS and AFF systems, only drives with equal or larger capacities can be added to an existing RAID group. Larger capacity drives are right sized. If you are creating new RAID groups, the new RAID group size should match the existing RAID group size to make sure that the overall aggregate performance does not deteriorate.

Matching an ONTAP Select Disk to the Corresponding ESX Disk

ONTAP Select disks are usually labeled NET x.y. You can use the following ONTAP command to obtain the disk UUID:

```
<system name>::> disk show NET-1.1
                  Disk: NET-1.1
                 Model: Micron_5100_MTFD
         Serial Number: 1723175C0B5E
                   UID:
500A0751:175C0B5E:00000000:00000000:00000000:00000000:00000000:00000000:00000000:00000000
                   BPS: 512
         Physical Size: 894.3GB
              Position: shared
Checksum Compatibility: advanced_zoned
             Aggregate: -
                  Plex: -This UID can be matched with the device UID displayed in the ‘storage
devices’ tab for the ESX host
```



In the ESXi shell, you can enter the following command to blink the LED for a given physical disk (identified by its `naa.unique-id`).

```
esxcli storage core device set -d <naa_id> -l=locator -L=<seconds>
```


Replacing Failed Drives when Using Software RAID

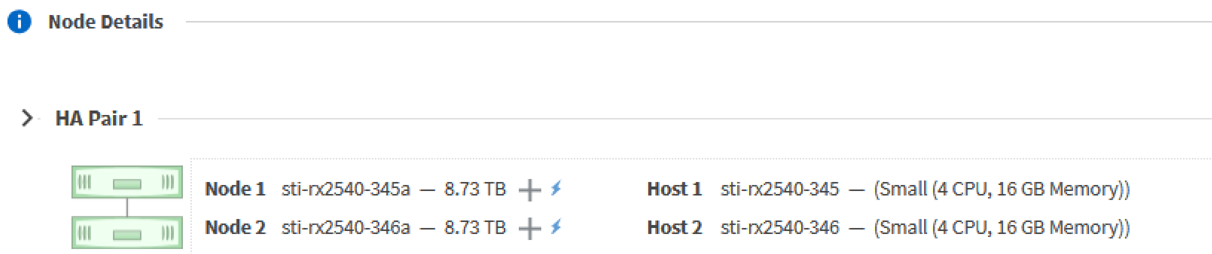
In a manner similar to ONTAP on FAS and AFF, ONTAP Select uses a spare drive (if one is available) and starts the rebuild process automatically. If no spare drive is available, one should be added to the ONTAP Select node.

Both the removal of the failed drive and the addition of a new drive (marked as a spare) must be performed through ONTAP Deploy. A sample workflow is provided below:

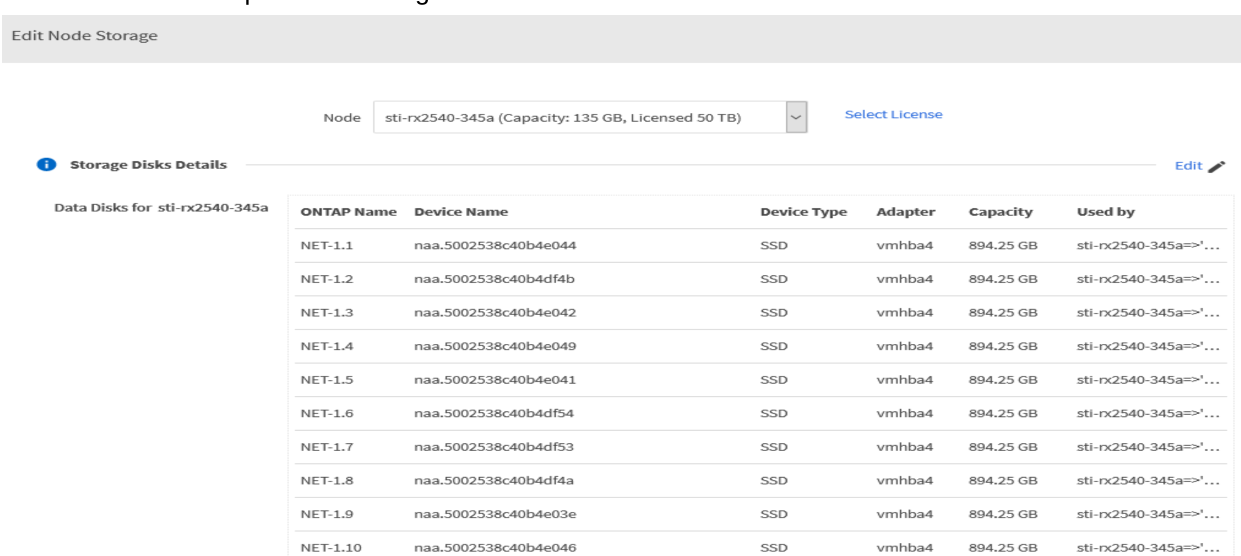
1. From the ONTAP Select CMD, identify the disk that has failed.

```
<cluster name>::> storage disk show -container-type broken
Usable           Disk    Container   Container
Disk                   Size Shelf Bay Type    Type        Name      Owner
---------------- ---------- ----- --- ------- ----------- --------- --------
NET-1.4              893.3GB     -   - SSD     broken      -         sti-rx2540-346a
```

2. From the ONTAP Deploy GUI, perform the following actions to remove the broken drive:
   a. Go to the Clusters page and click the relevant cluster.

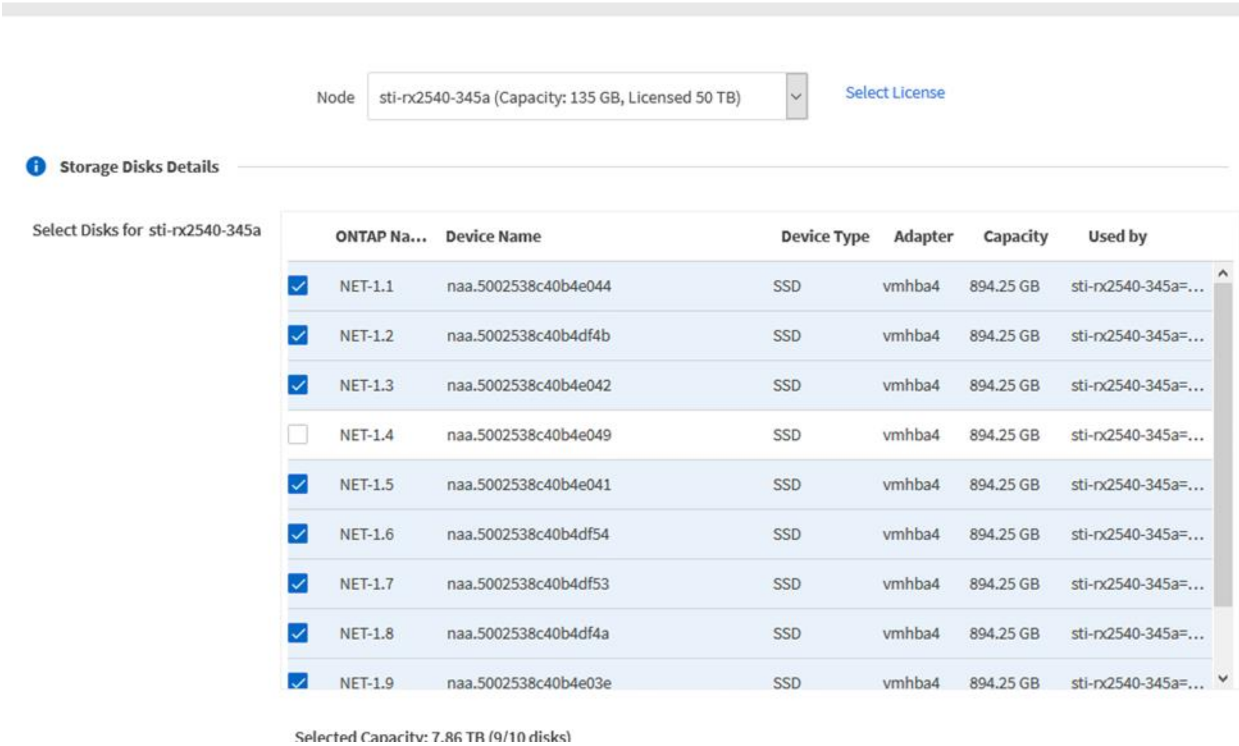


   b. Click + to expand the storage view.



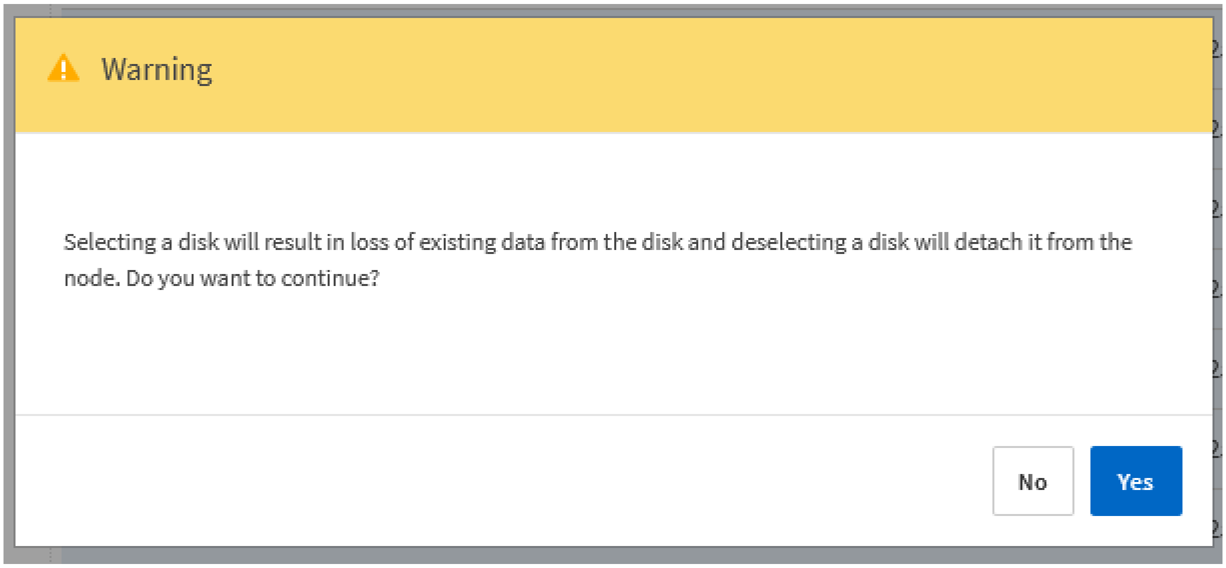
ONTAP Name	Device Name	Device Type	Adapter	Capacity	Used by
NET-1.1	naa.5002538c40b4e044	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.2	naa.5002538c40b4df4b	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.3	naa.5002538c40b4e042	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.4	naa.5002538c40b4e049	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.5	naa.5002538c40b4e041	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.6	naa.5002538c40b4df54	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.7	naa.5002538c40b4df53	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.8	naa.5002538c40b4df4a	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.9	naa.5002538c40b4e03e	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.10	naa.5002538c40b4e046	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...

   c. Click Edit to make changes to the attached disks and uncheck the broken drive.

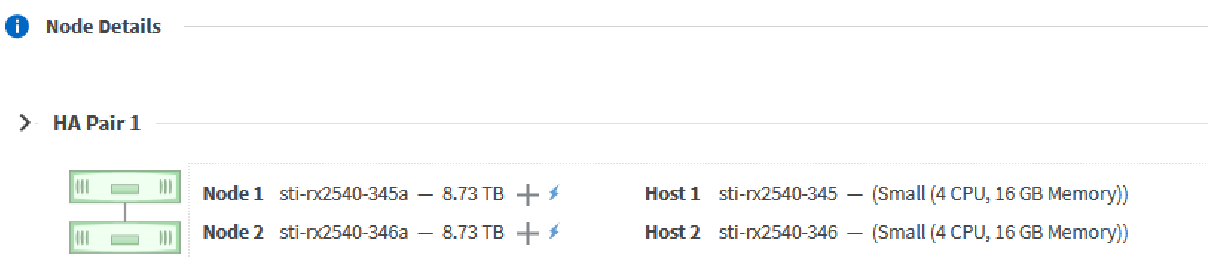


d. Provide the cluster credentials and click Edit Storage.

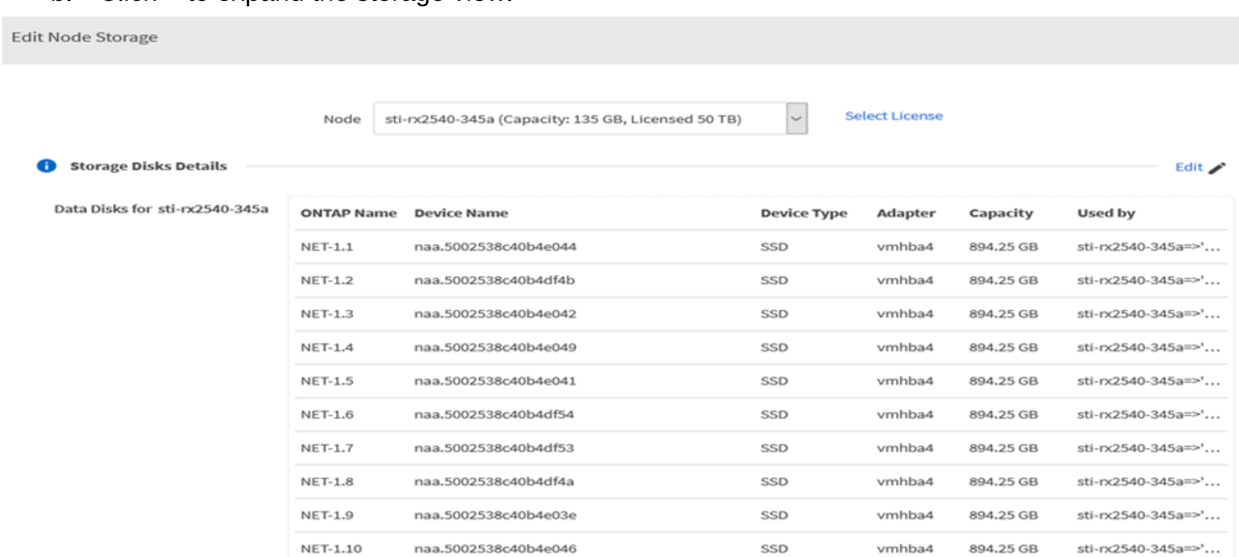
e. Confirm the operation.



3. From the ONTAP Deploy GUI, perform the following actions to add a spare drive:
   a. Go to the Clusters page and click the relevant cluster.

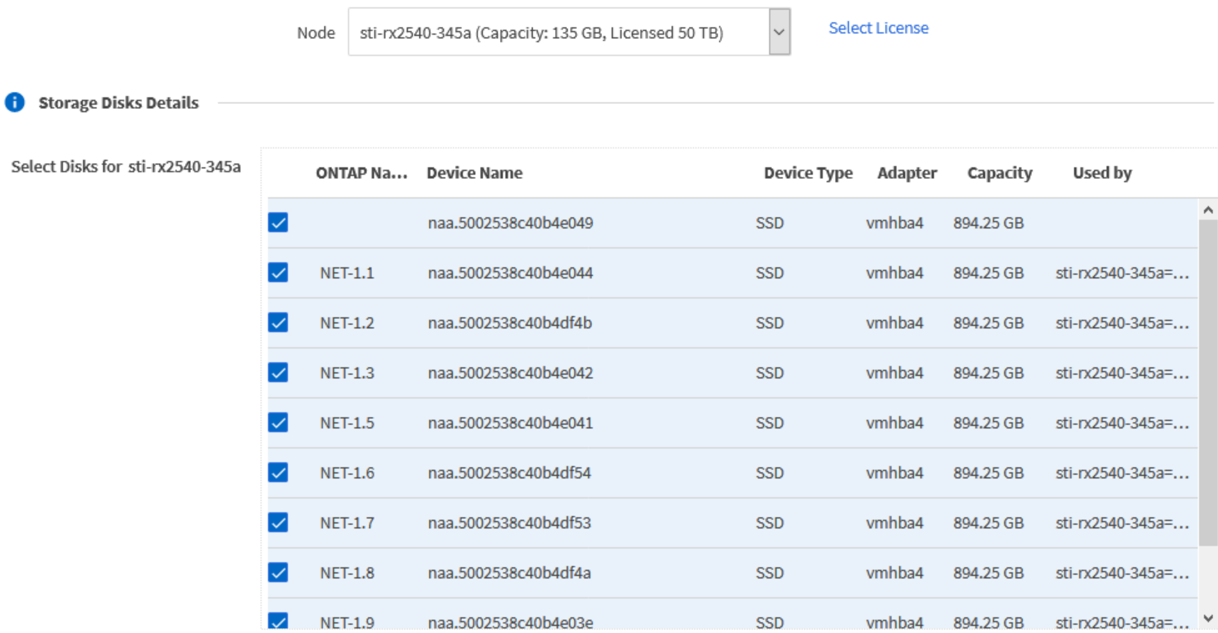


   b. Click + to expand the storage view.



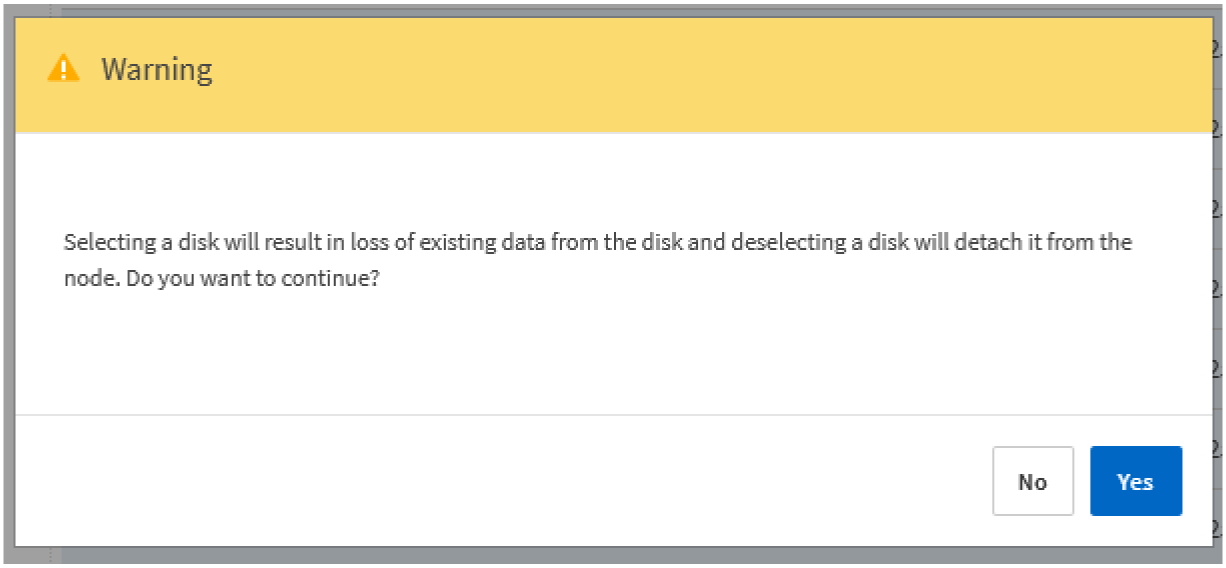
ONTAP Name	Device Name	Device Type	Adapter	Capacity	Used by
NET-1.1	naa.5002538c40b4e044	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.2	naa.5002538c40b4df4b	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.3	naa.5002538c40b4e042	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.4	naa.5002538c40b4e049	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.5	naa.5002538c40b4e041	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.6	naa.5002538c40b4df54	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.7	naa.5002538c40b4df53	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.8	naa.5002538c40b4df4a	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.9	naa.5002538c40b4e03e	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...
NET-1.10	naa.5002538c40b4e046	SSD	vmhba4	894.25 GB	sti-rx2540-345a=>'...

   c. Click Edit, check that the new drive is available, and then select it.



d. Provide the cluster credentials and click Edit Storage.

e. Confirm the operation.



Note: All drive addition and removal operations should be performed with ONTAP Deploy. Attaching a drive to the ONTAP Select VM using vSphere is not supported.

Multiple Drive Failures when Using Software RAID

It is possible for a system to encounter a situation in which multiple drives are in a failed state at the same time. The behavior of the system depends on the aggregate RAID protection and the number of failed drives.

A RAID4 aggregate can survive one disk failure, a RAID-DP aggregate can survive two disk failures, and a RAID-TEC aggregate can survive three disks failures.

If the number of failed disks is less than the maximum number of failures that RAID type supports, and if a spare disk is available, the reconstruction process starts automatically. If spare disks are not available, the aggregate serves data in a degraded state until spare disks are added.

If the number of failed disks is more than the maximum number of failures that the RAID type supports, then the local plex is marked as failed, and the aggregate state is degraded. Data is served from the second plex residing on the HA partner. This means that any I/O requests for node 1 are sent through cluster interconnect port e0e (iSCSI) to the disks physically located on node 2. If the second plex also fails, then the aggregate is marked as failed and data is unavailable.

A failed plex must be deleted and recreated for the proper mirroring of data to resume. Note that a multi-disk failure resulting in a data aggregate being degraded also results in a root aggregate being degraded. ONTAP Select uses the root-data-data (RDD) partitioning schema to split each physical drive into a root partition and two data partitions. Therefore, losing one or more disks might impact multiple aggregates, including the local root or the copy of the remote root aggregate, as well as the local data aggregate and the copy of the remote data aggregate.

```
rx254023::> aggr plex delete -aggregate aggr_22a -plex plex0

Warning: Deleting plex "plex0" of mirrored aggregate "aggr_22a" in a non-shared
         HA configuration will disable its synchronous mirror protection and
         may disable negotiated takeover of node "sdot-rx2540-22a" when
         aggregate "aggr_22a" is online.
Do you want to continue? {y|n}: y
[Job 53] Job succeeded: DONE

rx254023::> aggr mirror -aggregate aggr_22a
Info: Disks would be added to aggregate "aggr_22a" on node "sdot-rx2540-22a" in
```

```
     the following manner:

     Second Plex

       RAID Group rg0, 2 disks (advanced_zoned checksum, raid4)
        Position   Disk                      Type                  Size
        ---------- ------------------------- ---------- ---------------
        shared     NET-1.5                   SSD                      -
        shared     NET-1.6                   SSD                424.6GB

     Aggregate capacity available for volume use would be 363.0GB.

Do you want to continue? {y|n}: y

Mirroring of aggregate "aggr_22a" has been initiated. 1 disk needs to be zeroed before it can be
added to the aggregate. The process has been initiated. Once zeroing completes on this disk, all
disks will be added at once. Note that if the system reboots before the disk zeroing is complete,
an inactive plex will exist with no disks. The inactive plex can be removed manually. Use the
"storage aggregate plex delete" command to delete the inactive plex.

rx254023::> aggr show-status aggr_22a
Owner Node: sdot-rx2540-22a

Aggregate: aggr_22a (online, raid4, mirrored) (advanced_zoned checksums)

 Plex: /aggr_22a/plex1 (online, normal, active, pool1)
RAID Group /aggr_22a/plex1/rg0 (normal, advanced_zoned checksums)
                                                            Usable Physical
   Position Disk                        Pool Type     RPM     Size     Size Status
   -------- --------------------------- ---- ----- ------ -------- -------- ----------
   shared   NET-3.4                      1   SSD        -  418.0GB  894.3GB (normal)
   shared   NET-3.6                      1   SSD        -  418.0GB  894.3GB (normal)

 Plex: /aggr_22a/plex3 (online, normal, active, pool0)
  RAID Group /aggr_22a/plex3/rg0 (normal, advanced_zoned checksums)
                                                             Usable Physical
    Position Disk                        Pool Type     RPM     Size     Size Status
    -------- --------------------------- ---- ----- ------ -------- -------- ----------
    shared   NET-1.5                      0   SSD        -  418.0GB  894.3GB (normal)
    shared   NET-1.6                      0   SSD        -  418.0GB  894.3GB (normal)

4 entries were displayed.
```

Note: In order to test or simulate one or multiple drive failures, use the `storage disk fail -disk NET-x.y -immediate` command. If there is a spare in the system, the aggregate will begin to reconstruct. You can check the status of the reconstruction using the command “storage aggregate show”. You can remove the simulated failed drive using ONTAP Deploy. Note that ONTAP has marked the drive as Broken. The drive is not actually broken and can be added back using ONTAP Deploy. In order to erase the Broken label, enter the following commands in the ONTAP Select CLI:

```
set diag
disk unfail -disk NET-x.y -spare true
       disk show -broken
```

The output for the last command should be empty.

Virtualized NVRAM

NetApp FAS systems are traditionally fitted with a physical NVRAM PCI card. This card is a high-performing card containing nonvolatile flash memory that provides a significant boost in write performance. It does this by granting ONTAP the ability to immediately acknowledge incoming writes back to the client. It can also schedule the movement of modified data blocks back to slower storage media in a process known as destaging.

Commodity systems are not typically fitted with this type of equipment. Therefore, the functionality of the NVRAM card has been virtualized and placed into a partition on the ONTAP Select system boot disk. It is for this reason that placement of the system virtual disk of the instance is extremely important. For environments using ESX 6.5, ONTAP Select 9.5 uses a virtual NVME driver for accessing the system disks regardless of whether the underlying disk is SSD or NVMe. However, NetApp only supports NVMe for the physical system disk.

Best Practice
NetApp recommends using ESX 6.5 U2 or later and an NVMe disk for the datastore hosting the system disks. This configuration provides the best performance for the NVRAM partition. Note that when installing on ESX 6.5 U2 and higher, ONTAP Select utilizes the vNVME driver regardless of whether the system disk resides on an SSD or on an NVME disk. This sets the VM hardware level to 13, which is compatible with ESX 6.5 and newer.

2.5 High Availability Architecture

Although customers are starting to move application workloads from enterprise-class storage appliances to software-based solutions running on commodity hardware, the expectations and needs around resiliency and fault tolerance have not changed. An HA solution providing a zero recovery point objective (RPO) protects the customer from data loss due to a failure from any component in the infrastructure stack.

A large portion of the SDS market is built on the notion of shared-nothing storage, with software replication providing data resiliency by storing multiple copies of user data across different storage silos. ONTAP Select builds on this premise by using the synchronous replication features (RAID SyncMirror) provided by ONTAP to store an extra copy of user data within the cluster. This occurs within the context of an HA pair. Every HA pair stores two copies of user data: one on storage provided by the local node, and one on storage provided by the HA partner. Within an ONTAP Select cluster, HA and synchronous replication are tied together, and the functionality of the two cannot be decoupled or used independently. As a result, the synchronous replication functionality is only available in the multinode offering.

Note: In an ONTAP Select cluster, synchronous replication functionality is a function of HA implementation, not a replacement for the asynchronous SnapMirror or SnapVault replication engines. Synchronous replication cannot be used independently from HA.

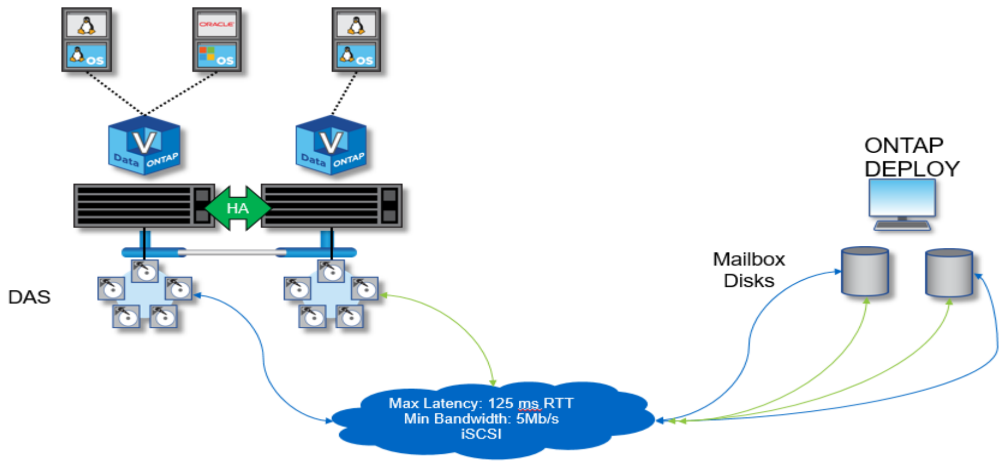
There are two ONTAP Select HA deployment models: the multinode clusters (four, six, or eight nodes) and the two-node clusters. The salient feature of a two-node ONTAP Select cluster is the use of an external mediator service to resolve split-brain scenarios. The ONTAP Deploy VM serves as the default mediator for all the two-node HA pairs that it configures.

There are minimum version requirements for these HA configurations:

- Four-node HA is supported with all ONTAP Select and ONTAP Deploy releases.
- Two-node HA requires the minimum versions of ONTAP Select 9.2 and ONTAP Deploy 2.4.
- Six-node and eight-node clusters require minimum versions of ONTAP Select 9.3 and ONTAP Deploy 2.7.

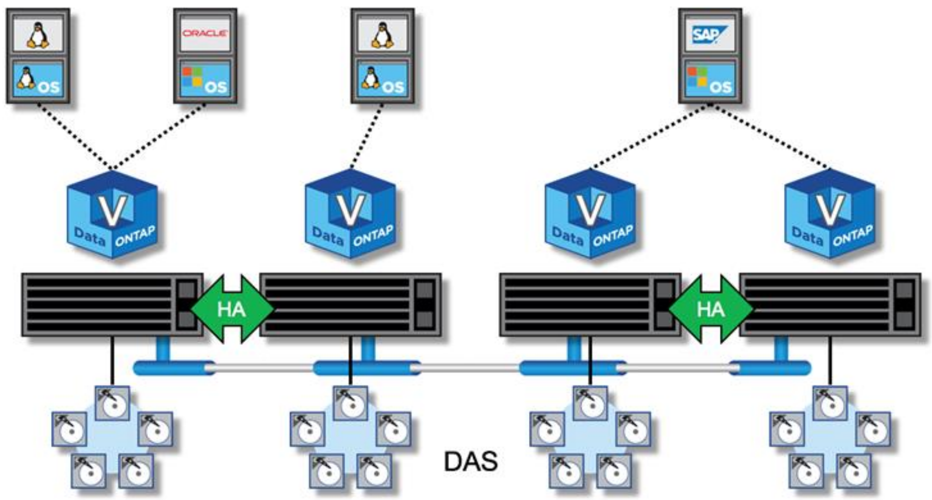
The two architectures are represented in Figure 9 and Figure 10.

Figure 9) Two-node ONTAP Select cluster with remote mediator and using local-attached storage.



Note: The two-node ONTAP Select cluster is composed of one HA pair and a mediator. Within the HA pair, data aggregates on each cluster node are synchronously mirrored, and, in the event of a failover, there is no loss of data.

Figure 10) Four-node ONTAP Select cluster using local-attached storage.



Note: The four-node ONTAP Select cluster is composed of two HA pairs. Six-node and eight-node clusters are composed of three and four HA pairs, respectively. Within each HA pair, data aggregates on each cluster node are synchronously mirrored, and, in the event of a failover, there is no loss of data.

Note: Only one ONTAP Select instance can be present on a physical server when using DAS storage. ONTAP Select requires unshared access to the local RAID controller of the system and is designed to manage the locally attached disks, which would be impossible without physical connectivity to the storage.

Two-Node HA Versus Multinode HA

Unlike FAS arrays, ONTAP Select nodes in an HA pair communicate exclusively over the IP network. That means that the IP network is a single point of failure (SPOF), and protecting against network partitions and split-brain scenarios becomes an important aspect of the design. The multinode cluster can sustain single-node failures because the cluster quorum can be established by the three or more surviving nodes. The two-node cluster relies on the mediator service hosted by the ONTAP Deploy VM to achieve the same result. The minimum version of the ONTAP Deploy VM required to support a two-node cluster with the mediator service is 2.4.

The heartbeat network traffic between the ONTAP Select nodes and the ONTAP Deploy mediator service is minimal and resilient so that the ONTAP Deploy VM can be hosted in a different data center than the ONTAP Select two-node cluster.

Note: The ONTAP Deploy VM becomes an integral part of a two-node cluster when serving as the mediator for that cluster. If the mediator service is not available, the two-node cluster continues serving data, but the storage failover capabilities of the ONTAP Select cluster are disabled. Therefore, the ONTAP Deploy mediator service must maintain constant communication with each ONTAP Select node in the HA pair. A minimum bandwidth of 5Mbps and a maximum round-trip time (RTT) latency of 125ms are required to allow proper functioning of the cluster quorum.

If the ONTAP Deploy VM acting as a mediator is temporarily or potentially permanently unavailable, a secondary ONTAP Deploy VM (minimum version 2.4) can be used to restore the two-node cluster quorum. This results in a configuration in which the new ONTAP Deploy VM is unable to manage the ONTAP Select nodes, but it successfully participates in the cluster quorum algorithm. The communication between the ONTAP Select nodes and the ONTAP Deploy VM is done by using the iSCSI protocol over IPv4. The ONTAP Select node management IP address is the initiator, and the ONTAP Deploy VM IP address is the target. Therefore, it is not possible to support IPv6 addresses for the node management IP addresses when creating a two-node cluster. The ONTAP Deploy hosted mailbox disks are automatically created and masked to the proper ONTAP Select node management IP addresses at the time of two-node cluster creation. The entire configuration is automatically performed during setup, and no further administrative action is required. The ONTAP Deploy instance creating the cluster is the default mediator for that cluster.

An administrative action is required if the original mediator location must be changed. It is possible to recover a cluster quorum even if the original ONTAP Deploy VM is lost. However, NetApp recommends that you back up the ONTAP Deploy database after every two-node cluster is instantiated.

For a complete list of steps required to configure a new mediator location, see the ONTAP Select 9 Installation and Cluster Deployment Guide.

Two-Node HA Versus Two-Node Stretched HA (MetroCluster SDS)

Starting with ONTAP Select 9.3 and ONTAP Deploy, it is possible to stretch a two-node, active/active HA cluster across larger distances and potentially place each node in a different data center. The only distinction between a two-node cluster and a two-node stretched cluster (also referred to as MetroCluster SDS) is the network connectivity distance between nodes.

The two-node cluster is defined as a cluster for which both nodes are located in the same data center within a distance of 300m. In general, both nodes have uplinks to the same network switch or set of interswitch link (ISL) network switches.

Two-node MetroCluster SDS is defined as a cluster for which nodes are physically separated (different rooms, different buildings, and different data centers) by more than 300m. In addition, each node’s uplink connections are connected to separate network switches. The MetroCluster SDS does not require dedicated hardware. However, the environment should adhere to requirements for latency (a maximum of 5ms for RTT and 5ms for jitter, for a total of 10ms) and physical distance (a maximum of 10km).

MetroCluster SDS is a premium feature and requires a Premium license. The Premium license supports the creation of both small and medium VMs, as well as HDD and SSD media.

Note: Starting with ONTAP Select 9.5 and ONTAP Deploy 2.10. MetroCluster SDS is supported with both local attached storage (DAS) and shared storage (vNAS). Note that vNAS configurations usually have a higher innate latency because of the network between the ONTAP Select VM and shared storage. MetroCluster SDS configurations must provide a maximum of 10ms of latency between the nodes, including the shared storage latency. In other words, only measuring the latency between the Select VMs is not adequate because shared storage latency is not negligible for these configurations.

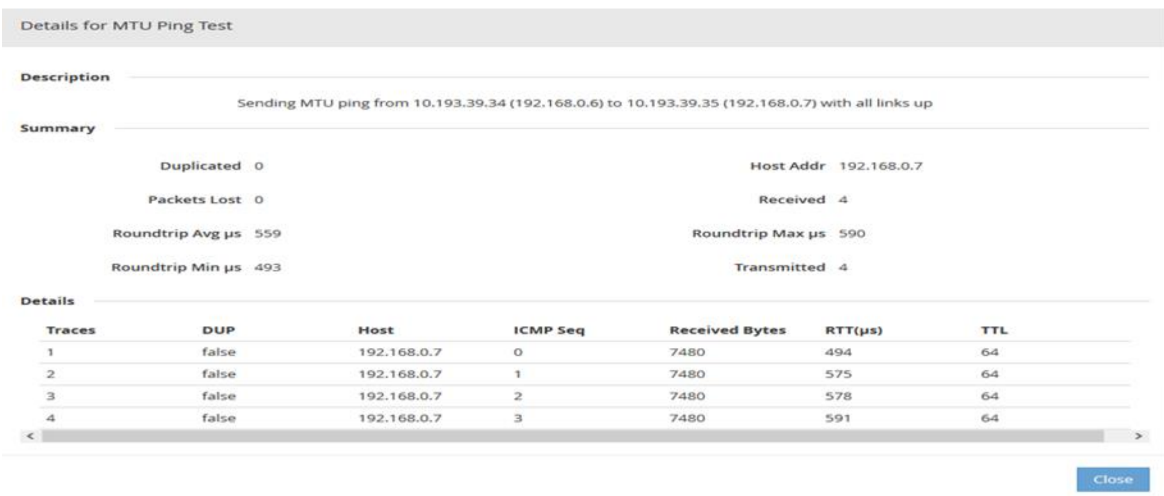
Two-Node Stretched HA (MetroCluster SDS) Best Practices

Before you create a MetroCluster SDS, use the ONTAP Deploy connectivity checker functionality to make sure that the network latency between the two data centers falls within the acceptable range.

1. After installing ONTAP Deploy, define two ESX hosts (one in each data center) that are used to measure the latency between the two sites.
2. Select Administration (top of screen) > Network > Connectivity Checker (left panel). The default settings are appropriate.

Note: The connectivity checker does not mark the test as failed if the latency exceeds 10ms. Therefore, you must check the value of the latency instead of the status of the connectivity checker test run.

The following example shows a connectivity checker output in which the latency between nodes is under 1ms.



Note: The connectivity checker does not check the latency between the ONTAP Select VM and the storage. When using external storage for MetroCluster SDS, the VM-to-storage latency is not negligible and the total latency must be under 10ms RTT.

The connectivity checker has the additional benefit of making sure that the internal network is properly configured to support a large MTU size. Starting with ONTAP Select 9.5 and ONTAP Deploy 2.10.1, the default MTU size is determined by querying the upstream vSwitch. However, the default MTU value can be manually overwritten to account for network overlay protocol overhead. The internal network MTU can be configured to between 7,500 and 9,000. This is a requirement for all HA traffic, whether the ONTAP Select cluster consist of two, four, six, or eight nodes.

There is an extra caveat when using virtual guest tagging (VGT) and two-node clusters. In two-node cluster configurations, the node management IP address is used to establish early connectivity to the mediator before ONTAP is fully available. Therefore, only external switch tagging (EST) and virtual switch tagging (VST) tagging is supported on the port group mapped to the node management LIF (port e0a). Furthermore, if both the management and the data traffic are using the same port group, only EST and VST are supported for the entire two-node cluster.

Synchronous Replication

The ONTAP HA model is built on the concept of HA partners. As explained earlier in this document, ONTAP Select extends this architecture into the nonshared commodity server world by using the RAID SyncMirror (RSM) functionality that is present in ONTAP to replicate data blocks between cluster nodes, providing two copies of user data spread across an HA pair.

Starting with ONTAP Deploy 2.7 and ONTAP Select 9.3, a two-node cluster with a mediator can be used to span two data centers. For more information, see the section "Two-Node HA Versus Two-Node Stretched HA (MetroCluster SDS)."

Mirrored Aggregates

An ONTAP Select cluster is composed of two to eight nodes. Each HA pair contains two copies of user data, synchronously mirrored across nodes over an IP network. This mirroring is transparent to the user, and it is a property of the data aggregate, automatically configured during the data aggregate creation process.

All aggregates in an ONTAP Select cluster must be mirrored for data availability in the event of a node failover and to avoid an SPOF in case of hardware failure. Aggregates in an ONTAP Select cluster are built from virtual disks provided from each node in the HA pair and use the following disks:

- A local set of disks (contributed by the current ONTAP Select node)
- A mirrored set of disks (contributed by the HA partner of the current node)

Note: The local and mirror disks used to build a mirrored aggregate must be the same size. These aggregates are referred to as plex 0 and plex 1 (to indicate the local and remote mirror pairs, respectively). The actual plex numbers can be different in your installation.

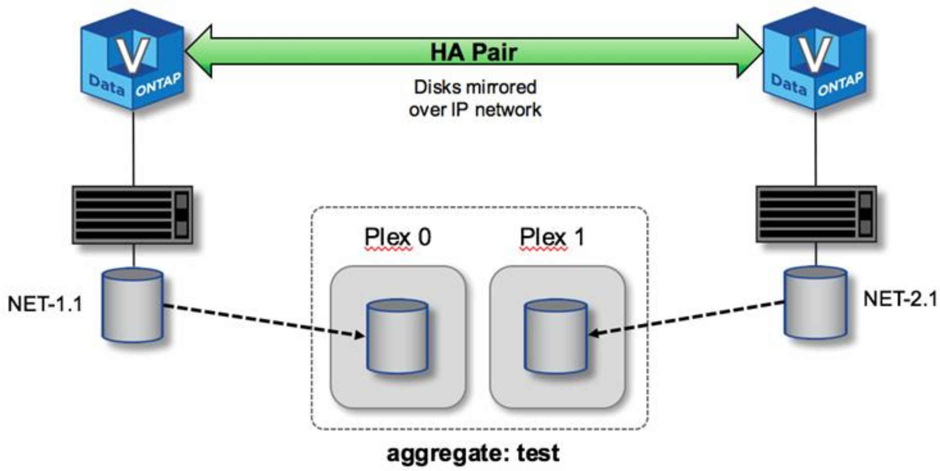
This approach is fundamentally different from the way standard ONTAP clusters work. This applies to all root and data disks within the ONTAP Select cluster. The aggregate contains both local and mirror copies of data. Therefore, an aggregate that contains N virtual disks offers N/2 disks' worth of unique storage, because the second copy of data resides on its own unique disks.

Figure 11 shows an HA pair within a four-node ONTAP Select cluster. Within this cluster is a single aggregate (test) that uses storage from both HA partners. This data aggregate is composed of two sets of virtual disks: a local set, contributed by the ONTAP Select owning cluster node (Plex 0), and a remote set, contributed by the failover partner (Plex 1).

Plex 0 is the bucket that holds all local disks. Plex 1 is the bucket that holds mirror disks, or disks responsible for storing a second replicated copy of user data. The node that owns the aggregate contributes disks to Plex 0, and the HA partner of that node contributes disks to Plex 1.

In Figure 11, there is a mirrored aggregate with two disks. The contents of this aggregate are mirrored across our two cluster nodes, with local disk NET-1.1 placed into the Plex 0 bucket and remote disk NET-2.1 placed into the Plex 1 bucket. In this example, aggregate test is owned by the cluster node to the left and uses local disk NET-1.1 and HA partner mirror disk NET-2.1.

Figure 11) ONTAP Select mirrored aggregate.



Note: When an ONTAP Select cluster is deployed, all virtual disks present on the system are automatically assigned to the correct plex, requiring no additional step from the user regarding disk assignment. This prevents the accidental assignment of disks to an incorrect plex and provides optimal mirror disk configuration.

Best Practice
Although the existence of the mirrored aggregate is needed to provide an up-to-date (RPO 0) copy of the primary aggregate, care should be taken that the primary aggregate does not run low on free space. A low-space condition in the primary aggregate can cause ONTAP to delete the common NetApp Snapshot™ copy used as the baseline for storage giveback. This works as designed to accommodate client writes. However, the lack of a common Snapshot copy on failback requires the ONTAP Select node to do a full baseline from the mirrored aggregate. This operation can take a significant amount of time in a shared-nothing environment. A good baseline for monitoring aggregate space utilization is up to 85%.

Write Path

Synchronous mirroring of data blocks between cluster nodes and the requirement for no data loss with a system failure have a significant impact on the path an incoming write takes as it propagates through an ONTAP Select cluster. This process consists of two stages:

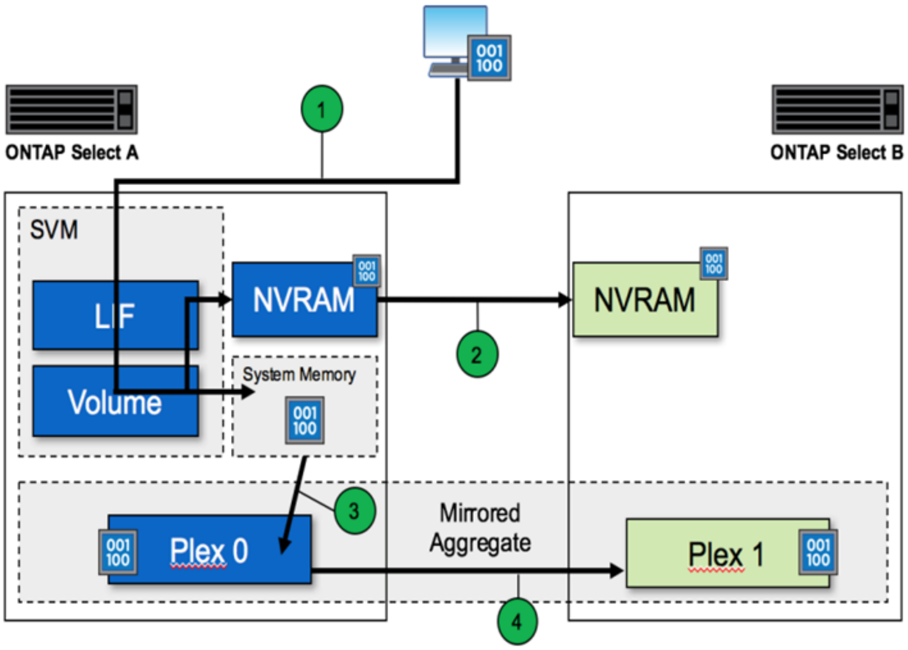
- Acknowledgment
- Destaging

Writes to a target volume occur over a data LIF and are committed to the virtualized NVRAM partition, present on a system disk of the ONTAP Select node, before being acknowledged back to the client. On an HA configuration, an additional step occurs, because these NVRAM writes are immediately mirrored to the HA partner of the target volume’s owner before being acknowledged. This process makes sure of the file system consistency on the HA partner node, if there is a hardware failure on the original node.

After the write has been committed to NVRAM, ONTAP periodically moves the contents of this partition to the appropriate virtual disk, a process known as destaging. This process only happens once, on the cluster node owning the target volume, and does not happen on the HA partner.

Figure 12 shows the write path of an incoming write request to an ONTAP Select node.

Figure 12) ONTAP Select write path workflow.



Incoming write acknowledgment includes the following steps:

1. Writes enter the system through a logical interface owned by ONTAP Select node A.
2. Writes are committed to the NVRAM of node A and mirrored to the HA partner, node B.
3. After the I/O request is present on both HA nodes, the request is then acknowledged back to the client.

ONTAP Select destaging from NVRAM to the data aggregate (ONTAP CP) includes the following steps:

1. Writes are destaged from virtual NVRAM to virtual data aggregate.
2. Mirror engine synchronously replicates blocks to both plexes.

Disk Heartbeating

Although the ONTAP Select HA architecture leverages many of the code paths used by the traditional FAS arrays, some exceptions exist. One of these exceptions is in the implementation of disk-based heartbeating, a nonnetwork-based method of communication used by cluster nodes to prevent network isolation from causing split-brain behavior. A split-brain scenario is the result of cluster partitioning, typically caused by network failures, whereby each side believes the other is down and attempts to take over cluster resources.

Enterprise-class HA implementations must gracefully handle this type of scenario. ONTAP does this through a customized, disk-based method of heartbeating. This is the job of the HA mailbox, a location on physical storage that is used by cluster nodes to pass heartbeat messages. This helps the cluster determine connectivity and therefore define quorum in the event of a failover.

On FAS arrays, which use a shared storage HA architecture, ONTAP resolves split-brain issues in the following ways:

- SCSI persistent reservations
- Persistent HA metadata
- HA state sent over HA interconnect

However, within the shared-nothing architecture of an ONTAP Select cluster, a node is only able to see its own local storage and not that of the HA partner. Therefore, when network partitioning isolates each side of an HA pair, the preceding methods of determining cluster quorum and failover behavior are unavailable.

Although the existing method of split-brain detection and avoidance cannot be used, a method of mediation is still required, one that fits within the constraints of a shared-nothing environment. ONTAP Select extends the existing mailbox infrastructure further, allowing it to act as a method of mediation in the event of network partitioning. Because shared storage is unavailable, mediation is accomplished through access to the mailbox disks over NAS. These disks are spread throughout the cluster, including the mediator in a two-node cluster, using the iSCSI protocol. Therefore, intelligent failover decisions can be made by a cluster node based on access to these disks. If a node can access the mailbox disks of other nodes outside of its HA partner, it is likely up and healthy.

Note: The mailbox architecture and disk-based heartbeating method of resolving cluster quorum and split-brain issues are the reasons the multinode variant of ONTAP Select requires either four separate nodes or a mediator for a two-node cluster.

HA Mailbox Posting

The HA mailbox architecture uses a message post model. At repeated intervals, cluster nodes post messages to all other mailbox disks across the cluster, including the mediator, stating that the node is up and running. Within a healthy cluster at any point in time, a single mailbox disk on a cluster node has messages posted from all other cluster nodes.

Attached to each Select cluster node is a virtual disk that is used specifically for shared mailbox access. This disk is referred to as the mediator mailbox disk, because its main function is to act as a method of cluster mediation in the event of node failures or network partitioning. This mailbox disk contains partitions for each cluster node and is mounted over an iSCSI network by other Select cluster nodes. Periodically, these nodes post health statuses to the appropriate partition of the mailbox disk. Using network-accessible mailbox disks spread throughout the cluster allows you to infer node health through a reachability matrix. For example, cluster nodes A and B can post to the mailbox of cluster node D, but not to the mailbox of node C. In addition, cluster node D cannot post to the mailbox of node C, so it is likely that node C is either down or network isolated and should be taken over.

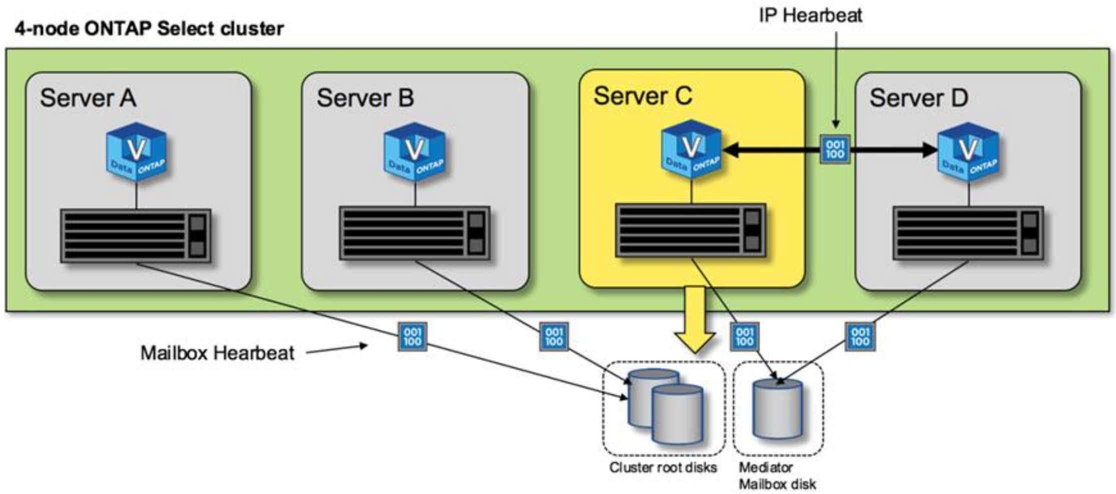
HA Heartbeating

Like with NetApp FAS platforms, ONTAP Select periodically sends HA heartbeat messages over the HA interconnect. Within the ONTAP Select cluster, this is performed over a TCP/IP network connection that exists between HA partners. Additionally, disk-based heartbeat messages are passed to all HA mailbox disks, including mediator mailbox disks. These messages are passed every few seconds and read back periodically. The frequency with which these are sent and received allows the ONTAP Select cluster to detect HA failure events within approximately 15 seconds, the same window available on FAS platforms. When heartbeat messages are no longer being read, a failover event is triggered.

Figure 13 shows the process of sending and receiving heartbeat messages over the HA interconnect and mediator disks from the perspective of a single ONTAP Select cluster node, node C.

Note: Network heartbeats are sent over the HA interconnect to the HA partner, node D, while disk heartbeats use mailbox disks across all cluster nodes, A, B, C, and D.

Figure 13) HA heartbeating in a four-node cluster: steady state.



HA Failover and Giveback

During a failover operation, the surviving node assumes the data serving responsibilities for its peer node using the local copy of its HA partner's data. Client I/O can continue uninterrupted, but changes to this data must be replicated back before giveback can occur. Note that ONTAP Select does not support a forced giveback because this causes changes stored on the surviving node to be lost.

The sync back operation is automatically triggered when the rebooted node rejoins the cluster. The time required for the sync back depends on several factors. These factors include the number of changes that must be replicated, the network latency between the nodes, and the speed of the disk subsystems on each node. It is possible that the time required for sync back will exceed the auto give back window of 10 minutes. In this case, a manual giveback after the sync back is required. The progress of the sync back can be monitored using the following command:

```
storage aggregate status -r -aggregate <aggregate name>
```

3 Deployment and Management

This section describes the deployment and management aspects of the ONTAP Select product.

3.1 ONTAP Select Deploy

The ONTAP Select cluster is deployed using specialized tooling that provides the administrator with the ability to build the ONTAP cluster and manage various aspects of the virtualized server. This utility, called ONTAP Select Deploy, comes packaged inside of an installation VM along with the ONTAP Select OS image. Bundling the deployment utility and ONTAP Select bits inside a single VM allows NetApp to include all the necessary support libraries and modules. Bundling also helps reduce the complexity of the interoperability matrix between various versions of ONTAP Select and the hypervisor.

The ONTAP Deploy application can be accessed through the following methods:

- CLI
- REST API
- GUI

The ONTAP Deploy CLI is shell-based and immediately accessible upon connection to the installation VM using SSH. Navigation of this shell is like navigation of the ONTAP shell, with commands bundled into groups that provide related functionality (for example, network create, network show, and network delete).

For automated deployments and integration into existing orchestration frameworks, ONTAP Deploy can also be invoked programmatically through a REST API. All functionality available through the shell-based CLI is available through the API. The entire list of API calls is documented using the Open API Specification (originally known as Swagger Specification).

ONTAP Deploy 2.8 uses v3 of the API. This version is not backward compatible with the prior versions of the API used with older ONTAP Deploy releases.

Deploy Upgrades

The Deploy utility can be upgraded separately from the Select cluster. Similarly, the Select cluster can be upgraded separately from the Deploy utility. See the upgrade section for the ONTAP Deploy and ONTAP Select interoperability matrix. Starting with ONTAP Deploy 2.8, an N-2 upgrade path is enforced. In other words, ONTAP Deploy direct upgrades are supported for the two prior releases of ONTAP Deploy. This has no bearing on the versions of ONTAP Select that are running in the client environment and no ONTAP upgrade is required.

Server Preparation

Although ONTAP Deploy provides the user with functionality that enables configuration of portions of the underlying physical server, there are several requirements that must be met before attempting to manage the server. This can be thought of as a manual preparation phase, because many of the steps are difficult to orchestrate through automation. This preparation phase involves the following tasks:

1. For local storage, configure the RAID controller and attached local storage, or, if using ONTAP software RAID, verify that the correct drive type and number of drives is available.
2. For VSAN or external array-hosted datastores, verify that the configurations are supported by VMware HCL, and follow the specific vendor best practices.
3. Verify physical network connectivity to the server. For external arrays, network resiliency, speed, and throughput are critical to the performance of the ONTAP Select VM.
4. Install the hypervisor.
5. Configure the virtual networking constructs (vSwitches and port groups).

Note: After the ONTAP Select cluster has been deployed, the appropriate ONTAP management tooling should be used to configure storage virtual machines (SVMs), LIFs, volumes, and so on. ONTAP Deploy does not provide this functionality.

The ONTAP Deploy utility and ONTAP Select software are bundled together into a single VM, which is then made available as an .OVA file for VMware vSphere. The bits are available from the NetApp Support site.

This installation VM runs the Debian Linux OS and has the following properties:

- Two vCPUs
- 4GB RAM
- 40GB virtual disk

ONTAP Select Deploy Placement in the Environment

Careful consideration should be given to the placement of the ONTAP Deploy installation VM, because the Deploy VM is used to verify hypervisor minimum requirements, deploy ONTAP Select clusters, and apply the license. Optionally, it can be used to troubleshoot network connectivity between Select nodes during the setup process.

ONTAP Deploy must also be able to communicate with the ONTAP Select node and cluster management IP addresses as follows:

- Ping
- SSH (port 22)
- SSL (port 443)

ONTAP Deploy uses the VMware VIX API to communicate with the vCenter and/or the ESX host as follows:

- HTTPS/SOAP on TCP port 443. This is the port for secure HTTP over TLS/SSL.
- Secondly, a connection to the ESX host is opened on a socket on TCP port 902. Data going over this connection is encrypted with SSL.
- In addition, ONTAP Deploy issues a ping command to verify that there is an ESX host responding at the IP address specified by the user.

VM Placement

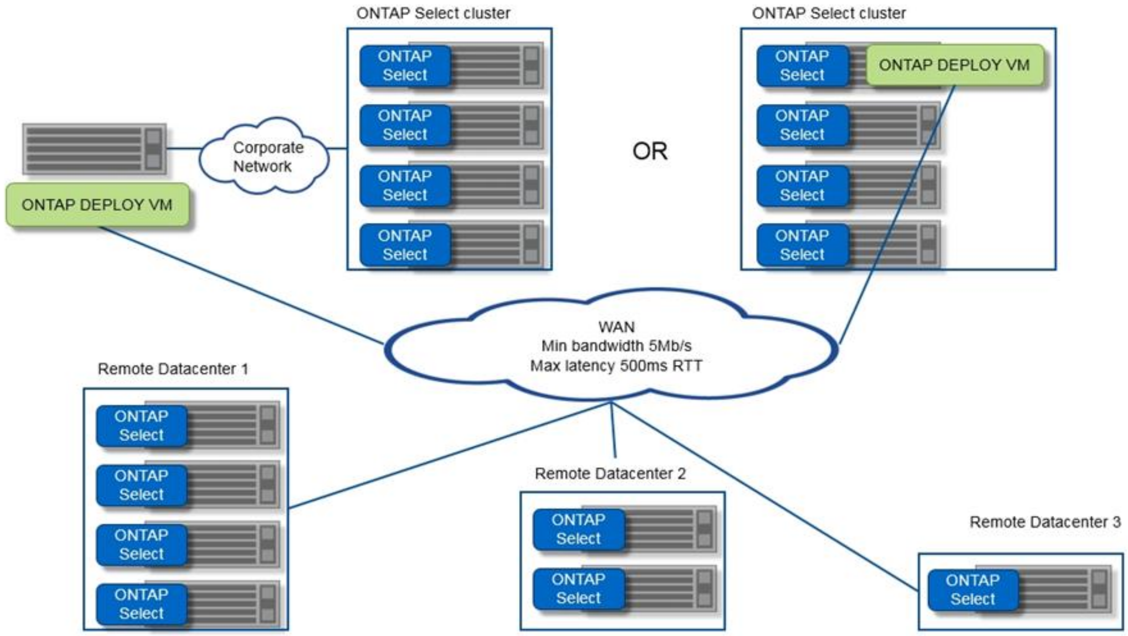
The ONTAP Select installation VM can be placed on any virtualized server in the customer environment. For four-node clusters, the ONTAP Deploy VM can be collocated on the same host as an ONTAP Select instance or on a separate virtualized server. For two-node clusters for which the ONTAP Deploy VM is also the cluster mediator, the collocation model is not supported because it would become a cluster SPOF.

The ONTAP Deploy VM can be installed in the same data center as the ONTAP Select cluster, or it can be centrally deployed in a core data center. The only requirement is that network connectivity exists between the ONTAP Deploy VM and the targeted ESX host and the future ONTAP Select cluster management IP address.

Note: Creating an ONTAP Select cluster over the WAN can take a considerably longer amount of time because the copying of the ONTAP Select binary files depends on the latency and bandwidth available between data centers. The maximum supported latency for creating remote ONTAP Select clusters is 500ms RTT. Deploying a two-node ONTAP Select cluster is supported on a WAN network in which the maximum latency and minimum bandwidth can support the more stringent mediator service traffic (minimum throughput 5Mbps; maximum latency 125ms RTT).

Figure 14 shows these deployment options.

Figure 14) ONTAP Select installation VM placement.



Note: Collocating the ONTAP Deploy VM and one of the ONTAP Select instances is not supported for two-node clusters.

Multiple ONTAP Select Deploy Instances

Depending on the complexity of the environment, it might be beneficial to have more than one ONTAP Deploy instance managing the ONTAP Select environment. For this scenario, make sure that each ONTAP Select cluster is managed by a single ONTAP Deploy instance. ONTAP Deploy stores cluster metadata within an internal database, so managing an ONTAP Select cluster using multiple ONTAP Deploy instances is not recommended.

When deciding whether to use multiple installation VMs, keep in mind that while ONTAP Deploy attempts to create unique MAC addresses by using a numeric hash based on the IP address of the installation VM, the uniqueness of the MAC address can only occur within that Deploy instance. Because there is no communication across Deploy instances, it is theoretically possible for two separate instances to assign multiple ONTAP Select network adapters with the same MAC address.

Best Practice
To eliminate the possibility of having multiple Deploy instances assign duplicate MAC addresses, one Deploy instance per layer-2 network should be used to manage an existing Select cluster or node or to create a new Select cluster or node.

Note: Each ONTAP Deploy instance can generate up to 64,000 unique MAC addresses. Each ONTAP Select node consumes four MAC addresses for its internal communication network schema. Each Deploy instance is also limited to managing 100 Select clusters and 400 hosts (a host is equivalent to one hypervisor server).

For two-node clusters, the ONTAP Deploy VM that creates the cluster is also the default mediator, and it requires no further configuration. However, it is critical that the mediator service is continuously available for proper functioning of the storage failover capabilities. For configurations in which the network latency, bandwidth, or other infrastructure issues require the repositioning of the mediator service closer to the ONTAP Select two-node cluster, another ONTAP Deploy VM can be used to host the mediator mailboxes temporarily or permanently.

Best Practice
The ONTAP Select two-node cluster should be carefully monitored for EMS messages indicating that storage failover is disabled. These messages indicate a loss of connectivity to the mediator service and should be rectified immediately.

3.2 Licensing ONTAP Select

When deploying ONTAP Select in a production environment, you must license the storage capacity used by the cluster nodes. Each ONTAP Select license is based on a flexible, consumption-based licensing model designed to allow customers to only pay for the storage they need. With ONTAP Select's original capacity tiers model, you must purchase a separate license for each node. Beginning with ONTAP Select 9.5 using Deploy 2.10, you now have the option of using capacity pools licensing instead. In both cases, you must use ONTAP Select Deploy to apply the licenses to the ONTAP Select nodes that are created by each instance of the Deploy utility.

Feature Evolution

The features and functionality of ONTAP Select licensing have continued to evolve. As mentioned above, ONTAP Select 9.5 using Deploy 2.10 now includes support for capacity pools licensing. Several changes were introduced with ONTAP Deploy 2.8 and ONTAP Select 9.4. One such change is that the ONTAP Select root aggregate no longer counts against the capacity license.

Also, the cluster create workflow in the web user interface now requires you to have a capacity license file at the time of deployment. With the capacity tiers model, it is no longer possible to create a production cluster using a serial number and then to apply a capacity license in the future. There is a CLI override available for the rare case in which a production serial number is available, but the corresponding license file is not yet available. In these situations, a valid license file must be applied within 30 days.

The biggest change introduced in ONTAP Deploy 2.8 and ONTAP Select 9.4 involves the license enforcement mechanism. With earlier versions of ONTAP Select, the virtual machines in a license violation situation reboot at midnight every day. The updated enforcement mechanism relies on blocking the aggregate operations (aggregate create and aggregate online). Although takeover operations are allowed, the giveback is blocked until the node comes into compliance with its capacity license.

When using a datastore to store the user data (in other words when using a hardware RAID controller as opposed to ONTAP software RAID) the user has the option to consume only a portion of a datastore. This functionality can be useful when the server capacity exceeds the desired Select license.

Allocation Characteristics and Overhead

The capacity license relates to the total size of the virtual data disks (VMDKs) attached to the ONTAP Select VM when using hardware RAID controllers. It also relates to the size of the data aggregates when using ONTAP software RAID.

In the case of multinode clusters, the per-node capacity license must cover both the active data on that node as well as the RAID SyncMirror copy of the active data on its HA peer.

Note: The actual amount of data stored on ONTAP Select is not relevant in the capacity license conversation; it can vary depending on data type and storage efficiency ratios. The amount of raw storage (defined as physical spindles inside the server) is also irrelevant because the datastore in which Select is installed can consume only a portion of the total space. For VSAN and external storage arrays, there is an additional aspect to keep in mind. The total space consumed by the

ONTAP Select VM varies depending on FTT/FTM and storage efficiency settings enabled at the VSAN and external storage array level. In these configurations, the ONTAP Select capacity license is not an indication of how much physical space the ONTAP Select VM consumes.

Administration

You can manage the capacity licenses through the Deploy web user interface by clicking the Administration tab and then clicking Licenses. You can also display all the nodes in a cluster and the respective licensing status using the system license show-status CLI command.

Common Characteristics for the Storage Capacity Licenses

The capacity tier and capacity pool licenses have several common characteristics, including the following:

- Storage capacity for a license is purchased in 1TB increments.
- Both the standard and premium performance tiers are supported.
- The nodes in an HA pair must have the same storage and license capacity.
- You must upload the license files to the Deploy administration utility, which then applies the licenses based on the type.

However, there are also several differences between the licensing models as described below.

3.3 Capacity Tiers Licensing

Capacity Tiers is the original licensing model provided with ONTAP Select. It continues to be supported with the latest ONTAP Select releases.

Storage Capacity Assigned to each ONTAP Select Node

With Capacity Tiers, you must purchase a license for each ONTAP Select node, and there is no concept of a cluster-level license. The assigned capacity is based on the purchase agreement. Any unused capacity cannot be moved to a different ONTAP Select node. The number of licenses would only exceed the number of nodes if a customer has purchased additional licenses for nodes that they are preparing to deploy.

Summary of Licensing Characteristics

The Capacity Tiers licensing model has the following characteristics:

- **License serial number.** The license serial number is a nine-digit number generated by NetApp for each node. Each license is locked to a specific ONTAP Select node with a matching serial number.
- **Node serial number.** The node serial number is nine digits long and is the same as the license serial number.
- **License duration.** The license is perpetual, and renewal is not required.

3.4 Capacity Pools Licensing

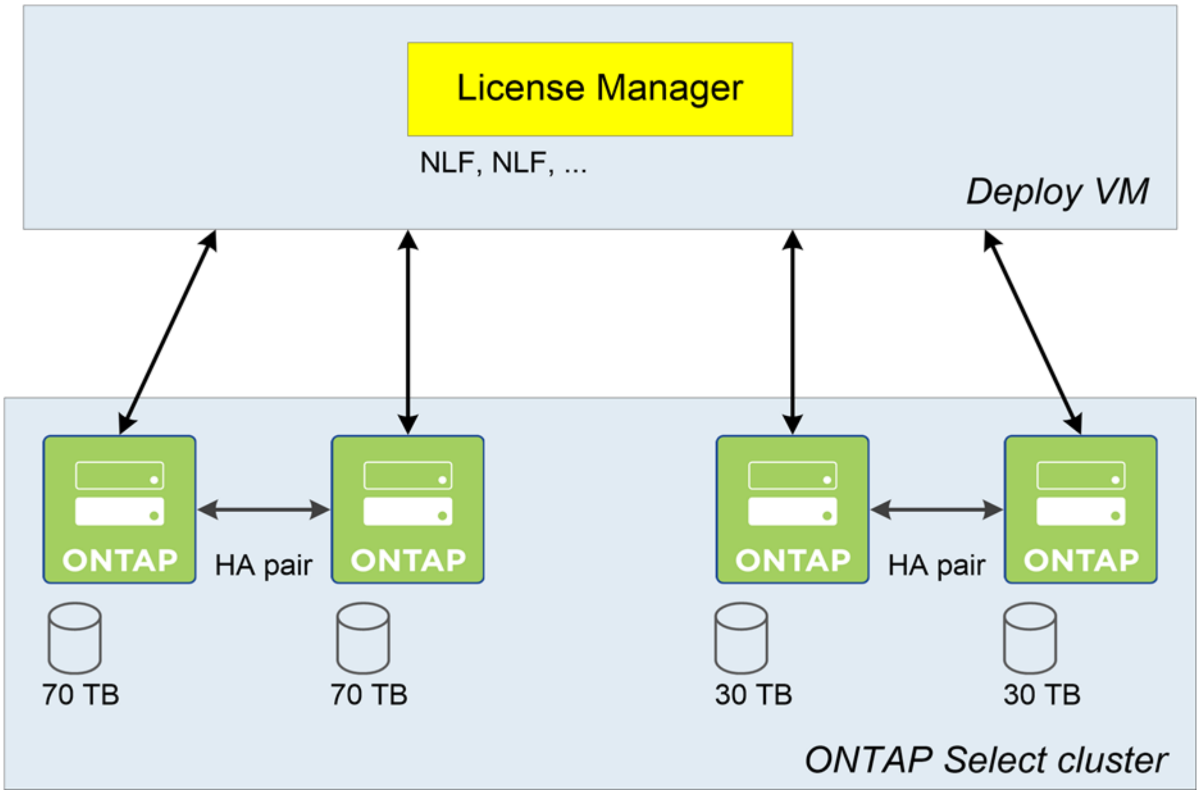
Capacity Pools is a new licensing model provided beginning with ONTAP Select 9.5 using Deploy 2.10. It provides an alternative to the capacity tiers model. The Capacity Pools licensing model provides several benefits, including the following:

- Storage capacity shared across one or more nodes
- More efficient allocation of storage capacity
- Significantly reduced administrative overhead and lower cost
- Improved usage metrics

Leasing Storage Capacity from a Shared Pool

Unlike the Capacity Tiers model, with Capacity Pools, you purchase a license for each shared pool. The nodes then lease capacity as needed from the single pool they are associated with. The License Manager (LM)—a new software component introduced with ONTAP Select Deploy 2.10—manages the Capacity Pool licenses and leases. LM is bundled with the Deploy utility as shown in Figure 15.

Figure 15) License Manager.



Every time a data aggregate is created, expanded, or changed, the ONTAP Select node must locate an active capacity lease or request a new lease from the LM. If a valid lease cannot be acquired, the data aggregate operation fails. The lease duration for each pool can be configured to between one hour and seven days, with a default of 24 hours. Leases are automatically renewed by the node. If a lease is not renewed for some reason, it expires and the capacity is returned to the pool.

Locking a Capacity Pool License to a License Manager Instance

After purchasing a capacity pool license from NetApp, you must associate the license serial number with a specific instance of LM. This is done through the License Lock ID (LLID), a unique 128-bit number identifying each LM instance (and therefore each Deploy instance). You can locate the LLID for your Deploy instance on the web user interface on the Administration page under System Settings. The LLID is also available in the Add Licenses section of the Getting Started page.

You provide both the license serial number and LLID when generating the license file. You can then upload the license file to the Deploy utility, so that the capacity pool can be used with new ONTAP Select cluster deployments.

Summary of the Licensing Characteristics

The Capacity Pools licensing model has the following characteristics:

- **License serial number.** The license serial number is a nine-digit number generated by NetApp for each capacity pool. Each license is locked to a specific License Manager instance.
- **Node serial number.** The node serial number is twenty digits long, is generated by the License Manager, and is assigned to the node.
- **License duration.** The license is valid for a limited term (such as one year) and must be renewed.

3.5 Modifying ONTAP Select Cluster Properties

ONTAP Select cluster properties such as cluster name, cluster management IP address, and node management IP address can be modified using ONTAP management tools such as System Manager. ONTAP Deploy is not notified when such modifications occur. Therefore, subsequent ONTAP Deploy management operations targeted at the ONTAP Select cluster fail. In a virtualized environment, the ONTAP Select VM name can also be changed, which would similarly result in ONTAP Deploy no longer being able to communicate with an ONTAP Select cluster.

Starting with ONTAP Deploy 2.6, the cluster refresh functionality allows ONTAP Deploy to recognize the following changes made to the ONTAP Select cluster:

- Networking configuration (IPs, netmasks, gateway, DNS, and NTP)
- ONTAP Select cluster or node names
- ONTAP Select version
- ONTAP Select VM name and state

The cluster refresh functionality works for any ONTAP Select node that is online and available (but has not been modified) at the time of upgrading to ONTAP Deploy 2.6. In other words, the older version of ONTAP Deploy must have knowledge of and access to the ONTAP Select node so that the ONTAP Deploy upgrade process can append some uniquely identifying information to that VM's metadata. After this unique identifier is stored in the VM's metadata and the ONTAP Deploy database, future changes to the ONTAP Select cluster or node properties can be synchronized with the ONTAP Deploy database by the cluster refresh operation. This process provides continued communication between ONTAP Deploy and the modified ONTAP Select VM.

3.6 ONTAP Management

Because ONTAP Select runs ONTAP, it supports all common NetApp management tools. As a result, after the product is deployed and ONTAP is configured, it can be administered using the same set of applications that a system administrator would use to manage FAS storage arrays. There is no special procedure required to build out an ONTAP configuration, such as creating SVMs, volumes, LIFs, and so on.

There are, however, several ONTAP Select management tasks that require the use of ONTAP Deploy. ONTAP Deploy is the only method to create Select clusters. Therefore, issues encountered during cluster creation can only be investigated using ONTAP Deploy. ONTAP Deploy communicates with the ONTAP Select clusters it created using the information configured at the time of deployment. This information includes the ESX host name or IP address and the ONTAP Select cluster management IP address. For two-node ONTAP Select clusters, the node management IP addresses are used for the iSCSI mediator traffic.

Changing the ONTAP Select node management IP addresses for two-node clusters after deployment results in an immediate loss of storage failover capabilities for that ONTAP Select cluster. A new mediator location on the same or a different ONTAP Deploy VM must be configured immediately.

The ability to change the ESX host name or IP address is not supported except for a VMware HA or vMotion. ONTAP Deploy attempts to rehost the ONTAP Select VM, as long as the new ESX host is managed by the same VMware vCenter Server.

After the cluster creation, ONTAP Deploy can be used to complement the other NetApp management tools for troubleshooting purposes.

The ONTAP Deploy CLI provides options for troubleshooting that are not available in the GUI. Most commands include a show option, which allows you to gather information about the environment.

The ONTAP Deploy logs can contain valuable information to help troubleshoot cluster setup issues. The ONTAP Deploy GUI and CLIs allow you to generate a NetApp AutoSupport® bundle containing the ONTAP Deploy logs. The GUI also allows you to download the bundle for immediate inspection.

Finally, the Deploy GUI can be used to invoke node-specific AutoSupport bundles.

ONTAP Deploy plays an important role in the quorum service for two-node clusters as well as troubleshooting of the environment. Therefore, the ONTAP Deploy database should be backed up regularly and after every change in the environment. Currently, it is not possible to rediscover an ONTAP Select cluster that was created by a different instance of ONTAP Deploy. Also, having an unmanaged cluster results in the loss of some important troubleshooting functionality. The ONTAP Deploy configuration database can be backed up by running the `deploy backup create` command from the ONTAP Deploy CLI.

4 Network Design Considerations

This section covers the various network configurations and best practices that should be considered when building an ONTAP Select cluster. Like the design and implementation of the underlying storage, care should be taken when making network design decisions. These choices have a significant impact on both the performance and resiliency of the ONTAP Select cluster. ONTAP Select 9.5 introduces support for the VMXNET3 driver, and it is the default driver for all new installations. Prior versions of ONTAP Select use the E1000 driver. Upgrading to ONTAP Select 9.5 does NOT automatically change the network driver. A manual procedure that includes an ONTAP Select node reboot is required. Contact NetApp Technical Support for further instructions.

Note: The ESX E1000 driver reports the speed as 1Gbps but it does not affect the actual throughput that the host can provide. Indeed, 10Gb NICs are fully supported at line speed. However, there are significant performance improvements when switching from the E1000 driver to the VMXNET3 driver. Therefore, NetApp recommends making the switch after upgrading to ONTAP Select 9.5.

In traditional FAS systems, ifgroups are used to provide aggregate throughput and fault tolerance using a single, logical, virtualized network interface configured on top of multiple physical network interfaces. ONTAP Select leverages the underlying hypervisor's virtualization of multiple physical network interfaces to achieve the same goals of throughput aggregation and resiliency. Therefore, the NICs that ONTAP Select manages are logical constructs, and configuring additional ifgroups does not achieve the goals of throughput aggregation or recovering from hardware failures. As a matter of fact, ifgroups are not supported with ONTAP Select.

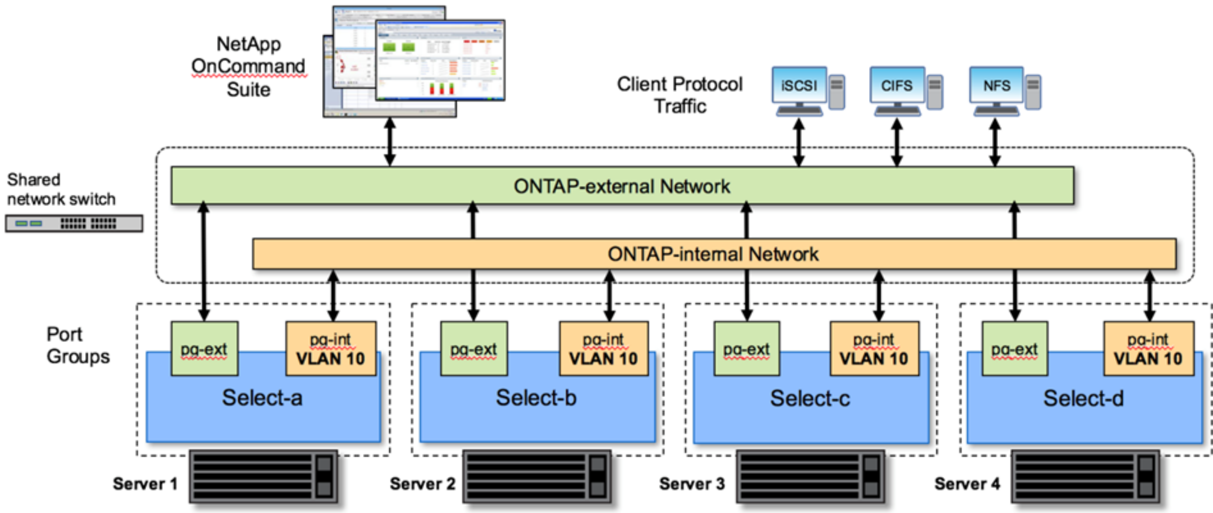
4.1 Network Configuration: Multinode

The multinode ONTAP Select network configuration consists of two networks. These are an internal network, responsible for providing cluster and internal replication services, and an external network, responsible for providing data access and management services. End-to-end isolation of traffic that flows within these two networks is extremely important in allowing you to build an environment that is suitable for cluster resiliency.

These networks are represented in Figure 16, which shows a four-node ONTAP Select cluster running on a VMware vSphere platform. Six- and eight-node clusters have a similar network layout.

Note: Each ONTAP Select instance resides on a separate physical server. Internal and external traffic is isolated using separate network port groups, which are assigned to each virtual network interface and allow the cluster nodes to share the same physical switch infrastructure.

Figure 16) Overview of an ONTAP Select multinode cluster network configuration.



Each ONTAP Select VM contains seven virtual network adapters (six adapters in versions before ONTAP Select 9.3) presented to ONTAP as a set of seven network ports, e0a through e0g. Although ONTAP treats these adapters as physical NICs, they are in fact virtual and map to a set of physical interfaces through a virtualized network layer. As a result, each hosting server does not require six physical network ports.

Note: Adding virtual network adapters to the ONTAP Select VM is not supported.

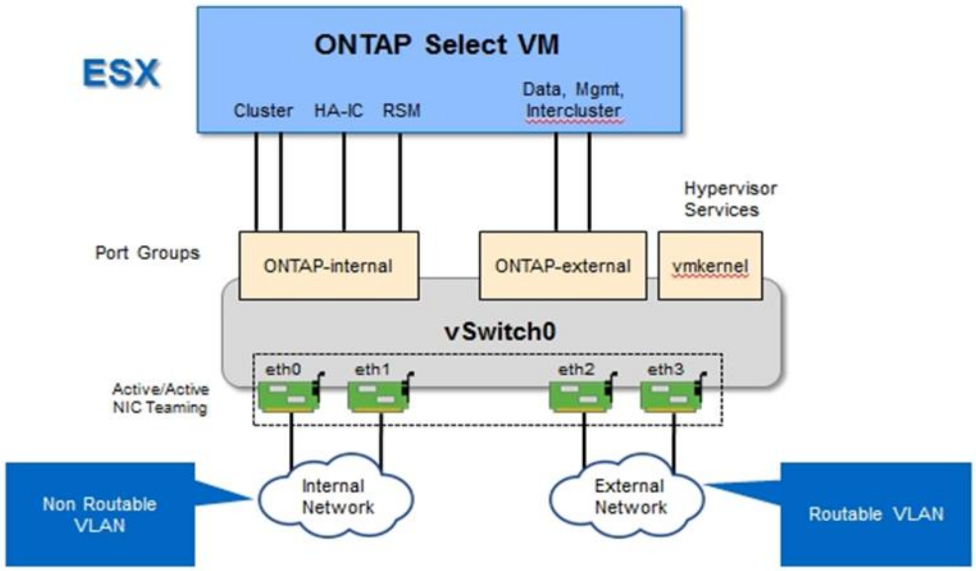
These ports are preconfigured to provide the following services:

- **e0a, e0b, and e0g.** Management and data LIFs
- **e0c, e0d.** Cluster network LIFs
- **e0e.** RSM
- **e0f.** HA interconnect

Ports e0a, e0b, and e0g reside on the external network. Although ports e0c through e0f perform several different functions, collectively they compose the internal Select network. When making network design decisions, these ports should be placed on a single layer-2 network. There is no need to separate these virtual adapters across different networks.

The relationship between these ports and the underlying physical adapters is illustrated in Figure 17, which depicts one ONTAP Select cluster node on the ESX hypervisor.

Figure 17) Network configuration of a single node that is part of a multinode ONTAP Select cluster.



Segregating internal and external traffic across different physical NICs prevents latencies from being introduced into the system due to insufficient access to network resources. Additionally, aggregation through NIC teaming makes sure that failure of a single network adapter does not prevent the ONTAP Select cluster node from accessing the respective network.

Note that both the external network and internal network port groups contain all four NIC adapters in a symmetrical manner. The active ports in the external network port group are the standby ports in the internal network. Conversely, the active ports in the internal network port group are the standby ports in the external network port group.

LIF Assignment

With the introduction of IPspaces, ONTAP port roles have been deprecated. Like FAS arrays, ONTAP Select clusters contain both a default IPspace and a cluster IPspace. By placing network ports e0a, e0b, and e0g into the default IPspace and ports e0c and e0d into the cluster IPspace, those ports have essentially been walled off from hosting LIFs that do not belong. The remaining ports within the ONTAP Select cluster are consumed through the automatic assignment of interfaces providing internal services. They are not exposed through the ONTAP shell, as is the case with the RSM and HA interconnect interfaces.

Note: Not all LIFs are visible through the ONTAP command shell. The HA interconnect and RSM interfaces are hidden from ONTAP and are used internally to provide their respective services.

The network ports and LIFs are explained in detail in the following sections.

Management and Data LIFs (e0a, e0b, and e0g)

ONTAP ports e0a, e0b, and e0g are delegated as candidate ports for LIFs that carry the following types of traffic:

- SAN/NAS protocol traffic (CIFS, NFS, and iSCSI)
- Cluster, node, and SVM management traffic
- Intercluster traffic (SnapMirror and SnapVault)

Note: Cluster and node management LIFs are automatically created during ONTAP Select cluster setup. The remaining LIFs can be created post deployment.

Cluster Network LIFs (e0c, e0d)

ONTAP ports e0c and e0d are delegated as home ports for cluster interfaces. Within each ONTAP Select cluster node, two cluster interfaces are automatically generated during ONTAP setup using link local IP addresses (169.254.x.x).

Note: These interfaces cannot be assigned static IP addresses, and additional cluster interfaces should not be created.

Cluster network traffic must flow through a low-latency, nonrouted layer-2 network. Due to cluster throughput and latency requirements, the ONTAP Select cluster is expected to be physically located within proximity (for example, multipack, single data center). Building four-node, six-node, or eight-node stretch cluster configurations by separating HA nodes across a WAN or across significant geographical distances is not supported. A stretched two-node configuration with a mediator is supported.

For details, see the section "MetroCluster Software Defined Storage (Two-Node Stretched Cluster High Availability)."

Note: To make sure of maximum throughput for cluster network traffic, this network port is configured to use jumbo frames (7500 to 9000 MTU). For proper cluster operation, verify that jumbo frames are enabled on all upstream virtual and physical switches providing internal network services to ONTAP Select cluster nodes.

RAID SyncMirror Traffic (e0e)

Synchronous replication of blocks across HA partner nodes occurs using an internal network interface residing on network port e0e. This functionality occurs automatically, using network interfaces configured by ONTAP during cluster setup, and requires no configuration by the administrator.

Note: Port e0e is reserved by ONTAP for internal replication traffic. Therefore, neither the port nor the hosted LIF is visible in the ONTAP CLI or in System Manager. This interface is configured to use an automatically generated link local IP address, and the reassignment of an alternate IP address is not supported. This network port requires the use of jumbo frames (7500 to 9000 MTU).

HA Interconnect (e0f)

NetApp FAS arrays use specialized hardware to pass information between HA pairs in an ONTAP cluster. Software-defined environments, however, do not tend to have this type of equipment available (such as InfiniBand or iWARP devices), so an alternate solution is needed. Although several possibilities were considered, ONTAP requirements placed on the interconnect transport required that this functionality be emulated in software. As a result, within an ONTAP Select cluster, the functionality of the HA interconnect (traditionally provided by hardware) has been designed into the OS, using Ethernet as a transport mechanism.

Each ONTAP Select node is configured with an HA interconnect port, e0f. This port hosts the HA interconnect network interface, which is responsible for two primary functions:

- Mirroring the contents of NVRAM between HA pairs
- Sending/receiving HA status information and network heartbeat messages between HA pairs

HA interconnect traffic flows through this network port using a single network interface by layering remote direct memory access (RDMA) frames within Ethernet packets.

Note: In a manner similar to the RSM port (e0e), neither the physical port nor the hosted network interface is visible to users from either the ONTAP CLI or from System Manager. As a result, the

IP address of this interface cannot be modified, and the state of the port cannot be changed. This network port requires the use of jumbo frames (7500 to 9000 MTU).

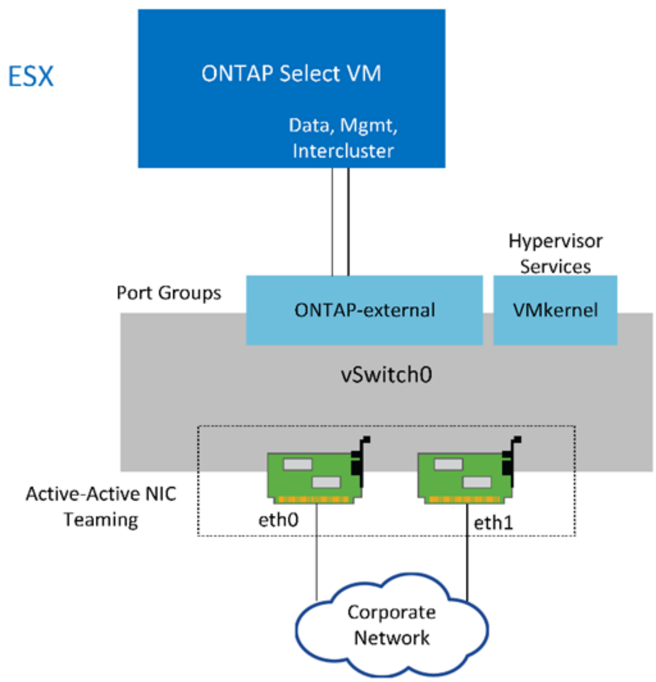
4.2 Network Configuration: Single Node

Single-node ONTAP Select configurations do not require the ONTAP internal network, because there is no cluster, HA, or mirror traffic. Unlike the multinode version of the ONTAP Select product, each ONTAP Select VM contains three virtual network adapters (two for releases before ONTAP Select 9.3), presented to ONTAP network ports e0a, e0b, and e0c.

These ports are used to provide the following services: management, data, and intercluster LIFs.

The relationship between these ports and the underlying physical adapters can be seen in Figure 18, which depicts one ONTAP Select cluster node on the ESX hypervisor.

Figure 18) Network configuration of single-node ONTAP Select cluster.



Note: Even though two adapters are sufficient for a single-node cluster, NIC teaming is still required.

LIF Assignment

As explained in the multinode LIF assignment section of this document, IPspaces are used by ONTAP Select to keep cluster network traffic separate from data and management traffic. The single-node variant of this platform does not contain a cluster network. Therefore, no ports are present in the cluster IPspace.

Note: Cluster and node management LIFs are automatically created during ONTAP Select cluster setup. The remaining LIFs can be created post deployment.

Management and Data LIFs (e0a, e0b, and e0c)

ONTAP ports e0a, e0b, and e0g are delegated as candidate ports for LIFs that carry the following types of traffic:

- SAN/NAS protocol traffic (CIFS, NFS, and iSCSI)

- Cluster, node, and SVM management traffic
- Intercluster traffic (SnapMirror and SnapVault)

4.3 Networking: Internal and External

ONTAP Select Internal Network

The internal ONTAP Select network, which is only present in the multinode variant of the product, is responsible for providing the ONTAP Select cluster with cluster communication, HA interconnect, and synchronous replication services. This network includes the following ports and interfaces:

- **e0c, e0d.** Hosting cluster network LIFs
- **e0e.** Hosting the RSM LIF
- **e0f.** Hosting the HA interconnect LIF

The throughput and latency of this network are critical in determining the performance and resiliency of the ONTAP Select cluster. Network isolation is required for cluster security and to make sure that system interfaces are kept separate from other network traffic. Therefore, this network must be used exclusively by the ONTAP Select cluster.

Note: Using the Select internal network for traffic other than Select cluster traffic, such as application or management traffic, is not supported. There can be no other VMs or hosts on the ONTAP internal VLAN.

Network packets traversing the internal network must be on a dedicated VLAN-tagged layer-2 network. This can be accomplished by completing one of the following tasks:

- Assigning a VLAN-tagged port group to the internal virtual NICs (e0c through e0f) (VST mode)
- Using the native VLAN provided by the upstream switch where the native VLAN is not used for any other traffic (assign a port group with no VLAN ID, that is, EST mode)

In all cases, VLAN tagging for internal network traffic is done outside of the ONTAP Select VM.

Note: Only ESX standard and distributed vSwitches are supported. Other virtual switches or direct connectivity between ESX hosts are not supported. The internal network must be fully opened; NAT or firewalls are not supported.

Within an ONTAP Select cluster, internal traffic and external traffic are separated using virtual layer-2 network objects known as port groups. Proper vSwitch assignment of these port groups is extremely important, especially for the internal network, which is responsible for providing cluster, HA interconnect, and mirror replication services. Insufficient network bandwidth to these network ports can cause performance degradation and even affect the stability of the cluster node. Therefore, four-node, six-node, and eight-node clusters require that the internal ONTAP Select network use 10Gb connectivity; 1Gb NICs are not supported. Tradeoffs can be made to the external network, however, because limiting the flow of incoming data to an ONTAP Select cluster does not affect its ability to operate reliably.

A two-node cluster can use either four 1Gb ports for internal traffic or a single 10Gb port instead of the two 10Gb ports required by the four-node cluster. In an environment in which conditions prevent the server from being fit with four 10Gb NIC cards, two 10Gb NIC cards can be used for the internal network and two 1Gb NICs can be used for the external ONTAP network.

Internal Network Validation and Troubleshooting

Starting with Deploy 2.2, the internal network in a multinode cluster can be validated by using the network connectivity checker functionality. This function can be invoked from the Deploy CLI running the `network connectivity-check start` command.

Run the `network connectivity-check show --run-id X (X` is a number) command to view the output of the test.

This tool is only useful for troubleshooting the internal network in a multinode Select cluster. The tool should not be used to troubleshoot single-node clusters (including vNAS configurations), ONTAP Deploy to ONTAP Select connectivity, or client-side connectivity issues.

Starting with Deploy 2.5, the cluster create wizard (part of the ONTAP Deploy GUI) includes the internal network checker as an optional step available during the creation of multinode clusters. Given the important role that the internal network plays in multinode clusters, making this step part of the cluster create workflow improves the success rate of cluster create operations.

Starting with ONTAP Deploy 2.10, the MTU size used by the internal network can be set between 7,500 and 9,000. The network connectivity checker can also be used to test MTU size between 7,500 and 9,000. The default MTU value is set to the value of the virtual network switch. That default would have to be replaced with a smaller value if a network overlay like VXLAN is present in the environment.

ONTAP Select External Network

The ONTAP Select external network is responsible for all outbound communications by the cluster and, therefore, is present on both the single-node and multinode configurations. Although this network does not have the tightly defined throughput requirements of the internal network, the administrator should be careful not to create network bottlenecks between the client and ONTAP VM, because performance issues could be mischaracterized as ONTAP Select problems.

Note: In a manner similar to internal traffic, external traffic can be tagged at the vSwitch layer (VST) and at the external switch layer (EST). In addition, the external traffic can be tagged by the ONTAP Select VM itself in a process known as VGT. See the section "Data and Management Separation" for further details.

Table 6 highlights the major differences between the ONTAP Select internal and external networks.

Table 6) Internal versus external network quick reference.

Description	Internal Network	External Network
Network services	• Cluster • HA/IC • RAID SyncMirror (RSM)	• Data management • Intercluster (SnapMirror and SnapVault)
Network isolation	Required	Optional
Frame size (MTU)	7,500 to 9,000[2]	• 1,500 (default) • 9,000 (supported)
IP address assignment	Autogenerated	User-defined
DHCP support	No	No

[1]ONTAP Select 9.3 supports a single 10Gb link for two-node clusters; however, it is a NetApp best practice to make sure of hardware redundancy through NIC aggregation.

[2]Requires ONTAP Select 9.5 and ONTAP Deploy 2.10.

NIC Teaming

To make sure that the internal and external networks have both the necessary bandwidth and resiliency characteristics required to provide high performance and fault tolerance, physical network adapter teaming is recommended. Starting with ONTAP Select 9.3, two-node cluster configurations with a single

10Gb link are supported. However, the NetApp recommended best practice is to make use of NIC teaming on both the internal and the external networks of the ONTAP Select cluster.

Best Practice
If a NIC has multiple application-specific integrated circuits (ASICs), select one network port from each ASIC when building network constructs through NIC teaming for the internal and external networks.

MAC Address Generation

The MAC addresses assigned to all ONTAP Select network ports are generated automatically by the included deployment utility. The utility uses a platform-specific, organizationally unique identifier (OUI) specific to NetApp to make sure there is no conflict with FAS systems. A copy of this address is then stored in an internal database within the ONTAP Select installation VM (ONTAP Deploy), to prevent accidental reassignment during future node deployments. At no point should the administrator modify the assigned MAC address of a network port.

4.4 Supported Network Configurations

Server vendors understand that customers have different needs and choice is critical. As a result, when purchasing a physical server, there are numerous options available when making network connectivity decisions. Most commodity systems ship with various NIC choices that provide single-port and multiport options with varying permutations of 1Gb and 10Gb ports. Care should be taken when selecting server NICs because the choices provided by server vendors can have a significant impact on the overall performance of the ONTAP Select cluster.

Link aggregation is a common network construct used to aggregate bandwidth across multiple physical adapters. LACP is a vendor-neutral standard that provides an open protocol for network endpoints that bundle groupings of physical network ports into a single logical channel. ONTAP Select can work with port groups that are configured as a Link Aggregation Group (LAG). However, starting with ONTAP Select 9.6 and ONTAP Deploy 2.12, NetApp recommends using the individual physical ports as simple uplink (trunk) ports, avoiding the LAG configuration. For both standard and distributed vSwitches, the best practices listed in Table 7 should be followed:

Table 7) Network configuration support matrix.

Server Environment	Select Configuration	Best Practices
• Standard or distributed vSwitch • 4 x 10Gb ports or • 4 x 1Gb ports • The physical uplink switch does not support or is not configured for LACP and supports large MTU size on all ports.[1]	• Do not use any LACP channels. • All the ports must be owned by the same vSwitch. The vSwitch must support a large MTU size.[1] • ONTAP Deploy 2.12 supports configurations with up to four port groups, two for the internal network and two for the external network. For best performance, all four port groups should be used. The procedure to switch from a single port group per network to two port groups per network is detailed in Section 4.5.	• The load-balancing policy at the port-group level is Route Based on Originating Virtual Port ID. • VMware recommends that STP be set to Portfast on the switch ports connected to the ESXi hosts.
• Standard or distributed vSwitch • 2 x 10Gb ports • The physical uplink switch does not support or is not configured for LACP and supports large MTU size on all ports.[1]	• Do not use any LACP channels. • The internal network must use a port group with 1 x 10Gb active and 1 x 10Gb standby.[1] • The external network uses a separate port group. The active port is the standby port for the internal port group. The standby port is the active port for the internal network port group. • All the ports must be owned by the same vSwitch. The vSwitch must support a large MTU size.[1]	• The load-balancing policy at the port group level is Route Based on Originating Virtual Port ID. • VMware recommends that the STP be set to Portfast on the switch ports connected to the ESXi hosts.

[1]Starting with ONTAP Select 9.5 and ONTAP Deploy 2.10, the internal network supports an MTU size between 7,500 and 9,000.

Because the performance of the ONTAP Select VM is tied directly to the characteristics of the underlying hardware, increasing the throughput to the VM by selecting 10Gb-capable NICs results in a higher-performing cluster and a better overall user experience. When cost or form factor prevents the user from designing a system with four 10Gb NICs, two 10Gb NICs can be used. There are a number of other configurations that are also supported. For two-node clusters, 4 x 1Gb ports or 1 x 10Gb ports are supported. For single node clusters, 2 x 1Gb ports are supported. See Table 8 for minimum requirements and recommendations.

Table 8) Network minimum and recommended configurations.

	Minimum Requirements	Recommendations
Single node clusters	2 x 1Gb	2 x 10Gb
Two node clusters / Metrocluster SDS	4 x 1Gb or 1 x 10Gb	2 x 10Gb

	Minimum Requirements	Recommendations
4/6/8 node clusters	2 x 10Gb	4 x 10Gb

4.5 VMware vSphere: vSwitch Configuration

ONTAP Select supports the use of both standard and distributed vSwitch configurations. This section describes the vSwitch configuration and load-balancing policies that should be used in both two-NIC and four-NIC configurations.

Standard vSwitch

All vSwitch configurations require a minimum of two physical network adapters bundled into a single NIC team. ONTAP Select 9.3 and newer releases support a single 10Gb link for two-node clusters. However, it is a NetApp best practice to make sure of hardware redundancy through NIC aggregation.

On a vSphere server, NIC teams are the aggregation construct used to bundle multiple physical network adapters into a single logical channel, allowing the network load to be shared across all member ports. It's important to remember that NIC teams can be created without support from the physical switch. Load-balancing and failover policies can be applied directly to a NIC team, which is unaware of the upstream switch configuration. In this case, policies are only applied to outbound traffic.

Note: Static port channels are not supported with ONTAP Select. LACP-enabled channels are supported with distributed vSwitches but using LACP LAGs may result in un-even load distribution across the LAG members.

Best Practice
To optimize load balancing across both the internal and the external ONTAP Select networks, use the Route Based on Originating Virtual Port load-balancing policy.

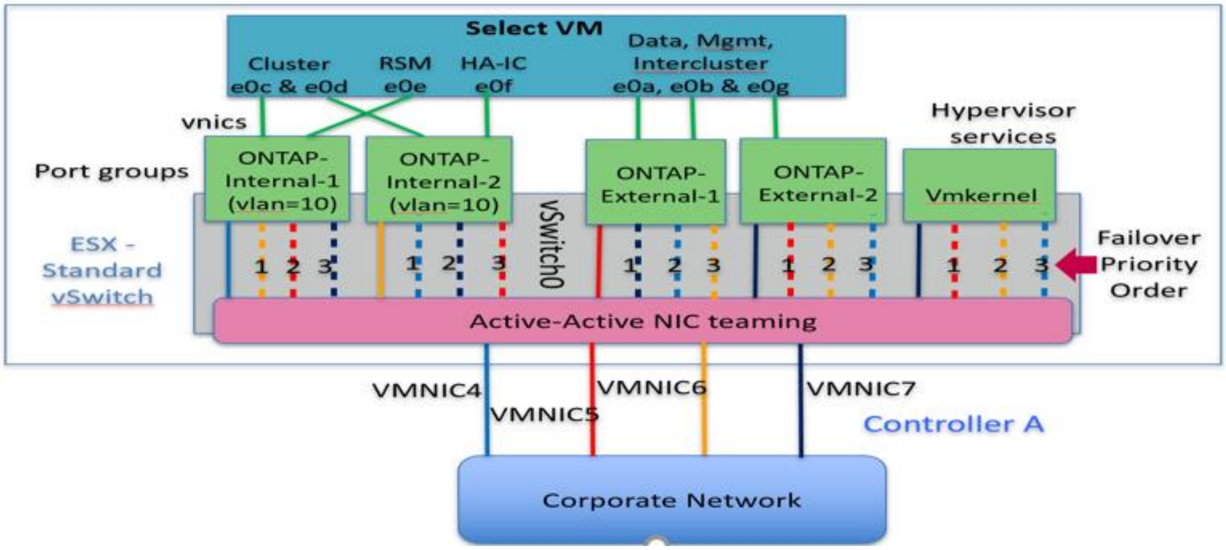
For single node clusters, ONTAP Deploy configures the ONTAP Select VM to use a port group for the external network and either the same port group or, optionally, a different port group for the cluster and node management traffic. For single node clusters, the desired number of physical ports can be added to the external port group as active adapters.

For multinode clusters, ONTAP Deploy configures each ONTAP Select VM to use one or two port groups for the internal network and separately, one or two port groups for the external network. The cluster and node management traffic can either use the same port group as the external traffic, or optionally a separate port group. The cluster and node management traffic cannot share the same port group with internal traffic.

Standard or Distributed vSwitch and Four Physical ports per Node

Starting with ONTAP Select 9.6 and ONTAP Deploy 2.12, four port groups can be assigned to each node in a multinode cluster. Each port group has a single active physical port and three standby physical ports as in Figure 19.

Figure 19) vSwitch with four physical ports per node.



The order of the ports in the standby list is important. Table 9 provides an example of the physical port distribution across the four port groups.

Table 9) Network minimum and recommended configurations.

Port Group	External 1	External 2	Internal 1	Internal 2
Active	vmnic0	vmnic1	vmnic2	vmnic3
Standby 1	vmnic1	vmnic0	vmnic3	vmnic2
Standby 2	vmnic2	vmnic3	vmnic0	vmnic1
Standby 3	vmnic3	vmnic2	vmnic1	vmnic0

Figure 20 and Figure 21 show the configurations of the external network port groups from the vCenter GUI (ONTAP-External and ONTAP-External2). Note that the active adapters are from different network cards. In this setup, vmnic 4 and vmnic 5 are dual ports on the same physical NIC, while vmnic 6 and vminc 7 are similarly dual ports on a separate NIC (vnmics 0 through 3 are not used in this example).The order of the standby adapters provides a hierarchical fail over with the ports from the internal network being last. The order of internal ports in the standby list is similarly swapped between the two external port groups.

Figure 20) Part 1: ONTAP Select external port group configurations.

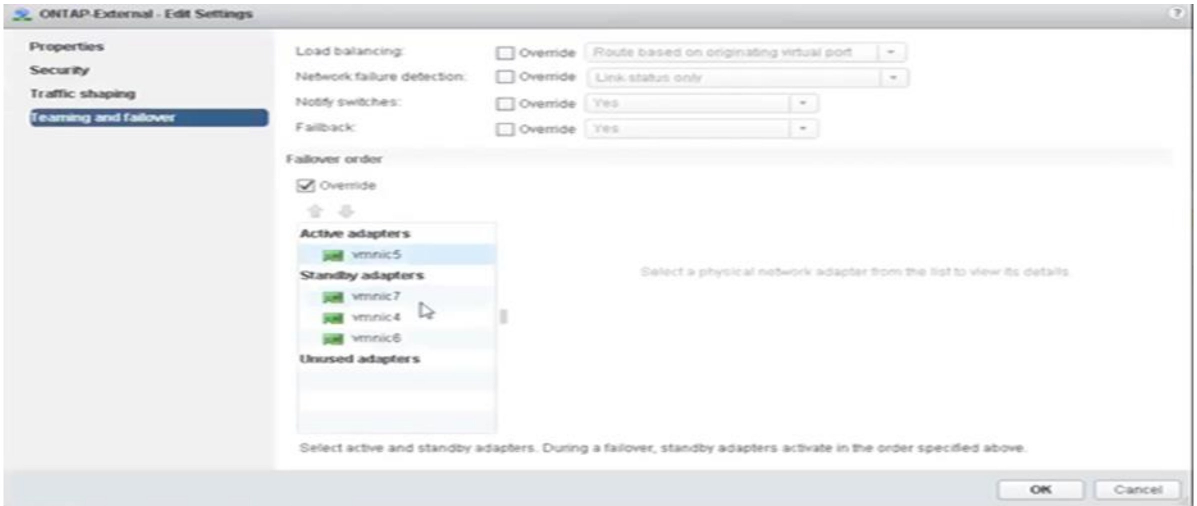
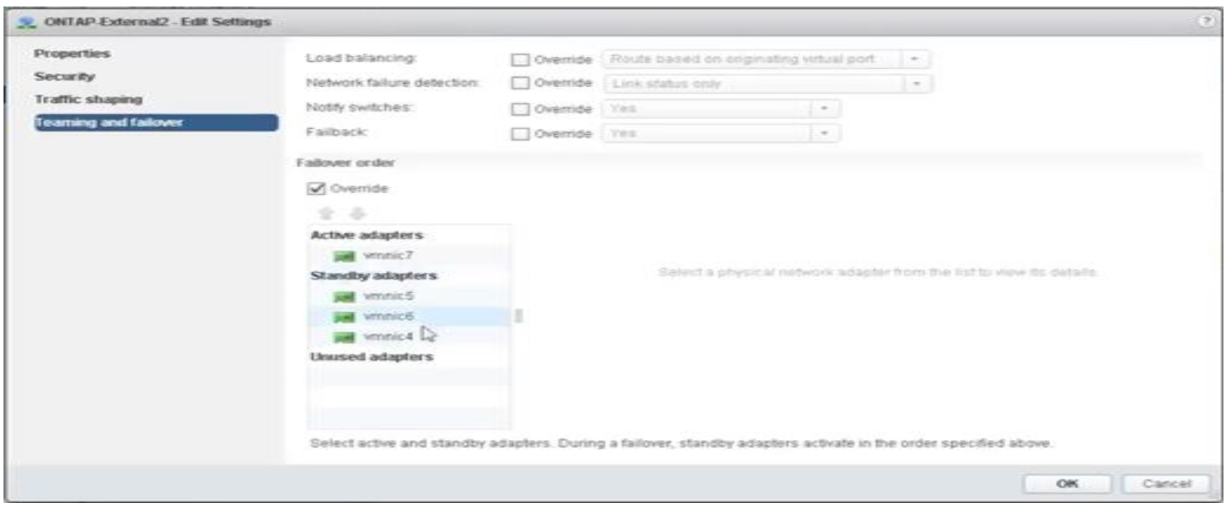


Figure 21) Part 2: ONTAP Select external port group configurations.



For readability, the assignments are as follows:

ONTAP-External	ONTAP-External2
Active adapters: vmnic5 Standby adapters: vmnic7, vmnic4, vmnic6	Active adapters: vmnic7 Standby adapters: vmnic5, vmnic6, vmnic4

Figure 22 and Figure 23 shows the configurations of the internal network port groups (ONTAP-Internal and ONTAP-Internal2). Note that the active adapters are from different network cards. In this setup, vmnic 4 and vmnic 5 are dual ports on the same physical ASIC, whereas vmnic 6 and vmnic 7 are similarly dual ports on a separate ASIC. The order of the standby adapters provides a hierarchical fail over with the ports from the external network being last. The order of external ports in the standby list is similarly swapped between the two internal port groups.

Figure 22) Part 1: ONTAP Select internal port group configurations.

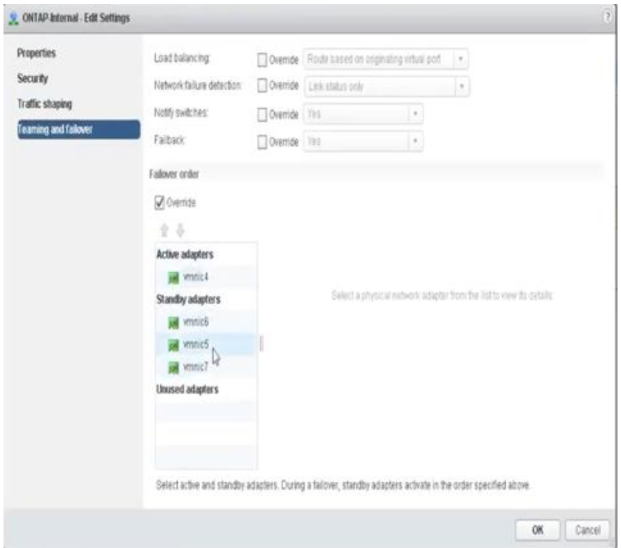
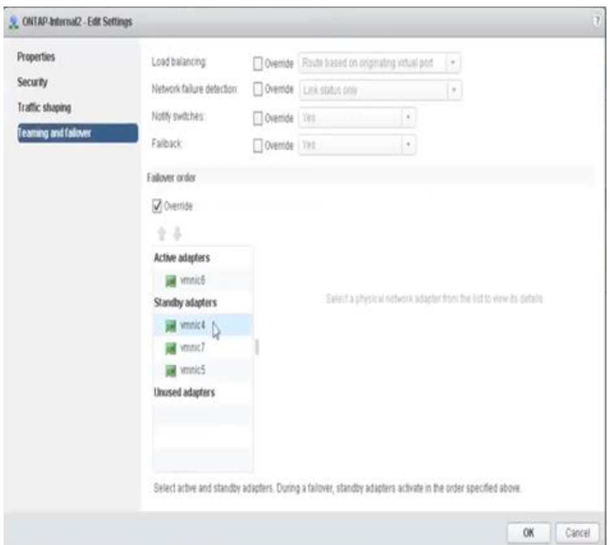


Figure 23) Part 2: ONTAP Select internal port groups.



For readability, the assignments are as follows:

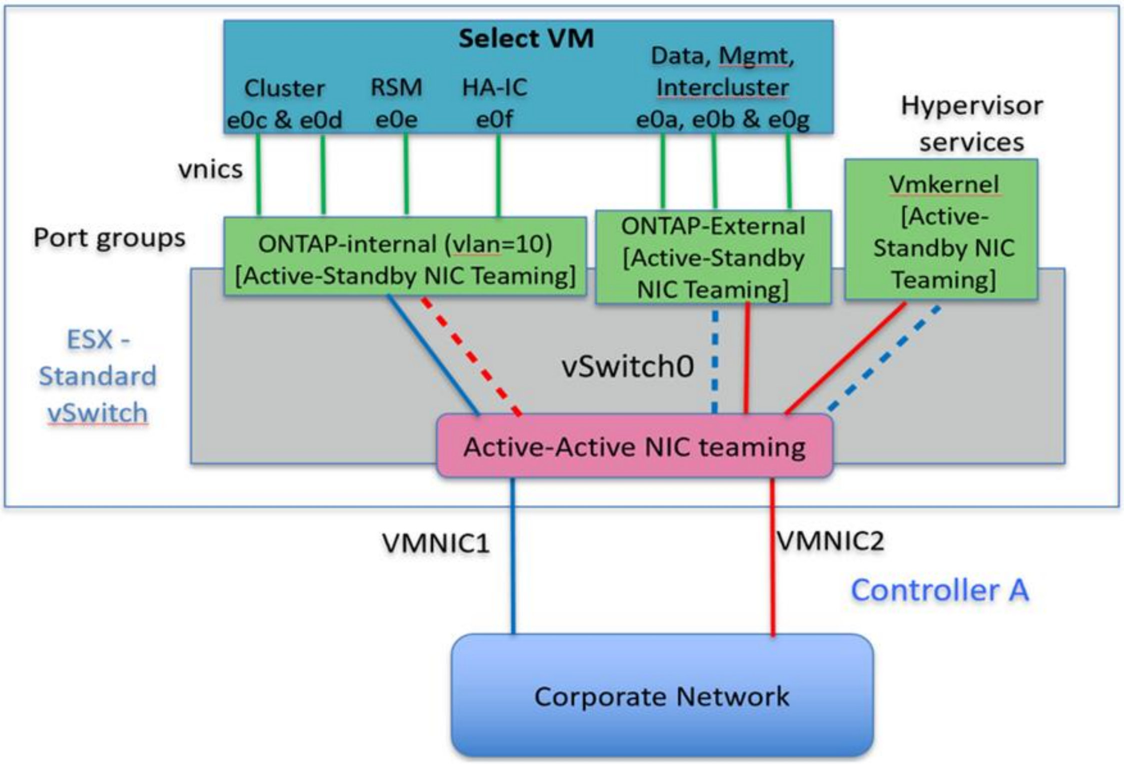
ONTAP-Internal	ONTAP-Internal2
Active adapters: vmnic4 Standby adapters: vmnic6, vmnic5, vmnic7	Active adapters: vmnic6 Standby adapters: vmnic4, vmnic7, vmnic5

Standard or Distributed vSwitch and Two Physical Ports per Node

When using two physical ports, each port group should have an active adapter and a standby adapter configured opposite to each other. The internal network is only present for multinode ONTAP Select clusters. For single-node clusters, both adapters can be configured as active in the external port group.

The following example shows the configuration of a vSwitch and the two port groups responsible for handling internal and external communication services for a multinode ONTAP Select cluster. The external network can use the internal network VMNIC in the event of a network outage because the internal network VMNICs are part of this port group and configured in standby mode. The opposite is the case for the external network. Alternating the active and standby VMNICs between the two port groups is critical for the proper failover of the ONTAP Select VMs during network outages.

Figure 24) vSwitch with two physical ports per node.



Distributed vSwitch with LACP

When using distributed vSwitches in your configuration, LACP can be used (though it is not a best practice) in order to simplify the network configuration. The only supported LACP configuration requires that all the VMNICs are in a single LAG. The uplink physical switch must support an MTU size between 7,500 to 9,000 on all the ports in the channel. The internal and external ONTAP Select networks should be isolated at the port group level. The internal network should use a nonroutable (isolated) VLAN. The external network can use either VST, EST, or VGT.

The following examples show the distributed vSwitch configuration using LACP.

Figure 25) LAG properties when using LACP.

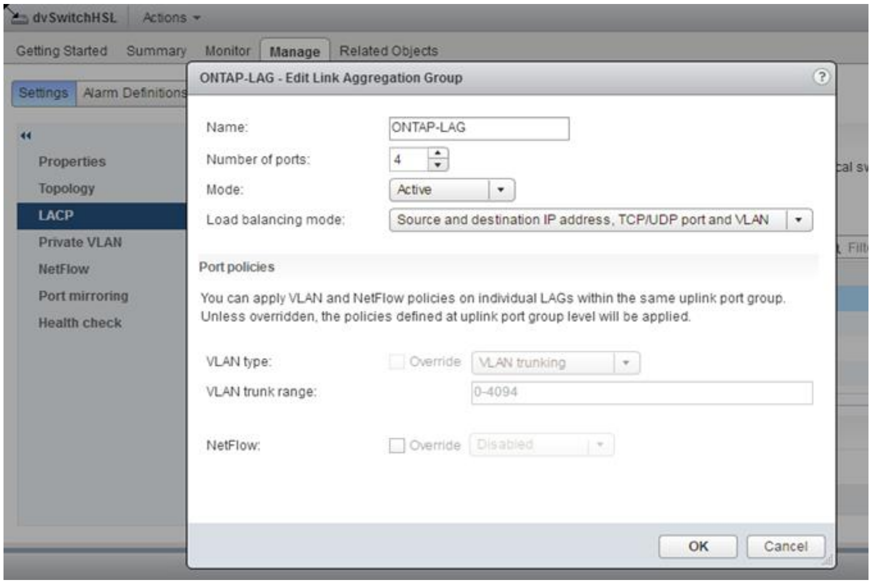


Figure 26) External port group configurations using a distributed vSwitch with LACP enabled.

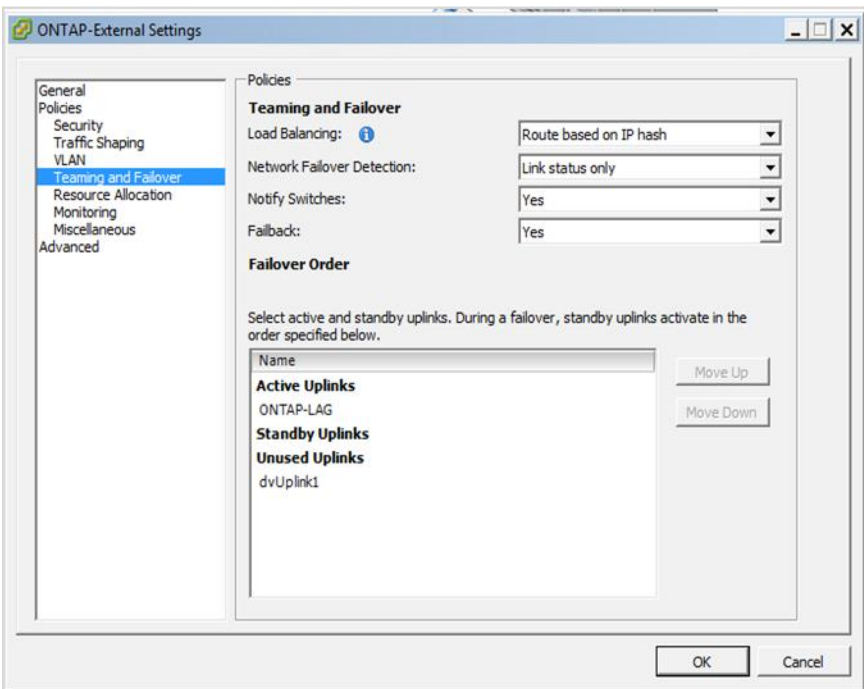
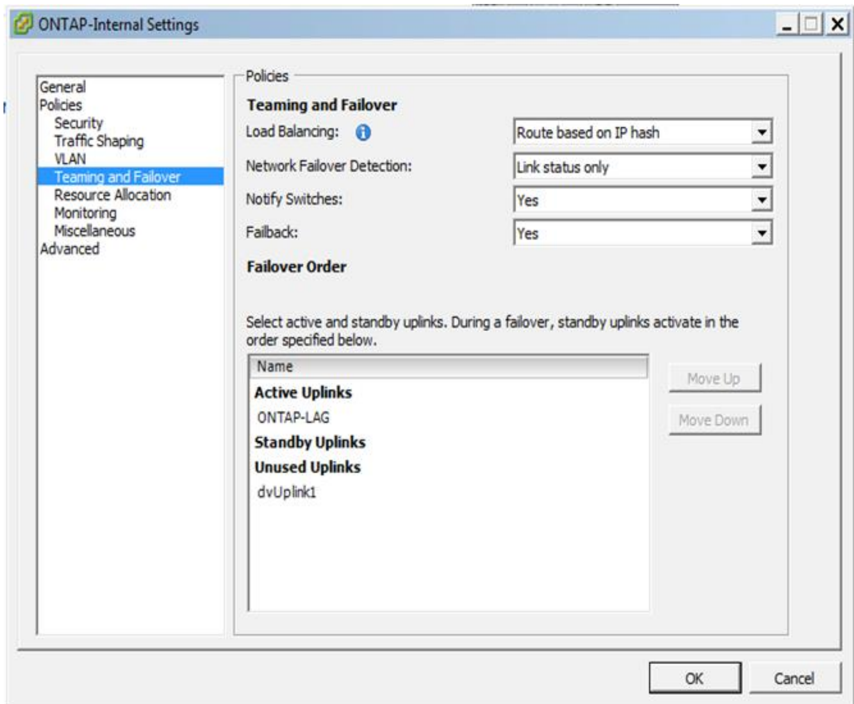


Figure 27) Internal port group configurations using a distributed vSwitch with LACP enabled.



Note: LACP requires that you configure the upstream switch ports as a port channel. Prior to enabling this on the distributed vSwitch, make sure that an LACP-enabled port channel is properly configured.

Best Practice

NetApp recommends that the LACP mode be set to active on both the ESX and the physical switches. Furthermore, the LACP timer should be set to fast (1 second) on the physical switch, ports, port channel interfaces and on the VMNICs.

When using a distributed vSwitch with LACP, NetApp recommends that you configure the load-balancing policy to Route Based on IP Hash on the port group and Source and Destination IP Address and TCP/UDP Port and VLAN on the LAG.

4.6 Physical Switch Configuration

Careful consideration should be taken when making connectivity decisions from the virtual switch layer to physical switches. Separation of internal cluster traffic from external data services should extend to the upstream physical networking layer through isolation provided by layer-2 VLANs.

This section covers upstream physical switch configurations based on single-switch and multiswitch environments.

Physical switch ports should be configured as trunkports. ONTAP Select external traffic can be separated across multiple layer-2 networks in one of two ways. One method is by using ONTAP VLAN-tagged virtual ports with a single port group. The other method is by assigning separate port groups in VST mode to management port e0a. You must also assign data ports to e0b and e0c/e0g depending on the ONTAP Select release and the single-node or multinode configuration. If the external traffic is separated across multiple layer-2 networks, the uplink physical switch ports should have those VLANs in its allowed VLAN list.

ONTAP Select internal network traffic occurs using virtual interfaces defined with link local IP addresses. Because these IP addresses are nonroutable, internal traffic between cluster nodes must flow across a single layer-2 network. Route hops between ONTAP Select cluster nodes are unsupported.

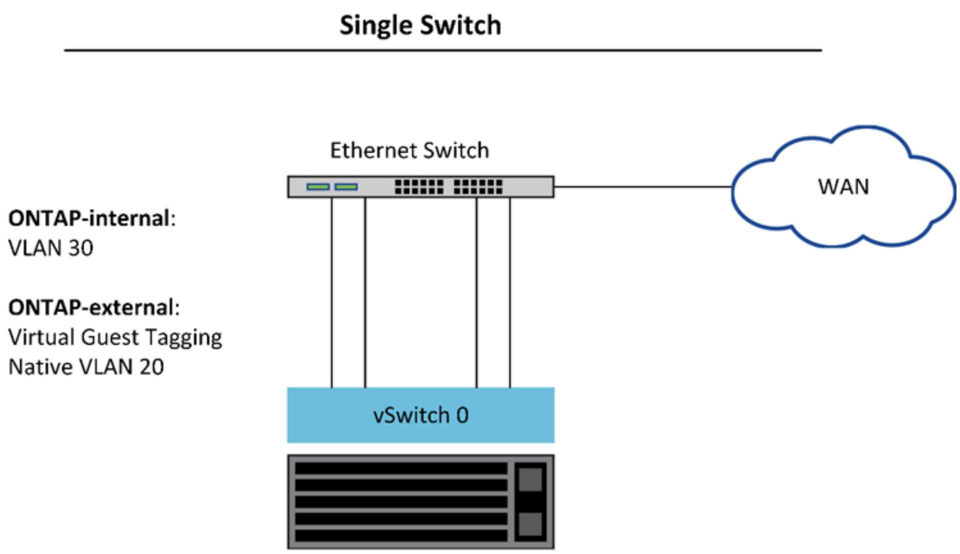
Best Practice
VMware recommends that STP be set to Portfast on the switch ports connected to the ESXi hosts. Not setting STP to Portfast on the switch ports can affect ONTAP Select's ability to tolerate uplink failures. When using LACP, the LACP timer should be set to fast (1 second). The load-balancing policy should be set to Route Based on IP Hash on the port group and Source and Destination IP Address and TCP/UDP port and VLAN on the LAG.

Shared Physical Switch

Figure 28 depicts a possible switch configuration used by one node in a multinode ONTAP Select cluster. In this example, the physical NICs used by the vSwitches hosting both the internal and external network port groups are cabled to the same upstream switch. Switch traffic is kept isolated using broadcast domains contained within separate VLANs.

Note: For the ONTAP Select internal network, tagging is done at the port group level. While the following example uses VGT for the external network, both VGT and VST are supported on that port group.

Figure 28) Network configuration using shared physical switch.



Note: In this configuration, the shared switch becomes a single point of failure. If possible, multiple switches should be used to prevent a physical hardware failure from causing a cluster network outage.

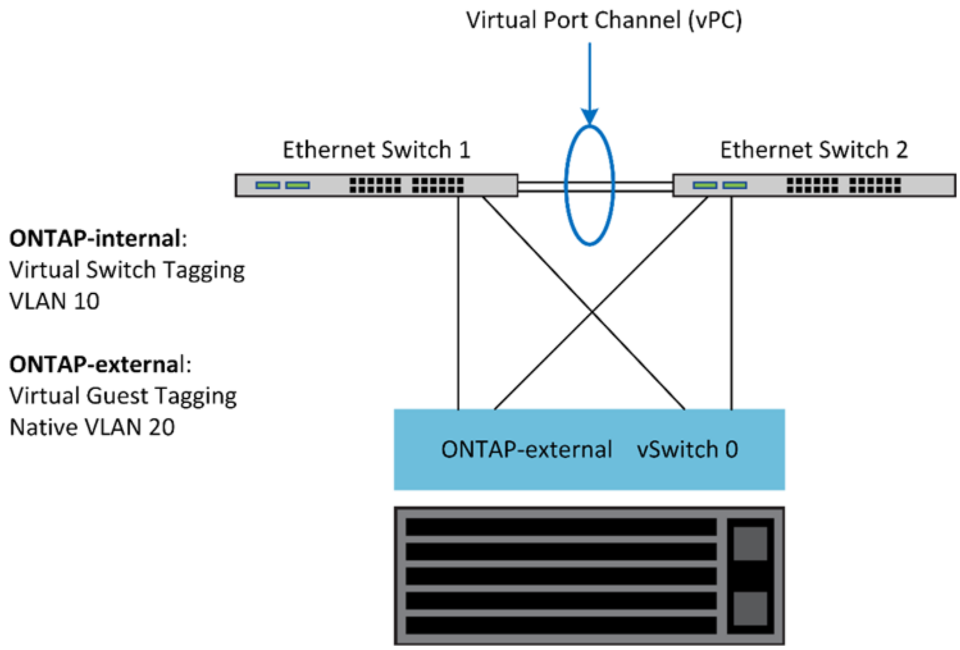
Multiple Physical Switches

When redundancy is needed, multiple physical network switches should be used. Figure 29 shows a recommended configuration used by one node in a multinode ONTAP Select cluster. NICs from both the internal and external port groups are cabled into different physical switches, protecting the user from a single hardware-switch failure. A virtual port channel is configured between switches to prevent spanning tree issues.

Best Practice

When sufficient hardware is available, NetApp recommends using the multiswitch configuration shown in Figure 29, due to the added protection against physical switch failures.

Figure 29) Network configuration using multiple physical switches.



4.7 Data and Management Separation

ONTAP Select external network traffic is defined as data (CIFS, NFS, and iSCSI), management, and replication (SnapMirror) traffic. Within an ONTAP cluster, each style of traffic uses a separate logical interface that must be hosted on a virtual network port. On the multinode configuration of ONTAP Select, these are designated as ports e0a and e0b/e0g. On the single node configuration, these are designated as e0a and e0b/e0c, while the remaining ports are reserved for internal cluster services.

NetApp recommends isolating data traffic and management traffic into separate layer-2 networks. In the ONTAP Select environment, this is done using VLAN tags. This can be achieved by assigning a VLAN-tagged port group to network adapter 1 (port e0a) for management traffic. Then you can assign a separate port group(s) to ports e0b and e0c (single-node clusters) and e0b and e0g (multinode clusters) for data traffic.

If the VST solution described earlier in this document is not sufficient, collocating both data and management LIFs on the same virtual port might be required. To do so, use a process known as VGT, in which VLAN tagging is performed by the VM.

Note: Data and management network separation through VGT is not available when using the ONTAP Deploy utility. This process must be performed after cluster setup is complete.

There is an additional caveat when using VGT and two-node clusters. In two-node cluster configurations, the node management IP address is used to establish connectivity to the mediator before ONTAP is fully available. Therefore, only EST and VST tagging is supported on the port group mapped to the node management LIF (port e0a). Furthermore, if both the management and the data traffic are using the same port group, only EST/VST are supported for the entire two-node cluster.

Both configuration options, VST and VGT, are supported. Figure 30 shows the first scenario, VST, in which traffic is tagged at the vSwitch layer through the assigned port group. In this configuration, cluster and node management LIFs are assigned to ONTAP port e0a and tagged with VLAN ID 10 through the assigned port group. Data LIFs are assigned to port e0b and either e0c or e0g and given VLAN ID 20 using a second port group. The cluster ports use a third port group and are on VLAN ID 30.

Figure 30) Data and management separation using VST.

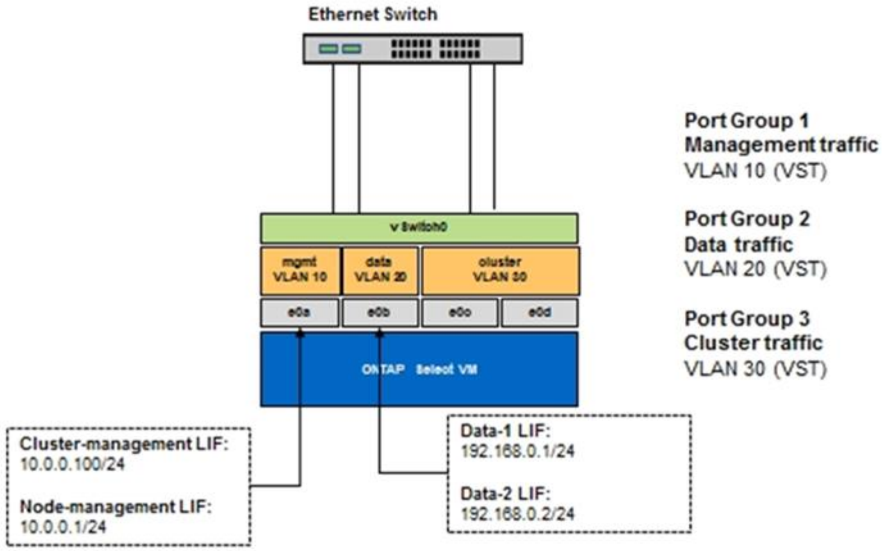
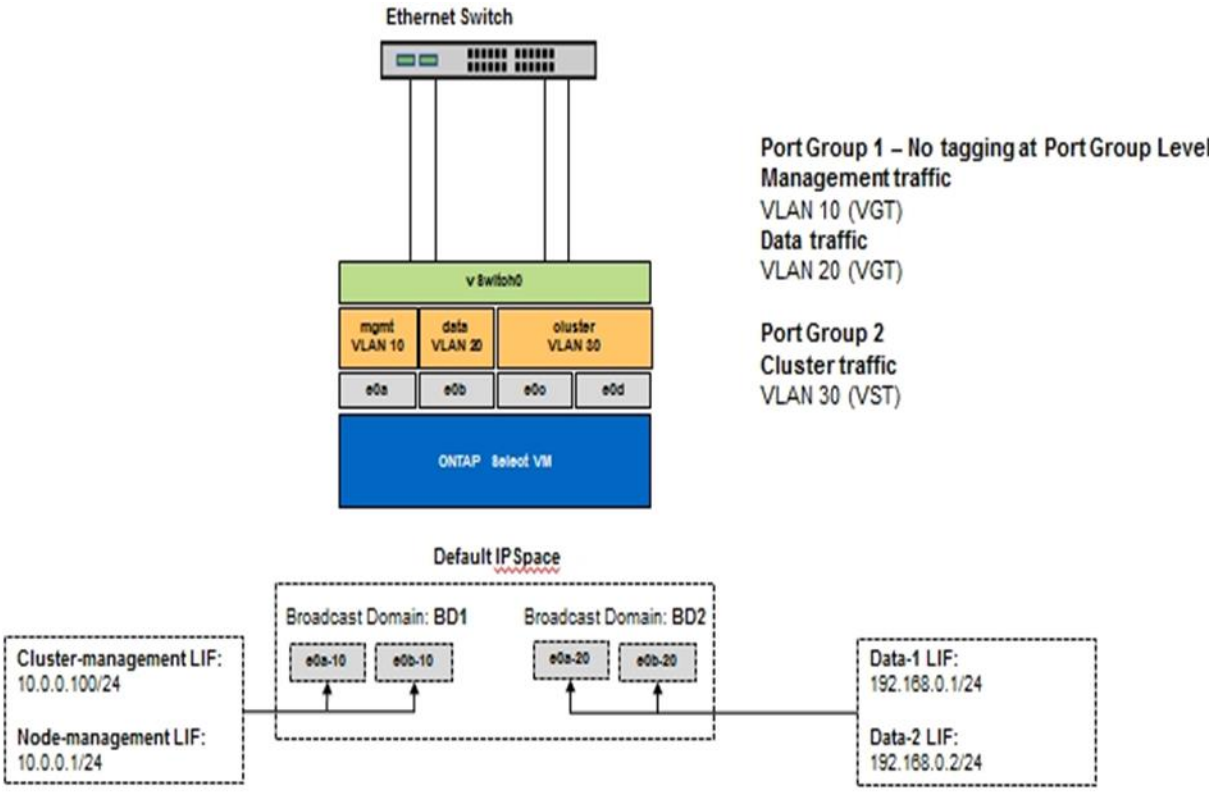


Figure 31 shows the second scenario, VGT, in which traffic is tagged by the ONTAP VM using VLAN ports that are placed into separate broadcast domains. In this example, virtual ports e0a-10/e0b-10/(e0c or e0g)-10 and e0a-20/e0b-20 are placed on top of VM ports e0a and e0b. This configuration allows network tagging to be performed directly within ONTAP, rather than at the vSwitch layer. Management and data LIFs are placed on these virtual ports, allowing further layer-2 subdivision within a single VM port. The cluster VLAN (VLAN ID 30) is still tagged at the port group.

Note: This style of configuration is especially desirable when using multiple IPspaces. Group VLAN ports into separate custom IPspaces if further logical isolation and multitenancy are desired.

Note: To support VGT, the ESXi/ESX host network adapters must be connected to trunk ports on the physical switch. The port groups connected to the virtual switch must have their VLAN ID set to 4095 to enable trunking on the port group.

Figure 31) Data and management separation using VGT.



Best Practice
If data traffic spans multiple layer-2 networks and the use of VLAN ports is required or when you are using multiple IPspaces, VGT should be used.

5 Use Cases

ONTAP Select is a flexible storage management solution that enables various use cases. This section describes some of these use cases.

5.1 Remote and Branch Offices

The ONTAP Select VM can be collocated with application VMs, making it an optimal solution for remote offices or branch offices (ROBOs). Using ONTAP Select to provide enterprise-class file services while allowing bidirectional replication to other ONTAP Select or FAS clusters enables resilient solutions to be built in low-touch or low-cost environments. ONTAP Select comes prepopulated with feature licenses for CIFS, NFS, and iSCSI protocol services as well as both SnapMirror and SnapVault replication technologies. Therefore, all of these features are available immediately upon deployment.

Starting with ONTAP Select 9.2 and ONTAP Deploy 2.4, all vSphere and VSAN licenses are now supported.

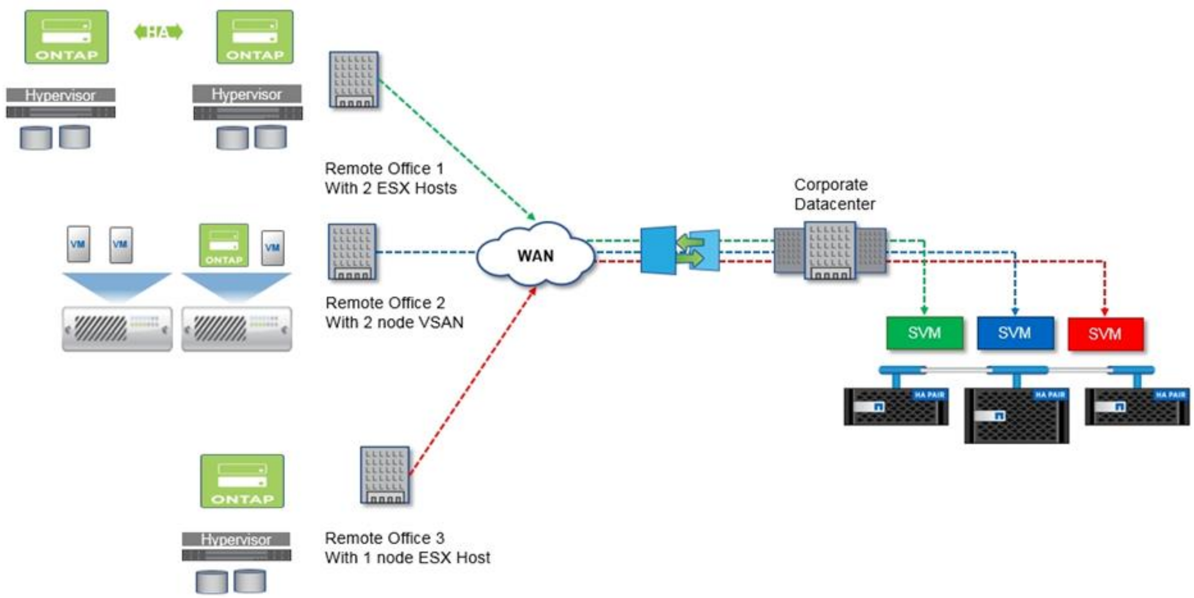
An ONTAP Select two-node cluster with a remote mediator is an attractive solution for small data centers. In this configuration, HA functionality is provided by ONTAP Select. The minimum networking requirement for a two-node ONTAP Select ROBO solution is four 1Gb links. Starting with ONTAP Select 9.3, a single 10Gb network connection is also supported.

The vNAS ONTAP Select solution running on VSAN (including the two-node VSAN ROBO configuration) is another option. In this configuration, the HA functionality is provided by VSAN. Finally, a single-node ONTAP Select cluster replicating its data to a core location can provide a set of robust enterprise data management tools on top of a commodity server.

Figure 32 depicts a common remote office configuration using ONTAP Select.

Schedule-driven SnapMirror relationships periodically replicate the data from the remote office to a single consolidated engineered storage array located in the main data center.

Figure 32) Scheduled backup of remote office to corporate data center.

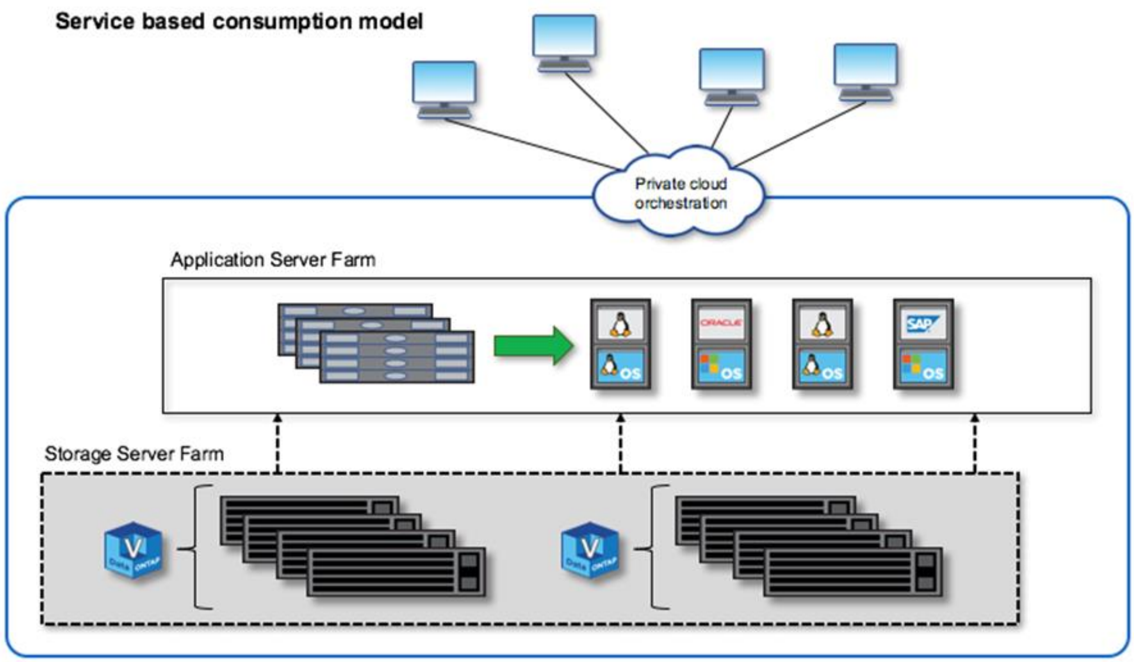


5.2 Private Cloud (Data Center)

Another common use case for ONTAP Select is providing storage services for private clouds built on commodity servers. Figure 33 shows how a storage farm provides compute and locally attached storage to the ONTAP Select VM, which provides storage services upstream to an application stack. The entire workflow, from the provisioning of SVMs to the deployment and configuration of application VMs, is automated through a private cloud orchestration framework.

This is a service-oriented private cloud model. Using the HA version of ONTAP Select creates the same ONTAP experience you would expect on higher-cost FAS arrays. Storage server resources are consumed exclusively by the ONTAP Select VM, with application VMs hosted on separate physical infrastructure.

Figure 33) Private cloud built on DAS.



5.3 MetroCluster Software Defined Storage (Two-Node Stretched Cluster High Availability)

Starting with ONTAP Select 9.3P2 and ONTAP Deploy 2.7, a two-node cluster can be stretched between two locations if certain minimum requirements are met. This architecture fits neatly in between hardware-based MetroCluster and single data-center clusters (hardware-defined or software-defined). The requirements for the ONTAP Select MetroCluster SDS highlight the general flexibility of software-defined storage solutions as well as the differences between it and the hardware-based MetroCluster SDS. No proprietary hardware is required.

Unlike MetroCluster, ONTAP Select uses existing network infrastructure and supports a network latency of up to 5ms RTT with a maximum jitter of up to 5ms, for a total of 10ms maximum latency. A maximum distance of 10km is also a requirement, although the latency profile is more important. Separation requirements in the market space have more to do with physical separation than the actual distance. In some instances, this can mean different buildings. In other instances, it can mean different rooms in the same building. Regardless of the actual physical placement, what defines a two-node cluster as a MetroCluster SDS is that each node uses a separate uplink switch.

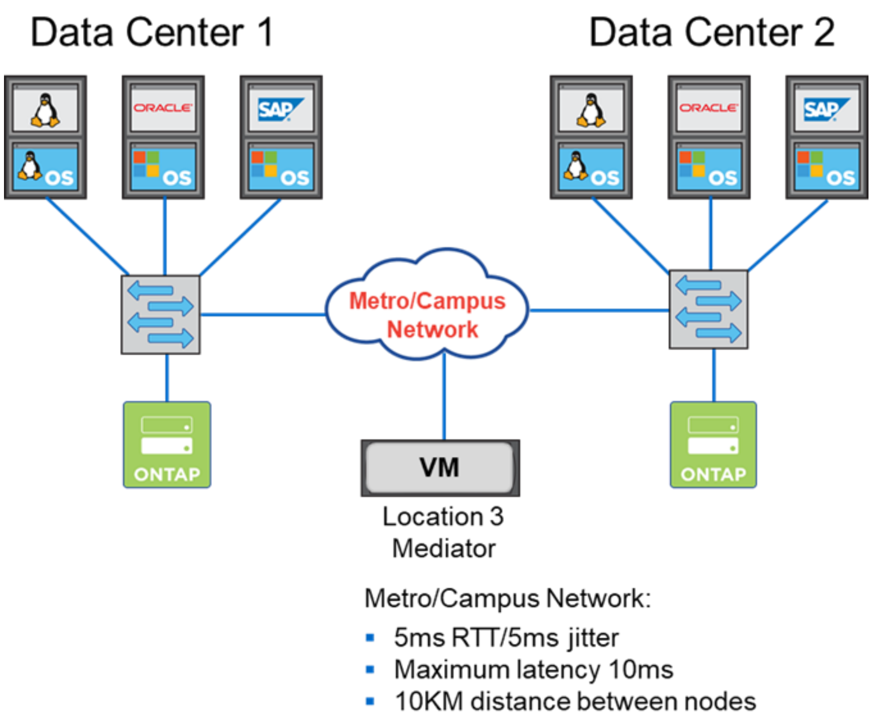
As part of the two-node HA configuration, a mediator is required to properly identify the active node during a failover and avoid any split-brain scenario in which both nodes remain independently active during a network partition. This operation is identical to the regular two-node HA configuration previously available. For proper protection and failover during site failure, the mediator should be in a different site from the two HA nodes. The maximum latency between the mediator and each ONTAP Select node cannot exceed 125ms.

With this solution, enterprise customers can confidently take advantage of the flexibility of a software-defined storage solution on commodity hardware. They can deploy with peace of mind knowing their data is protected with an enterprise-grade, 0 RPO solution.

ONTAP Select MetroCluster SDS provides the following benefits:

- MetroCluster SDS provides another dimension (data center to data center) of protection for ONTAP Select. Customers can now take advantage of this extra level of protection in addition to leveraging all the benefits of software-defined storage and ONTAP.
- MetroCluster SDS provides business-critical data protection with a 0 RPO and automatic failover. Both the data storage and the application access points are automatically switched over to the surviving data center or node with zero intervention from IT.
- MetroCluster SDS is cost effective. It takes advantage of the existing networking infrastructure to enable stretched resiliency between the HA pair, and no additional hardware is required. It also provides active/active data access and data center redundancy in the same cluster.

Figure 34) MetroCluster SDS.



For more best practices and other requirements, see the sections “Two-Node HA Versus Two-Node Stretched HA (MetroCluster SDS)” and “Two-Node Stretched HA (MetroCluster SDS) Best Practices.”

6 Upgrading ONTAP Select and ONTAP Deploy

This section contains important information about the maintenance of various aspects of an ONTAP Select cluster. It is possible to upgrade ONTAP Select and ONTAP Deploy independently of each other. Table 10 describes the support matrix for ONTAP Select and ONTAP Deploy.

Table 10) ONTAP Deploy versus ONTAP Select support matrix.

	Select 9.4	Select 9.5	Select 9.6
Deploy 2.9	Supported	Not supported	Not supported
Deploy 2.10	Supported	Supported	Not supported

	Select 9.4	Select 9.5	Select 9.6
Deploy 2.11	Supported	Supported	Not supported
Deploy 2.12	Supported	Supported	Supported

Note: ONTAP Deploy only manages the Select clusters that it has deployed. There is currently no functionality to discover ONTAP Select clusters installed using another instance of ONTAP Deploy. NetApp recommends backing up the ONTAP Deploy configuration every time a new cluster is deployed. Restoring the ONTAP Deploy database allows a new ONTAP Deploy instance to manage ONTAP Select clusters installed using another ONTAP Deploy VM. However, care should be taken so that one cluster is not managed by multiple ONTAP Deploy instances.

7 Increasing the ONTAP Select Capacity Using ONTAP Deploy

7.1 Increasing Capacity for ONTAP Select vNAS and DAS with Hardware RAID Controllers

ONTAP Deploy can be used to add and license additional storage for each node in an ONTAP Select cluster. The storage-add functionality in ONTAP Deploy is the only way to increase the storage under management, and directly modifying the ONTAP Select VM is not supported. The following figure shows the "+" icon that initiates the storage-add wizard.

The following considerations are important for the success of the capacity-expansion operation. Adding capacity requires the existing license to cover the total amount of space (existing plus new). A storage-add operation that results in the node exceeding its licensed capacity fails. A new license with sufficient capacity should be installed first.

If the extra capacity is added to an existing ONTAP Select aggregate, then the new storage pool (datastore) should have a performance profile similar to that of the existing storage pool (datastore). Note that it is not possible to add non-SSD storage to an ONTAP Select node installed with an AFF-like personality (flash enabled). Mixing DAS and external storage is also not supported.

If locally attached storage is added to a system to provide for additional local (DAS) storage pools, you must build an additional RAID group and LUN (or LUNs). Just as with FAS systems, care should be taken to make sure that the new RAID group performance is similar to that of the original RAID group if you are adding new space to the same aggregate. If you are creating a new aggregate, the new RAID group layout could be different if the performance implications for the new aggregate are well understood.

The new space can be added to that same datastore as an extent if the total size of the datastore does not exceed the ESX-supported maximum datastore size. Adding a datastore extent to the datastore in which ONTAP Select is already installed can be done dynamically and does not affect the operations of the ONTAP Select node.

If the ONTAP Select node is part of an HA pair, some additional issues should be considered.

In an HA pair, each node contains a mirror copy of the data from its partner. Adding space to node 1 requires that an identical amount of space is added to its partner, node 2, so that all the data from node 1 is replicated to node 2. In other words, the space added to node 2 as part of the capacity-add operation for node 1 is not visible or accessible on node 2. The space is added to node 2 so that node 1 data is fully protected during an HA event.

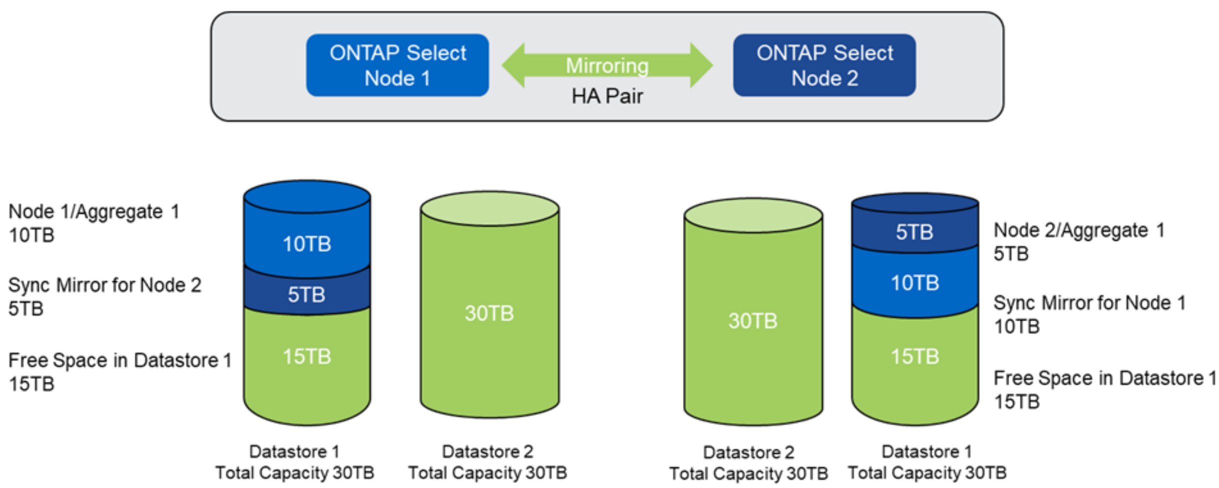
There is an additional consideration with regard to performance. The data on node 1 is synchronously replicated to node 2. Therefore, the performance of the new space (datastore) on node 1 must match the performance of the new space (datastore) on node 2. In other words, adding space on both nodes, but using different drive technologies or different RAID group sizes, can lead to performance issues. This is due to the RAID SyncMirror operation used to maintain a copy of the data on the partner node.

To increase user-accessible capacity on both nodes in an HA pair, two storage-add operations must be performed, one for each node. Each storage-add operation requires additional space on both nodes. The total space required on each node is equal to the space required on node 1 plus the space required on node 2.

Initial setup is with two nodes, each node having two datastores with 30TB of space in each datastore. ONTAP Deploy creates a two-node cluster, with each node consuming 10TB of space from datastore 1. ONTAP Deploy configures each node with 5TB of active space per node.

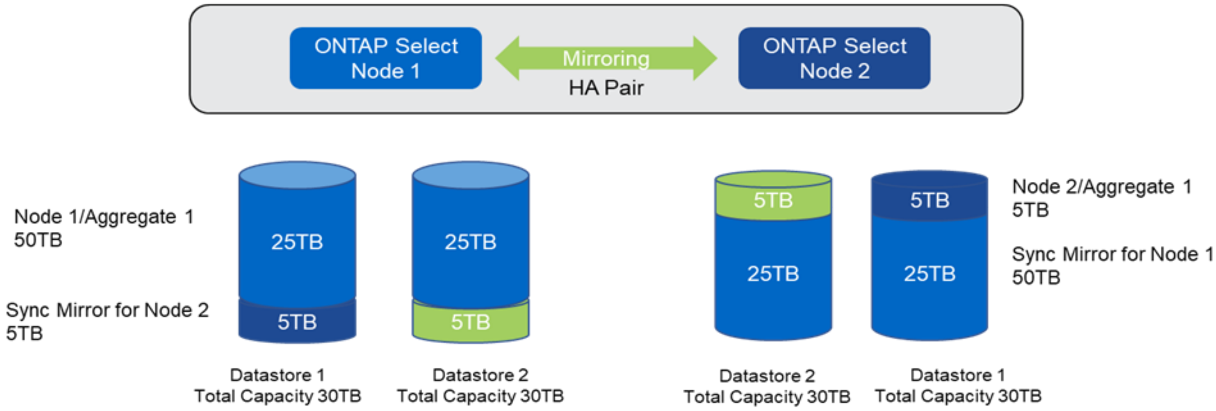
Figure 35 shows the results of a single storage-add operation for node 1. ONTAP Select still uses an equal amount of storage (15TB) on each node. However, node 1 has more active storage (10TB) than node 2 (5TB). Both nodes are fully protected as each node hosts a copy of the other node's data. There is additional free space left in datastore 1, and datastore 2 is still completely free.

Figure 35) Capacity distribution: allocation and free space after a single storage-add operation.



Two additional storage-add operations on node 1 consume the rest of datastore 1 and a part of datastore 2 (using the capacity cap). The first storage-add operation consumes the 15TB of free space left in datastore 1. Figure 36 shows the result of the second storage-add operation. At this point, node 1 has 50TB of active data under management, while node 2 has the original 5TB.

Figure 36) Capacity distribution: allocation and free space after two additional storage-add operation for node 1.



Starting with ONTAP Deploy 2.7 and ONTAP Select 9.3, the maximum VMDK size used during capacity add operations is 16TB. The maximum VMDK size used during cluster create operations continues to be 8TB. ONTAP Deploy creates correctly sized VMDKs depending on your configuration (a single-node or multinode cluster) and the amount of capacity being added. However, the maximum size of each VMDK should not exceed 8TB during the cluster create operations and 16TB during the storage-add operations.

7.2 Increasing capacity for ONTAP Select with Software RAID

The storage-add wizard can similarly be used to increase capacity under management for ONTAP Select nodes using software RAID. The wizard only presents those DAS SDD drives that are available and can be mapped as RDMs to the ONTAP Select VM.

Though it is possible to increase the capacity license by a single TB, when working with software RAID, it is not possible to physically increase the capacity by a single TB. Similar to adding disks to a FAS or AFF array, certain factors dictate the minimum amount of storage that can be added in a single operation.

Note that in an HA pair, adding storage to node 1 requires that an identical number of drives is also available on the node's HA pair (node 2). Both the local drives and the remote disks are used by one storage-add operation on node 1. That is to say, the remote drives are used to make sure that the new storage on node 1 is replicated and protected on node 2. In order to add locally usable storage on node 2, a separate storage-add operation and a separate and equal number of drives must be available on both nodes.

ONTAP Select partitions any new drives into the same root, data, and data partitions as the existing drives. The partitioning operation takes place during the creation of a new aggregate or during the expansion on an existing aggregate. The size of the root partition stripe on each disk is set to match the existing root partition size on the existing disks. Therefore, each one of the two equal data partition sizes can be calculated as the disk total capacity minus the root partition size divided by two. The root partition stripe size is variable, and it is computed during the initial cluster setup as follows. Total root space required (68GB for a single-node cluster and 136GB for HA pairs) is divided across the initial number of disks minus any spare and parity drives. The root partition stripe size is maintained to be constant on all the drives being added to the system.

If you are creating a new aggregate, the minimum number of drives required varies depending on the RAID type and whether the ONTAP Select node is part of an HA pair.

If adding storage to an existing aggregate, some additional considerations are necessary. It is possible to add drives to an existing RAID group, assuming that the RAID group is not at the maximum limit already. Traditional FAS and AFF best practices for adding spindles to existing RAID groups also apply here, and creating a hot spot on the new spindle is a potential concern. In addition, only drives of equal or larger data partition sizes can be added to an existing RAID group. As explained above, the data partition size is not the same as drive raw size. If the data partitions being added are larger than the existing partitions, the new drives is right-sized. In other words, a portion of capacity of each new drive remains unutilized.

It is also possible to use the new drives to create a new RAID group as part of an existing aggregate. In this case, the RAID group size should match the existing RAID group size.

7.3 Single-Node to Multinode Upgrade and Cluster Expansions

Upgrading from the single-node, non-HA version of ONTAP Select to the multinode scale-out version is not supported. Migrating from a single-node version to a multinode version requires the provisioning of a new ONTAP Select cluster using SnapMirror technology to copy existing data from the single-node cluster.

Expanding or reducing the number of nodes in a multinode ONTAP Select cluster is not a supported workflow at the time of writing this document.

7.4 Increasing and Decreasing the ONTAP Select VM Size

Starting with ONTAP Deploy 2.12 and ONTAP Select 9.6, it is possible to increase, but not decrease, the ONTAP Select VM size in conjunction with the application of the correct license for the new ONTAP Select instance. In other words, it is possible to move from the ONTAP Select small VM (using a Standard license) to the ONTAP Select medium VM (using a Premium license), to the ONTAP Select large VM (using a Premium XL license). If you are already using a larger license, there is no need to provide a new license. There are limitations that one should be aware of:

- In the current release, it is only possible to move up, not down, in VM size.
- In the current release, increasing the ONTAP Select VM size is only supported with a Capacity Tier type of license.
- In the current release, the large ONTAP Select VM is only supported for deployments using software RAID (direct-attached storage without a hardware raid controllers). ONTAP Deploy does not block the creation of a large VM on a system that is not using software RAID. A warning is displayed notifying the administrator that this configuration is not supported, but the operation is allowed to proceed.

The workflow to increase the ONTAP Select VM size is managed from ONTAP Deploy, and it requires a node reboot. If the node is part of an HA pair, ONTAP Deploy upgrades one node at the time and begins the second node upgrade only after the first node has finished the upgrade successfully and can take over services for node 2.

8 ONTAP Select Performance

The performance numbers described in this section are intended as a rough estimate of the performance of an ONTAP Select cluster and are not a performance guarantee. The performance of an ONTAP Select cluster can vary considerably due to the characteristics of the underlying hardware and configuration. As a matter of fact, the specific hardware configuration is the biggest factor in the performance of a particular ONTAP Select instance. Here are some of the factors that affect the performance of a specific ONTAP Select instance:

- **Core frequency.** In general, a higher frequency is preferable.
- **Single socket versus multisocket.** ONTAP Select does not use multisocket features, but the hypervisor overhead for supporting multisocket configurations accounts for some amount of deviation in total performance.
- **RAID card configuration and associated hypervisor driver.** The default driver provided by the hypervisor might need to be replaced by the hardware vendor driver.

- **Drive type and number of drives in the RAID group(s).**
- **Hypervisor version and patch level.**

This document includes performance comparisons only when the testing was performed on the exact same test bed to highlight the impact of a specific feature. In general, we document the hardware environment and run the highest performance configuration possible on that platform.

8.1 ONTAP Select Premium 9.4 HA Pair with Direct-Attached Storage (SSD)

Reference Platform

ONTAP Select 9.4 (Premium) hardware (per node):

- Cisco UCS C240 M4S2:
  - Intel Xeon CPU E5-2697 at 2.60GHz
  - 2 x sockets; 14 x CPUs per socket
  - 56 x logical CPUs (HT enabled)
  - 256GB RAM
  - VMware ESXi 6.5
  - Drives per host: 24 X371A NetApp 960GB SSD

Client hardware:

- 4 x NFSv3 IBM 3550m4 clients

Configuration information:

- 1,500 MTU for data path between clients and Select cluster
- No storage efficiency features in use (compression, deduplication, Snapshot copies, SnapMirror, and so on)

Table 11 lists the throughput measured against read/write workloads on an HA pair of ONTAP Select Premium nodes. Performance measurements were taken using the SIO load-generating tool.

Table 11) Performance results for a single node (part of a four-node medium instance) ONTAP Select 9.4 cluster on DAS (SSD).

Description	Sequential Read 64KiB	Sequential Write 64KiB	Random Read 8KiB	Random Write 8KiB	Random WR/ RD (50/50) 8KiB
ONTAP 9.4 Select Medium instance with DAS (SSD)	1045MBps 16,712 IOPS	251MBps 4016 IOPS	492MBps 62,912 IOPS	141MBps 18,048 IOPS	218MBps 27,840 IOPS

64K Sequential Read

Details:

- SIO direct I/O enabled
- 2 x data NIC
- 1 x data aggregate (2TB)
- 64 volumes; 64 SIO procs/threads
- 32 volumes per node (64 total)
- 1 x SIO procs per volume; 1 x SIO thread per file
- 1 x files per volume; files are 12000MB each

64K Sequential Write

Details:

- SIO direct I/O enabled
- 2 x data NIC
- 1 x data aggregate (2TB):
- 64 volumes; 128 SIO procs/threads
- 32 volumes per node (64 total)
- 2 x SIO procs per volume; 1 x SIO thread per file
- 2 x files per volume; files are 30720MB each

8K Random Read

Details:

- SIO direct I/O enabled
- 2 x data NIC
- 1 x data aggregate (2TB):
- 64 volumes; 64 SIO procs/threads
- 32 volumes per node (64 total)
- 1 x SIO procs per volume; 8 x SIO thread per file
- 1 x files per volume; files are 12228MB each

8K Random Write

Details:

- SIO direct I/O enabled
- 2 x data NIC
- 1 x data aggregate (2TB)
- 64 volumes; 64 SIO procs/threads
- 32 volumes per node (64 total)
- 1 x SIO procs per volume; 8 x SIO thread per file
- 1 x files per volume; files are 8192MB each

8K Random 50% Write 50% Read

Details:

- SIO direct I/O enabled
- 2 x data NIC
- 1 x data aggregate (2TB)
- 64 volumes; 64 SIO procs/threads
- 32 volumes per node (64 total)
- 1 x SIO procs per volume; 20 x SIO thread per file
- 1 x files per volume; files are 12228MB each

8.2 ONTAP Select Premium 9.5 HA Pair with Direct-Attached Storage (SSD)

Reference Platform

ONTAP Select 9.5 (Premium) hardware (per node):

- Cisco UCS C240 M4SX:
  - Intel Xeon CPU E5-2620 at 2.1GHz
  - 2 x sockets; 16 x CPUs per socket
  - 128GB RAM
  - VMware ESXi 6.5
  - Drives per host: 24 900GB SSD

Client hardware:

- 5 x NFSv3 IBM 3550m4 clients

Configuration information:

- 1,500 MTU for data path between clients and Select cluster
- No storage efficiency features in use (compression, deduplication, Snapshot copies, SnapMirror, and so on)

Table 12 lists the throughput measured against read/write workloads on an HA pair of ONTAP Select Premium nodes using both software RAID and hardware RAID. Performance measurements were taken using the SIO load-generating tool.

Table 12) Performance results for a single node (part of a four-node medium instance) ONTAP Select 9.5 cluster on DAS (SSD) with software RAID and hardware RAID.

Description	Sequential Read 64KiB	Sequential Write 64KiB	Random Read 8KiB	Random Write 8KiB	Random WR/ RD (50/50) 8KiB
ONTAP 9.5 Select Medium instance with DAS (SSD) hardware RAID	1,714MiBps	412MiBps	391MiBps	251MiBps	309MiBps
ONTAP 9.5 Select Medium instance with DAS (SSD) software RAID	1,674MiBps	360MiBps	451MiBps	223MiBps	293MiBps

64K Sequential Read

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (8 TB software RAID)
- 64 SIO procs, 1 thread per proc
- 32 volumes per node
- 1 x files per proc; files are 12000MB each

64K Sequential Write

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (4TB software RAID)
- 128 SIO procs, 1 thread per proc
- volumes per node 32 (hardware RAID), 16 (software RAID)
- 1 file per proc; files are 30720MB each

8K Random Read

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (4TB software RAID)
- 64 SIO procs, 8 thread per proc
- volumes per node 32
- 1 file per proc; files are 12228MB each

8K Random Write

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (4 TB software RAID)
- 64 SIO procs, 8 threads per proc
- volumes per node 32
- 1 file per proc; files are 8192MB each

8K Random 50% Write 50% Read

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (4TB software RAID)
- 64 SIO procs, 20 threads per proc
- volumes per node 32
- 1 file per proc; files are 12228MB each

8.3 ONTAP Select Premium 9.6 HA Pair with Direct-Attached Storage (SSD)

Reference Platform

ONTAP Select 9.6 (Premium XL) Hardware (per Node)

- FUJITSU PRIMERGY RX2540 M4:
  - Intel(R) Xeon(R) Gold 6142b CPU at 2.6 GHz
  - 32 physical cores (16 x 2 sockets), 64 logical
  - 256 GB RAM
  - Drives per host: 24 960GB SSD
  - ESX 6.5U1

Client Hardware

- 5 x NFSv3 IBM 3550m4 clients

Configuration Information

- SW RAID 1 x 9 + 2 RAID-DP (11 drives)
- 22+1 RAID-5 (RAID-0 in ONTAP) / RAID cache NVRAM
- No storage efficiency features in use (compression, deduplication, Snapshot copies, SnapMirror, and so on)

Table 12 lists the throughput measured against read/write workloads on an HA pair of ONTAP Select Premium nodes using both software RAID and hardware RAID. Performance measurements were taken using the SIO load-generating tool.

Table 13) Performance results for a single node (part of a four-node medium instance) ONTAP Select 9.5 cluster on DAS (SSD) with software RAID and hardware RAID.

Description	Sequential Read 64KiB	Sequential Write 64KiB	Random Read 8KiB	Random Write 8KiB	Random WR/RD (50/50) 8KiB
ONTAP 9.6 Select Large instance with DAS (SSD) software RAID	2171 MiBps	559 MiBps	954 MiBps	394 MiBps	564 MiBps
ONTAP 9.6 Select Medium instance with DAS (SSD) software RAID	2090 MiBps	592 MiBps	677 MiBps	335 MiBps	441 MiBps
ONTAP 9.6 Select Medium instance with DAS (SSD) hardware RAID	2038 MiBps	520 MiBps	578 MiBps	325 MiBps	399 MiBps

64K Sequential Read

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (8TB software RAID)

- 64 SIO procs, 1 thread per proc
- 32 volumes per node
- 1 x files per proc; files are 12000MB each

64K Sequential Write

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (4TB software RAID)
- 128 SIO procs, 1 thread per proc
- volumes per node 32 (hardware RAID), 16 (software RAID)
- 1 file per proc; files are 30720MB each

8K Random Read

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (4 TB hardware RAID)
- 64 SIO procs, 8 thread per proc
- volumes per node 32
- 1 file per proc; files are 12228MB each

8K Random Write

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (4TB software RAID)
- 64 SIO procs, 8 threads per proc
- volumes per node 32
- 1 file per proc; files are 8192MB each

8K Random 50% Write 50% Read

Details:

- SIO direct I/O enabled
- 2 nodes
- 2 x data NIC per node
- 1 x data aggregate per node (2TB hardware RAID), (4TB software RAID)
- 64 SIO procs, 20 threads per proc
- volumes per node 32

- 1 file per proc; files are 12228MB each

Where to Find Additional Information

To learn more about the information described in this document, refer to the following documents and/or websites:

- ONTAP Select product page
https://www.netapp.com/us/products/data-management-software/ontap-select-sds.aspx
- ONTAP Select Resources page
http://mysupport.netapp.com/ontapselect/resources
- ONTAP 9 Documentation Center
http://docs.netapp.com/ontap-9/index.jsp

Version History

Version	Date	Document Version History
Version 1.0	June 2016	Initial version.
Version 1.1	August 2016	Updated the networking sections 2.5 and 5.
Version 1.2	December 2016	• Added support for ONTAP Select 9.1 and OVF evaluation method. • Consolidated the networking section. • Consolidated the deploy section.
Version 1.3	March 2017	• Added support for ONTAP Deploy 2.3, external array, and VSAN. • Added support for SATA and NL-SAS along with datastore size considerations for larger capacity media. • Added IOPS metrics to performance table. • Added network checker for internal network troubleshooting.
Version 1.41	June 2017	• Added support for ONTAP Deploy 2.4, ONTAP Select 9.2, and 2-node clusters. • Added VSAN performance information.
Version 1.5	March 2018	Added support for ONTAP Deploy 2.7 and ONTAP Select 9.3.
Version 1.6	June 2018	Added support for ONTAP Deploy 2.8 and ONTAP Select 9.4.
Version 1.7	February 2019	Added support for ONTAP Deploy 2.10 and ONTAP Select 9.5.
Version 1.8	August 2019	Added support for ONTAP Deploy 2.11 and ONTAP Select 9.6.
Version 1.9	November 2019	Terminal content update. Further updates can be found at https://docs.netapp.com/us-en/ontap-select.

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

NetApp®