Technical Report

ONTAP Select on KVM

Product Architecture and Best Practices

Veena Kannan, NetApp
March 2019 | TR-4613

NetApp®

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

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. ONTAP Select extends the Data Fabric solution to the commodity server offerings likely existing in the customer's data center.

This document describes the best practices that should be followed when building an ONTAP Select cluster in a Kernel-Based Virtual Machine (KVM) hypervisor environment, from hardware selection to deployment and configuration. Additionally, it aims to answer the following questions:

- How is ONTAP Select different from the engineered FAS systems?
- 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 provide administrators with the ability to rapidly provision resources with a level of 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 them to react quickly to changing IT needs by dynamically shifting infrastructure resources to where they are needed most.

Software-defined infrastructures are built on three tenets:

- Flexibility
- Scalability
- Programmability

Software-Defined Storage

The shift toward software-defined infrastructures might be having its greatest affect 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 trend is especially evident within private cloud environments. These are enterprise-class, service-oriented architectures designed from the ground up with software definition in mind. 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 based on the premise of bundling compute, storage, and networking services. The rapid adoption of hyperconverged solutions over the past several years has highlighted the desire for simplicity and flexibility. However, as companies have replaced enterprise-class storage arrays with a more customized, make-your-own model by building storage management solutions on top of homegrown components, a set of new problems emerges.

In a commodity world in which data is fragmented across silos of direct-attached storage, data mobility and data management 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 using ONTAP and all of its storage management services. Decoupling ONTAP from the underlying hardware allows you to use enterprise-class file and replication services within a software-defined storage environment.

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

Running ONTAP as software on top of another application allows you to use much of the qualification work performed by the hypervisor, critical in helping rapidly expand our list of supported platforms. In addition, implementing ONTAP as a virtual machine (VM) allows customers to plug into existing management and orchestration frameworks. This configuration enables rapid provisioning and end-to-end automation, from deployment to ripping off the setup (for example, in a DevOps environment).

1.3 ONTAP Select Features

Table 1 highlights some of the major features of ONTAP Select.

Table 1) ONTAP Select configuration.

Description	ONTAP Select
Node count	Single node and 2, 4, 6 and 8-node high availability (HA)
VM CPU/memory	4 vCPUs/16GB (small instance) 8 vCPUs/64GB* (medium instance) (*requires Premium license)
Hypervisor	KVM and related packages (see Software Requirements)
HA	Yes
iSCSI/CIFS/NFS	Yes
NetApp SnapMirror® and NetApp SnapVault® technology	Yes
Compression	Yes
Capacity limit	Up to 400TB per node
Disk type	Serial-attached SCSI (SAS), near-line SAS (NL-SAS), SATA, or solid-state drive (SSD)* (*requires Premium license)
Minimum number of disks	8 SAS, NL-SAS, SATA, or 4 SSDs* (*requires Premium license)
Maximum number of disks	60
Select cluster size	Single node and 2, 4, 6 and 8-node HA
Hardware support	Wider support for major vendor offerings that meet minimum criteria

Note: A total of 60 virtual disks (as seen within the ONTAP Select VM) is supported. The maximum size of a virtual disk is 16TB.

1.4 ONTAP Select Standard Versus ONTAP Select Premium

The Premium license allows the user to configure either a small instance or a medium instance whereas the Standard license can only be used with a small instance. The difference between the small instance and medium instance consists of the number of resources reserved for each instance of Select. For example, the Premium VM consumes eight CPU cores and 64GB of RAM. More information can be found in section 2.1, “Virtual Machine Properties.”

The number of cores and amount of memory per ONTAP Select VM cannot be further modified. In addition, a Premium Select license is required when using SSDs for direct-attached storage (DAS)

configurations (hardware RAID controller or ONTAP Software RAID) or any NetApp ONTAP MetroCluster SDS constructs.

In a multinode 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 Select VMs should be identical.

Note: It is not possible to convert from a Standard license to a Premium license.

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. Therefore, a client can test the full solution, including ONTAP Deploy, the ONTAP Select setup product. Deploy checks and enforces all minimum requirements for ONTAP Select, which is useful for both documenting the procedure and vetting the environment for suitability.

Note that the direct deployment of raw ONTAP Select images without the use of the ONTAP Deploy utility is not supported.

Starting with Deploy 2.11 and ONTAP Select 9.5, conversion of evaluation to production licenses is supported.

Prior to Deploy 2.11, after 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.

1.6 ONTAP Select System 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.

Software Requirements

ONTAP Select requires that the hosting server meet the following software requirements. These versions have been qualified for use with the most typical deployment scenarios.

Note that the following packages might require the uninstallation of the default or preexisting packages and reinstallation of previous versions. Package dependencies must be resolved correctly to make sure of proper functioning of the virtualization environment. Use the `rhel-7-server-rpms` package and the `rhel7-server-openstack-7.0-rpms` repositories.

The following table (Table 2) summarizes the mandatory Linux versions and packages required for each version of ONTAP Select and Deploy:

Table 2) Mandatory Linux versions.

ONTAP Select 9.5 Deploy 2.10	ONTAP Select 9.4 Deploy 2.9	ONTAP Select 9.4 Deploy 2.8	ONTAP Select < 9.4 OR Deploy < 2.8	ONTAP Select
Red Hat Enterprise Linux	Red Hat Enterprise Linux 7.5/CentOS 7.5	Red Hat Enterprise Linux 7.4/CentOS 7.4	Red Hat Enterprise Linux 7.2, 7.3/CentOS 7.2, 7.3	RHEL/CentOS Linux versions

ONTAP Select 9.5 Deploy 2.10	ONTAP Select 9.4 Deploy 2.9	ONTAP Select 9.4 Deploy 2.8	ONTAP Select < 9.4 OR Deploy < 2.8	ONTAP Select
7.6/CentOS 7.6				
qemu-kvm-rhev2.9.0 (SW RAID), qemu-kvm1.5.3 (HW_RAID)	qemu-kvm-rhev2.9.0(SW_RAID) qemu-kvm1.5.3-156(HW_RAID)	qemu-kvm-rhev2.9.0(SW_RAID) qemu-kvm1.5.3-141(HW_RAID)	qemu-kvm-1.5.3126.el7_3.3.x86_64	Linux kernel module that enables mapping of a physical CPU to a virtual CPU. This installed module converts the underlying Linux OS into a type 1 hypervisor.
Libvirt4.5.0-10.e17_6.3	libvirt-3.9.0	libvirt-3.2.014.el7_3.5.x86_64/libvirtdaemon-3.2.0-14.el7/libvirt-snmp-0.0.35.el7.x86_64 / libvirtpython-3.2.0-3.el7_4.1.x86_64	libvirt-2.0.010.el7_3.5.x86_64	A collection of software that provides a convenient way to manage VMs and other virtualization functionality, such a storage and network interface management. These software pieces include a long-term stable C API, a daemon (libvirtd), and a command-line utility (virsh).
openvswitch.2.7.3-1.x86_64	openvswitch-2.7.3-1.x86_64	openvswitch-2.7.3-1.x86_64	openvswitch2.5.0-14	Open vSwitch (OVS) provides standard network bridging functions and support for the OpenFlow protocol for remote per-flow control of traffic.
virt-install 1.5.0-1.noarch	virt-install1.4.1-7.0.1.el7.noar ch	virt-install1.4.1-7.0.1.el7.noar ch	virt-install1.4.0-2	virt-install is a command-line tool for creating new container guests using the "libvirt" hypervisor management library (only required for Deploy).
lshw.B.02.18-12. x86_64.	lshw-B.02.18-12.el7.x86_64	lshw-B.02.187.el7.x86_64	lshw-B.02.1712	Small tool to extract detailed information about the hardware configuration of a machine. It can report the exact memory configuration, firmware version, mainboard configuration, CPU version and speed, cache configuration, bus speed, and so on.

ONTAP Select 9.5 Deploy 2.10	ONTAP Select 9.4 Deploy 2.9	ONTAP Select 9.4 Deploy 2.8	ONTAP Select < 9.4 OR Deploy < 2.8	ONTAP Select
Lsscsi. 0.27-6. x86_64	lsscsi.x86_640.27 -6	lsscsi.x86_640.27-4	lsscsi.x86_640.27-4	Lists information about SCSI devices attached to the system.
telnet-server. 0.17-64.x86_64	telnet-server0.17-64.el7.x86_64	telnet-server0.17-64.el7.x86_64	telnet-server-0.17-64.el7.x86_64	Provides Telnet services
Fence-agents-all-4.2.1-11.e17_6.7	fence-agentsall-4.0.1186.el7_4.1	fence-agentsall-4.0.1166.el7_4.0	fence-agentsall-4.0.1166.el7_4.1	Collection of cluster management scripts
lvm2. 2.02.180-10.e17_6.2. x86_64	lvm2-cluster-2.02.177-4.el7	lvm2-cluster-2.02.171-8.el6	lvm2-cluster2.02.171-8.el7	Clustered Logical Volume Manager (CLVM) is a set of clustering extensions to the Logical Volume Manager (LVM). These extensions allow a cluster of computers to manage shared storage (for example, on a SAN) using LVM.
Pacemaker-1.1.19	pacemaker-1.1.18-11.e17	pacemaker-1.1.16-11.e16	pacemaker-1.1.16-11.e16	Advanced, scalable cluster resource manager to provide Linux Host High-Availability. It has capabilities for managing resources and dependencies. It runs scripts at initialization, when machines go up or down, and when related resources fail. It can also be configured to periodically check resource health.
Pcs-0.9.165-6.e17	pcs-0.9.162-6.e16	pcs-0.9.158-6.e17	pcs-0.9.158-6.e16	Pacemaker configuration tool. It permits users to easily view, modify, and create pacemaker-based clusters.

Note: For Linux 7.4 version onwards, `openstack40_extras` might have to be enabled (changed to 1 from 0) on the `.repo` file for OpenStack to enable the downloading and installation of OpenStack related packages. `OpenVswitch-2.7.3` might have to be built from sources and installed from Red Hat Package Manager (RPM).

Best Practice

An upgrade to ONTAP 9.5 Deploy 2.10 must first involve an upgrade of the corresponding Linux version to RHEL/CentOS 7.6 along with the corresponding packages listed for Deploy 2.10. For this upgrade, the ONTAP Select VM might have to be migrated to another host temporarily before being brought back to the upgraded host.

Best Practice

During installation of RHEL, select a default security policy and software to be the virtualization host. The destination should be a local boot disk because the RAID logical unit numbers (LUNs) are used to create ONTAP Select storage pools. To install the listed packages, the `–obsoletes` flag might need to be used. Firewall rules must be created to open up some ports, such as for the console.

Hardware Requirements

For ONTAP Select, the physical host server must meet the following minimum requirements:

- Intel Xeon E5-26xx v3 (Haswell) CPU or greater
- Six cores (four for ONTAP Select, two for OS)
- 24GB RAM (16GB for ONTAP Select, and 8GB for the OS)
- A minimum of two 1Gb network interface card (NIC) ports for single-node clusters, a minimum of four 1Gb NIC ports for two-node clusters, and two 10GbE NIC ports (four recommended) for four-node clusters.
- Starting with ONTAP 9.3, a minimum of a single 10Gb NIC port is supported, but not recommended

An ONTAP Select Premium license supports both a small VM (minimum requirements earlier) and a medium VM. An ONTAP Select medium VM reserves eight cores and 64GB of RAM. Therefore, the server minimum requirements should be adjusted accordingly.

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

- 8–60 internal disks (SAS, NL-SAS, or SATA)
- 4–60 SSDs (Select Premium)
- A hardware RAID controller with a 512MB writeback cache and 12Gbps of throughput.
- Software RAID requires that no hardware RAID controller is present.

For shared storage (external arrays), the RAID controller is no longer a requirement. However, the following restrictions and best practices should be considered when using storage LUNs to host ONTAP Select.

- Support for external arrays requires ONTAP Select 9.3 Deploy 2.6 at a minimum
- Support for KVM Live Migration requires ONTAP Select 9.3 Deploy 2.6 at a minimum
- Multinode clusters on shared storage are supported starting with ONTAP Deploy 2.8 and ONTAP Select 9.4. For prior releases, only single node Select clusters are supported with external array-type LUNs. For multinode clusters, you must use local storage (DAS).
- The external array vendor must list Red Hat Linux 7.4/7.5/7.6 or CentOS 7.4/7.5/7.6 as an OS that supports the array.

ONTAP Select Feature Support

The ONTAP Select release offers full support for most ONTAP functionality, except for those features that have hardware-specific dependencies such as NetApp MetroCluster or FCoE.

The supported functionality includes:

- NFS, CIFS, and iSCSI
- NetApp SnapMirror and NetApp SnapVault
- NetApp FlexClone® technology
- NetApp SnapRestore® technology
- NetApp Volume Encryption (NVE)
- NetApp ONTAP FlexGroups
- NetApp SnapLock® Enterprise
- NetApp FabricPool (separate license)
- NetApp FlexCache® (separate license)
- NetApp SyncMirror (separate license)
- NetApp Data Availability Services (separate license)
- ONTAP 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 NetApp Active IQ® Unified Manager, NetApp OnCommand Insight (OCI), NetApp OnCommand Workflow Automation (WFA), and NetApp SnapCenter®. Use of SnapCenter, NetApp SnapManager®, or NetApp SnapDrive® with ONTAP Select requires server-based licenses.

Consult the Interoperability Matrix Tool (IMT) for a complete list of supported management applications.

Note that the following ONTAP features are not supported by ONTAP Select:

- Interface groups (ifgroups)
- Service Processor
- Hardware-centric features such as MetroCluster, FC SAN (FC/FCoE), and full disk encryption (FDE)
- NetApp Storage Encryption (NSE) drives

ONTAP Select Storage Efficiency Support

ONTAP Select provides storage efficiency options that are similar to the storage efficiency options present on FAS and All Flash FAS arrays. ONTAP Select supports SSD media. However, there are significant differences in default behavior between the releases, as well as between ONTAP Select Premium with SSD media and All Flash FAS arrays.

ONTAP Select also supports the full volume-level inline deduplication functionality.

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.

An AFF-like personality automatically enables the following storage efficiency features 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 an ONTAP Select instance has all the default storage efficiency policies enabled, run the following command on a newly created volume:

```
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
twonode95IP15::>
```

For ONTAP Select upgrades from 9.4 to 9.5, it is important that ONTAP Select 9.4 is installed on DAS SSD storage with a Premium license. In addition, the Enable Storage Efficiencies checkbox must be checked during the initial cluster installation using ONTAP Deploy. Enabling the AFF-like personality post ONTAP upgrade when the prior conditions are not met requires the manual creation of a boot argument and a node reboot. Contact technical support for further details.

ONTAP Deploy adds an additional configuration check during ONTAP Select cluster setup. This configuration check asks the user to confirm that DAS storage is of the type SSD. ONTAP Deploy enforces this check during setup, as well as during storage add operations. In other words, after an ONTAP Select Premium VM is configured for SSD storage, only local (DAS) SSD media can be added to that VM. There are several reasons for this, including the fact that ONTAP Select does not support multiple RAID controllers, nor does it support mixing media types on the same RAID controller. However, this enablement enforcement makes sure that SSD-appropriate storage efficiency options cannot be enabled on HDD-based datastores.

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	9.5 Premium (DAS SSD)	9.4[1] / 9.3 [2] Premium (DAS SSD)	9.5 / 9.4[1] / 9.3[2] Premium or Standard (DAS HDD)
Inline zero detection	Yes (default)	Yes Enabled by user on a per-volume basis.	Yes Enabled by user on a per-volume basis.
Volume inline deduplication	Yes (default)	Yes (recommended) Enabled by user on a per-volume basis	Not available
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.

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.
Background compression	Not supported	Not supported	Yes Enabled by user on a per-volume basis.
Compression scanner	Yes	Yes Enabled by user on a per-volume basis.	Yes
Inline data compaction	Yes (default)	Yes (recommended) Enabled by user on a per-volume basis.	Yes Enabled by user on a per-volume basis.
Compaction scanner	Yes	Yes Enabled by user on a per-volume basis.	Yes
Aggregate inline deduplication	Yes (default)	Yes (recommended) Enabled by user on a per-volume basis with space guarantee = none)	N/A
Volume background deduplication	Yes (default)	Yes (recommended)	Yes Enabled by user on a per-volume basis.
Aggregate background deduplication	Yes (default)	Yes (recommended) Enabled by user on a per-volume basis with space guarantee = none)	N/A

[1] ONTAP Select 9.4 on DAS SSDs (requires Premium license) allows existing data in an aggregate to be deduped using the 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.

Notes on upgrade behavior for DAS SSD configurations

On a system upgraded to ONTAP Select 9.5, storage efficiency values of existing volumes can be verified after the upgrade is complete. In addition, 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 the following points are true:

- Volumes with a space guarantee will not have inline data-compaction, aggregate inline deduplication and aggregate background deduplication enabled. These options can be enabled post upgrade.
- Volumes with a space guarantee will not have background compression enabled. This option can be enabled post upgrade.

- The storage efficiency policy on 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 the following points are true:

- Volumes with a space guarantee do not see any difference after upgrade.
- Volumes with a space guarantee have aggregate background deduplication turned on.
- Volumes with a storage policy of inline-only have their policy set to auto.
- Volumes with user-defined storage efficiency policies have no change in policy, except for volumes with a 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 to provide storage management services on a virtualized commodity server. The ONTAP Select product can be deployed two different ways:

- **Non-HA (single node).** The single-node version of ONTAP Select is well suited for storage infrastructures that provide their own storage resiliency such as RAID 5 or RAID 6 using a RAID controller. These systems offer data protection at the array layer. The single-node Select cluster can also be used as a remote solution in which the data is protected by ONTAP backup and replication features to a core location.
- **High availability (multinode).** The multinode version of the solution uses two, four, six, or eight ONTAP Select nodes and adds support for high-availability and ONTAP nondisruptive operations, all within a shared-nothing environment. Each individual node still uses hardware RAID controllers to provide resiliency for its own underlying storage.

Starting ONTAP 9.4 Deploy 2.8, a Software RAID option is available. See the section Software RAID.

When choosing a solution, resiliency requirements, environment restrictions, and cost factors should be considered.

Note: The single-node and multinode versions of ONTAP Select are deployment options, not separate products. However, the multinode solution requires the purchase of more node licenses, and 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.

2.1 Virtual Machine Properties

The ONTAP Select VM has a fixed set of properties that 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 shows the resources used by the ONTAP Select VM.

Table 4) ONTAP Select virtual machine properties.

Description	9 Single Node	Single Node
CPU/memory	4 cores/16GB RAM or	4 cores/16GB RAM or

Description	9 Single Node	Single Node
	8 cores/64GB RAM	8 cores/64GB RAM
Virtual network interfaces	3 (2 for ONTAP Select versions before 9.3)	7 (4 internal, 3 external; 6 for ONTAP Select versions before 9.3)
SCSI controllers	4	4
System boot disk	10GB	10GB
System core dump disk	120GB	120GB
Mailbox disk	N/A	556MB
Cluster root disk	68GB	68GB x 2 (because disk is mirrored)
Console access	Enter `virsh console <Deploy VM name>` in the Deploy VM and `admin user ssh` in the node management IP address	Enter `virsh console <Deploy VM>` in the Deploy VM, and `admin user ssh` into the node management IP or cluster-management IP address
Nonvolatile RAM (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)

Note: The core dump disk partition is separate from the system boot disk. The core file size is directly related to the amount of memory allocated to the ONTAP instance. Therefore, NetApp can support larger-sized memory instances in the future without requiring a redesign of the system boot disk.

Note: The NVRAM partition was separated as its own disk starting with ONTAP Select 9.5 installed on ESX 9.5 U2 and higher. Prior versions of ONTAP Select and ONTAP Select 9.5 installation that were upgraded from prior versions collocated the NVRAM partition on the boot disk.

When using local attached storage (DAS) in a hardware RAID configuration, ONTAP Select uses the hardware RAID controller cache to achieve a significant increase in write performance. Additionally, certain restrictions apply to the ONTAP Select VM. DAS software RAID configurations might optionally configure an NVMe drive as the service disk to achieve similar write performance benefits.

Specifically:

- Only one ONTAP Select VM from a cluster can reside on a single server (the host running KVM). Each single-node Select cluster can use only its own locally attached storage pools.
- Only direct-attached storage (DAS) is supported.

2.2 External Array Configurations

ONTAP Select clusters are supported on external array LUN types. In these configurations, back-end storage resiliency is assumed to be provided by the underlying infrastructure. A minimum requirement is that the underlying configuration is supported by Red Hat Linux. Therefore, RHEL or CentOS should be listed for the external array vendor as being compatible for the OS. The external array should also list RHEL and CentOS as the supported host OS.

ONTAP Select 9.3 and Deploy 2.6 enable support for Live Migration.

The following best practices should be considered when installing an ONTAP Select cluster on an external array LUN:

- The hosts (source and destination) involved in the VM migration should run Linux open-source components such as CLVM and Pacemaker (Predictive Cache Statistics [PCS] utility).
- External array configurations are supported with both ONTAP Select Standard and Premium.
- To avoid affecting other VMs using the same shared LUN, you must make sure that you have sufficient free capacity in that LUN. The LUN must be able to accommodate the true Select VM size as derived from the Select capacity settings.
- Live migration of ONTAP Select VMs is supported starting with ONTAP Select 9.3 and ONTAP Deploy 2.6. When an ONTAP Select VM moves to a different host during a live migration operation, the ONTAP Deploy 2.6 instance managing the ONTAP Select instance temporarily loses connectivity to the ONTAP Select VM. Therefore, the first operation fails with an error message stating that the ONTAP Select VM no longer exists on host `<hostname>`. The cluster refresh operation should then be used to re-establish connectivity with the ONTAP Select VM on the new host.
- FC, FCoE, and iSCSI are the only supported protocols for the back-end storage connectivity between a Red Hat Linux host and an external array.
- Hybrid arrays and All Flash arrays are supported with both ONTAP Select Standard and Premium.
- If array-side storage efficiency policies are natively supported, testing should be performed to determine whether to enable ONTAP Select storage efficiency features. NetApp does not recommend enabling storage efficiencies on both the external array-side and on ONTAP Select at the same time.
- Connectivity between the host and the external arrays should be through 10Gb iSCSI with no single point of failure. Jumbo frames are recommended.
- The ONTAP Select VM should have dedicated network ports for client traffic that do not overlap with the ports used for connectivity to the back-end array.

The following limitations should be considered when installing a single-node Select cluster on an external array-type datastore:

- Only one Select node per host is supported. Multiple single-node Select clusters can share an external array datastore as long as they are installed on separate hosts.
- ONTAP Deploy cluster refresh operations require that all hosts are managed by the same Deploy instance.
- A live migration operation can result in a situation in which two ONTAP Select VMs reside on the same Red Hat Linux host. This configuration is not currently supported, and ONTAP Deploy 2.6 is not able to re-establish management connectivity to the ONTAP Select VM until that VM is moved to another host.

NetApp FAS, NetApp SolidFire®, and NetApp E-Series arrays are supported behind the Red Hat Linux host. NetApp recommends following the NetApp Storage Best Practices documentation for the respective array.

Note: Configuring external storage shared between multiple hosts enables ONTAP Select VM migration because data is available on the same back-end storage from multiple hosts that share that storage.

2.3 RAID Services for Local Attached Storage

Some software-defined solutions require the presence of an SSD to act as a higher speed write-staging device. However, ONTAP Select uses a hardware RAID controller to achieve a write performance boost. A hardware RAID controller also provides the added benefit of protection against physical drive failures 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.

Starting from ONTAP 9.4 Deploy 2.8, a software RAID option will be made available (See the section “Software RAID”). This feature eliminates the need to use a hardware RAID controller by providing RAID protection choices through a software layer within ONTAP.

Note: ONTAP Select data aggregates are configured to use RAID 1 for virtual disks 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. NetApp supports as many of these options as possible, providing that they meet certain minimum requirements.

The RAID controller on the host must meet the following requirements:

- The HW 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, RAID 6).
- The drive cache should be set to disabled.
- The write policy should 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 and SSD drives, using a single RAID group of up to 24 drives allows ONTAP to reap the benefits of spreading incoming read requests across a higher number of disks. This feature provides a significant performance gain. With SAS/SSD configurations, performance testing was done against single LUN versus multi-LUN configurations. No significant differences were found, so, for simplicity's sake, NetApp recommend 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. We also recommend one spare per RAID group, although a global spare for all RAID groups can also be used.

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 include the following:

- **Write through.** 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.
- **Write around.** 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, and thus lower latency and higher throughput for mixed read/write workloads. However, without the presence of a BBU or nonvolatile flash technology, users operating in this mode run the risk of losing data during a power failure.

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

Best Practice

The server RAID controller must be configured to operate in writeback mode. If write workload performance issues are seen, check the controller settings and make sure that write through or write around is not enabled.

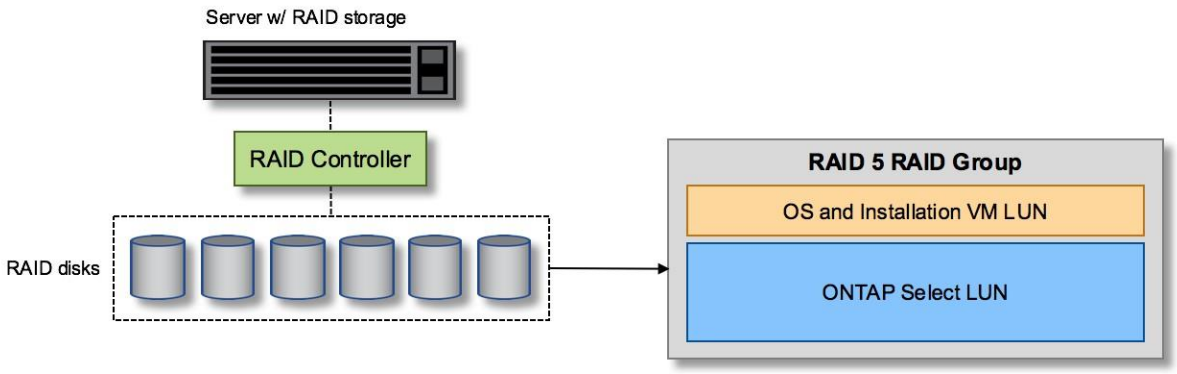
Local Disks Shared Between ONTAP Select and OS

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 another for the ONTAP Select VM.

Note: Dedicating one LUN for ONTAP Select assumes that the physical storage capacity of the system doesn't surpass the hypervisor-supported file system extent limits.

For example, a customer purchases a system with six internal SSDs and a single Smart Array RAID controller. All internal drives are managed by this RAID controller, and no other storage is present on the system (see Figure 1). No other storage is present on the system, so the hypervisor must share storage with the ONTAP Select node.

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



Note: The OS is Linux, and the installation VM is the Deploy VM.

Provisioning the OS LUNs from the same RAID group as ONTAP Select allows the hypervisor OS and any client VMs that are 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 configuration allows the administrator to re-create the OS LUN without affecting ONTAP Select.

Local Disks Split Between ONTAP Select and OS

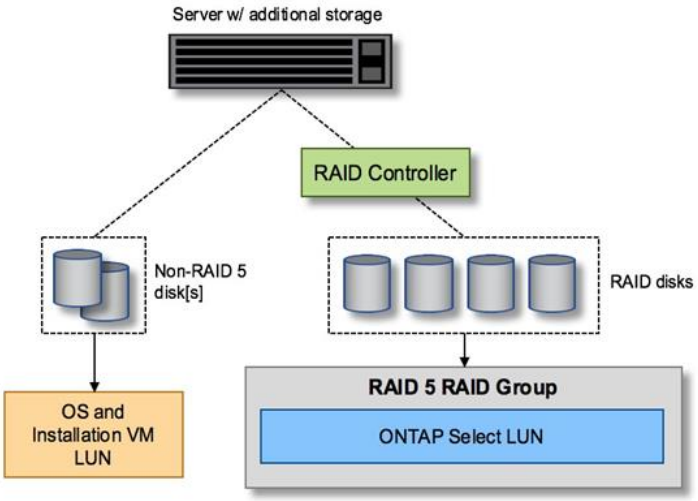
The other possible configuration provided by server vendors involves configuring the system with multiple RAID or disk controllers. In this configuration, a set of disks is managed by one disk controller, which

might or might not offer RAID services. Meanwhile, a second set of disks is managed by a hardware RAID controller that is able to offer RAID 5/6 services.

With this configuration, the set of spindles that sits behind the RAID controller that provides RAID 5/6 services should be used exclusively by the ONTAP Select VM. Depending on the total storage capacity under management, the disk spindles should be configured into one or more RAID groups and one or more LUNs. These LUNs are then used to create one or more volume groups, and all volumes are protected by the RAID controller.

The first set of disks is reserved for the hypervisor OS and any client VMs not using ONTAP storage. This configuration is depicted in Figure 2.

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



Multiple LUNs

There are two cases in which a single–RAID group and a single-LUN configuration must change. When using NL-SAS or SATA drives, the RAID group size must not exceed 12 drives. Additionally, a single LUN might become larger than the underlying hypervisor storage limits (either the individual file system extent maximum size or the total storage pool maximum size). Then the underlying physical storage must be broken up into multiple LUNs to permit successful file system creation.

Best Practice

Increasing the number of LUNs within a RAID group does not improve the performance of ONTAP Select. Multiple LUNs should only be used to follow best practices for SATA/NL-SAS configurations or to bypass hypervisor file system limitations.

2.4 LVM Limits and Storage Pool Considerations

The maximum device size with LVM is 8EB on 64-bit CPUs. If you create multiple RAID groups to improve the RAID rebuild time for SATA/NL-SAS drives, then multiple LUNs are provisioned.

When multiple LUNs are required, then make 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 must have a distinctly different performance profile, then we strongly recommend that you isolate these LUNs in a separate ONTAP aggregate.

Multiple physical volumes can be used to create a single volume group up to the maximum size of the volume group. 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 volume group.

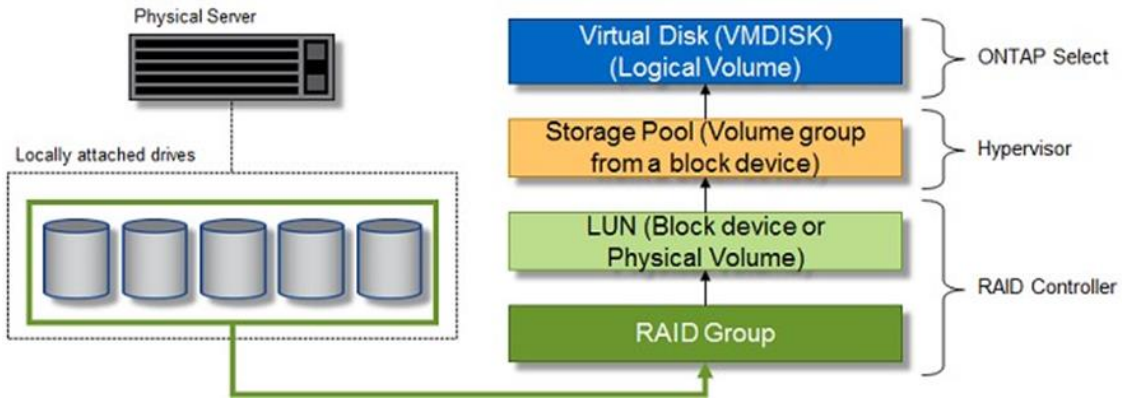
Alternatively, you can start by creating a single storage pool on a single LUN. When more space, which requires a larger ONTAP Select capacity license, is needed, a new storage pool (volume group) can be created. Using the Deploy utility, you can expose this additional storage to the ONTAP Select VM using the `storage add` command. The storage then shows up as an additional virtual disk. Both types of capacity extension operations are supported and can be achieved using the ONTAP Deploy storage add functionality. This functionality is covered in more detail in the section “Creating Volumes and Storage Pools.”

2.5 ONTAP Select Virtual Disks

At its core, ONTAP Select presents the ONTAP software 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. Note the following issues:

- RAID group and LUN configuration occurs from within the server's RAID controller software.
- Storage pool configuration happens from within the hypervisor.
- Virtual disks are created and owned by individual VMs: in this case, 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 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.

Because all virtual disks on the ONTAP Select VM are striped across the underlying physical disks, there is no performance gain in building configurations with a higher number of virtual disks. Additionally, shifting the responsibility of virtual disk creation and assignment from the administrator to the management tool prevents the user from inadvertently assigning a virtual disk to an incorrect storage pool.

ONTAP Select breaks up the underlying attached storage into equally sized virtual disks, none of which exceed 8TB in size. If the ONTAP Select node is part of an HA pair, a minimum of two virtual disks are

created on each cluster node. The virtual disks are then assigned to the local and mirror plex to be used within a mirrored aggregate.

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

Adding capacity to an ONTAP Select VM can create virtual disks of different sizes. However, unlike in FAS systems, virtual disks of different sizes can exist in the same aggregate. ONTAP Select uses a RAID 0 stripe across these virtual disks, which allows you to use all space in each virtual disk, regardless of its size.

Best Practice

In a manner similar to creating multiple LUNs, increasing the number of virtual disks used by the system provides no performance benefits for ONTAP Select.

2.6 Virtualized NVRAM

NetApp FAS systems are traditionally fitted with a physical NVRAM PCI card. This high-performance card contains nonvolatile flash memory that provides ONTAP with a significant boost in write performance in the following ways:

- Incoming writes back to the client are acknowledged immediately.
- The card schedules the movement of modified data blocks back to the slower storage media. This process is known as destaging.
- In a shared-nothing environment, NVRAM is used to store mirrored data on the partner node. The NVRAM is not normally used to flush the contents into the storage media. An exception is during failure conditions because doing so requires an additional read from the storage as virtual NVRAM (vNVRAM) is just another storage media. Instead the data is also stored in memory, which is then flushed.

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 for the instance is very important. In addition, the product requires the presence of a physical RAID controller with a resilient cache for locally attached storage configurations.

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.

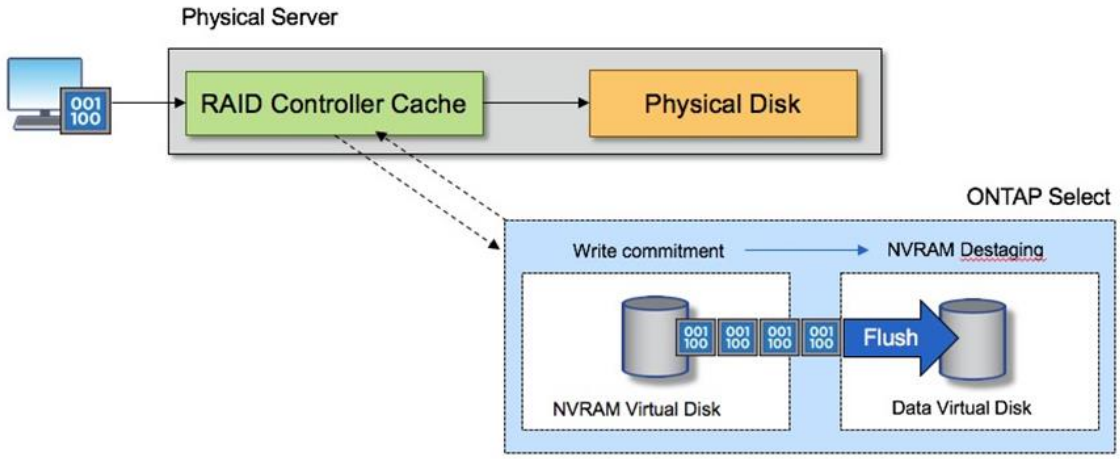
Incoming write requests to the ONTAP Select VM are sent to the VM's NVRAM partition. At the virtualization layer, this partition exists within an ONTAP Select system disk, a virtual disk attached to the ONTAP Select VM. At the physical layer, these requests are cached in the local RAID controller, like all block changes targeted to the underlying spindles. From here, the write is acknowledged back to the host. At this point, the block physically 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. This figure 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 virtual disk 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 the hypervisors are also 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 to the NVRAM partition. Therefore, when choosing a RAID controller, select one with the largest cache available. A larger cache enables less frequent disk flushing and an increase in performance for the ONTAP Select VM, the hypervisor, and any compute VMs collocated on the server.

2.7 Software RAID

In addition to hardware RAID, ONTAP Select 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 system such as FAS. The RAID layer performs drive parity calculations and provides protection against individual drive failures within an ONTAP Select node. With software RAID, ONTAP Select provides deployment agility and flexibility, better management of storage resources, and more control to administrators.

Refer to the section "Storage Configuration with Software RAID" for more information about how to configure storage in a software RAID environment.

A hardware RAID controller might not be available or might be undesirable in some environments, such as when ONTAP Select is deployed on 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:

- Available starting with the ONTAP 9.4 Deploy 2.8 release with the KVM hypervisor
- Requires a premium license
- Supports only SSD drives for root and data disks

- Separates system disks from root and data disks
- 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 multinode setup). The terms service disk and system disk are used interchangeably. Service disks are the individual disks that are used within ONTAP Select to service various items such as clustering, booting, and so on. The system disk is the single disk seen from the host that is used to allocate these service disks.

Note: Hardware RAID is not deprecated. Starting with Deploy 2.8, both hardware RAID and software RAID options are available.

Software RAID confers the following advantages:

- A choice of three different RAID types – RAID 4, NetApp RAID DP®, and NetApp RAID-TEC™
- Better administrative control because disks are seen as-is through Raw Device Mappings (raw device mapping (RDM)). See Appendix E “VirtIO Interface for Raw Device Mapping.”
- More granularity with pre-created partitions
- Flexible allocation of spares to accommodate disk failures
- More deployment options for ONTAP Select.
  - For example, small form-factor ruggedized systems that do not have real estate
  - Support for legacy infrastructure without a RAID controller
- Cost savings from not having a hardware RAID controller on the system
- Drives performance aspects of ONTAP Select

Note: NetApp recommends 8 to 12 drives for the RAID group size. The maximum number of drives supported per RAID group is 24.

Software RAID Requirements

Software RAID has strict requirements for the presence of RAID controllers in the system. The absence of a hardware RAID controller is ideal, but, if a system does have an existing RAID controller, the software RAID controller must satisfy one or more of the following requirements:

- A hardware RAID controller must be absent.
- Or you must disable any hardware RAID controller on the servers such that disks can be presented as belonging directly to a JBOD. This change can be made at the BIOS.
- Some customer systems could have a hardware RAID controller, but they should be configured to be in a SAS JBOD HBA mode. For example, some BIOS configurations enable the Advance Host Controller Interface mode in addition to RAID, which you can chose to enable JBOD mode. This creates a pass-through, so that the physical drives can be seen as-is on the host.

  Note: NetApp recommends a pass-through mode. Depending on the maximum number of drives supported by the controller, you might require an additional controller.

- With the SAS HBA mode, verify that the I/O controller (SAS HBA) is supported with a minimum speed of 6Gbps. However, NetApp recommends a speed of 12Gbps.

No other mode is supported. For example, some controllers only allow a RAID 0 support that can artificially enable disks to pass-through. However, the side effects can be undesirable.

Also be aware of these additional requirements:

- Supported size of physical disks: 200GB to 16TB
- Only supported on DAS configurations
- Only supported with the KVM hypervisor with RedHat/CentOS Linux 7.4
- Only supports SSDs
- Only available with a premium ONTAP Select capacity license

Note: Physical drives are still accessible directly from the host. Therefore, administrators must keep track of which drives are in use by the ONTAP Select VM and prevent inadvertent use of those drives on the host. For example, you should not mount volumes in the physical drives used by the ONTAP Select VM.

Note: Hot spares are optional but recommended. ONTAP gives out warnings and event management system (EMS) messages if hot spares are not specified during the creation of aggregates. Spare disks are required if a physical disk failure occurs and the disk must be replaced.

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

Licensing Consumption Changes

Note that the following licensing changes apply starting from Deploy 2.8 and affect both hardware RAID and SWRAID equally.

- Service (System) Disks are not counted ~ 130GB/node (System Disks -Boot Image/CoreDump, NVRAM, Mediator)
- Root (aggr0) disks are not counted ~ 70GB/node
- Parity and spare disks are not counted.
- Only space for data aggregates (and twice the space for mirrored aggregates) will be taken into account for the capacity license.
- Licensing enforcement is performed from within Deploy only for hardware RAID but not for software RAID. In addition, licensing enforcement for capacity will be done from within ONTAP during the data aggregate creation time in both the hardware RAID as well as the software RAID cases.

This has the advantage of making the licensing calculations intuitive and straight-forward.

Table 5 illustrates the minimum number of drives required in Software RAID configurations for each RAID type.

Table 5) Minimum number of physical disks for various Software RAID configurations.

Cluster Size	RAID Type	Minimum Drives Required	Layout (Disk Types)
Single node	RAID 4	4*	1 service**
			1 parity***
			1 parity***
	RAID DP	6*	1 service
			2 parity
			3 data
	RAID-TEC	8*	1 service
			3 parity
			4 data
Multinode (for each node)	RAID 4	7*	1 service
			2 x 1 parity
			2 x 2 data

Cluster Size	RAID Type	Minimum Drives Required	Layout (Disk Types)
	RAID DP	11*	1 service
			2 x 2 parity
			2 x 3 data
	RAID-TEC	15*	1 service
			2 x 3 parity
			2 x 4 data

* A spare disk is optional and is not counted toward the license (Add 1 above).

** A service (or system) disk is not counted toward the license.

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

Level	9 Single Node
RAID 4	Requires 1 parity drive and can survive 1 disk failure
RAID DP	Requires 2 parity drives and can survive 2 disk failures
RAID-TEC	Requires 3 parity drives and can survive 3 disk failures

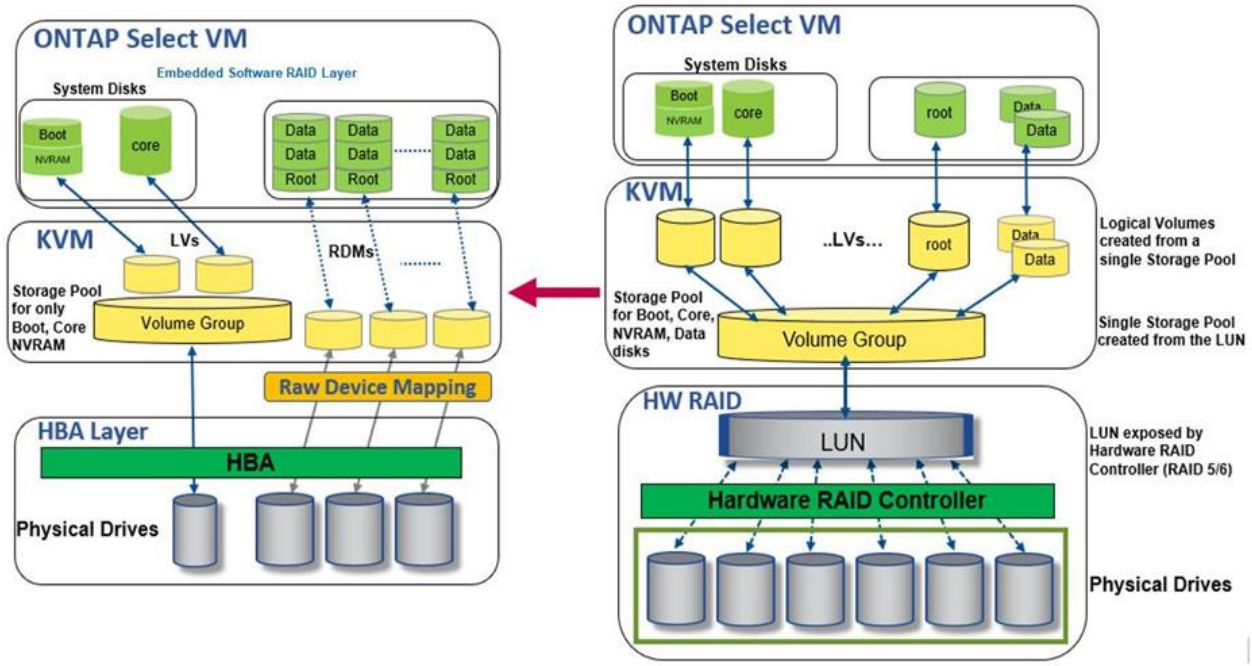
Comparing Software RAID to Hardware RAID

Table 6) Comparison between software RAID and hardware RAID.

Attributes	Software RAID	Hardware RAID
Back-end storage	DAS only	DAS and external array
Disk type	SSD only	SSD and HDD
RAID levels	RAID 5, RAID DP (double parity), and RAID-TEC (triple parity)	Underlying hardware RAID (usually RAID-5/6)
Storage pool (Volume Group) requirements from LUN	Storage pool (single LUN-SSD or NVMe) now only required for system disks**. Root (aggr0) and data disks are carved out of a node's allocated physical SSDs.	Single storage pool (LUN exposed by hardware RAID controller) for system, root, and data disks
LUN usage for storage pools	A dedicated SSD LUN required to create system disks. In addition, NVMe LUN option available for system disks.	Same LUN (HDD or SSD) used for system, root, and data disks.
License accounting	System disks and root aggregates (aggr0) are not counted towards license. Parity and spare drives are not counted towards license.	System disks not counted towards license (starting with ONTAP 9.4). ONTAP Select has no knowledge of parity or spare drives.
Data disks	Pre-created as two partitions on all physical disks allocated for a node.	Disks can only be carved out of the storage pool assigned during the initial cluster creation.

Figure 5 shows the basic architectural differences between software RAID and hardware RAID configurations.

Figure 5) Software RAID and Hardware RAID configurations.



The main change with a software RAID configuration is the use of raw disks through RDM. RDM contains 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 VMDISKS seen within the ONTAP Select VM.

The system disks (or service disks) are separated out through a dedicated storage pool. The root and data disks no longer require a storage pool with software RAID. An LVM storage pool using virsh is still required, but only for the system disks, because actual disks are used to create the root and data partitions. See the three-partition layout in Figure 5.

Partitions use ADPv2 schemes as shown below for a single node cluster and a node of an HA pair. (see the webpage ADPv1 and ADPv2 in a nutshell).

P indicates a parity drive. DP indicates a dual parity drive and S indicates a spare drive (Figure 6 and Figure 7).

Figure 6) Sample layout on a 12-disk system.

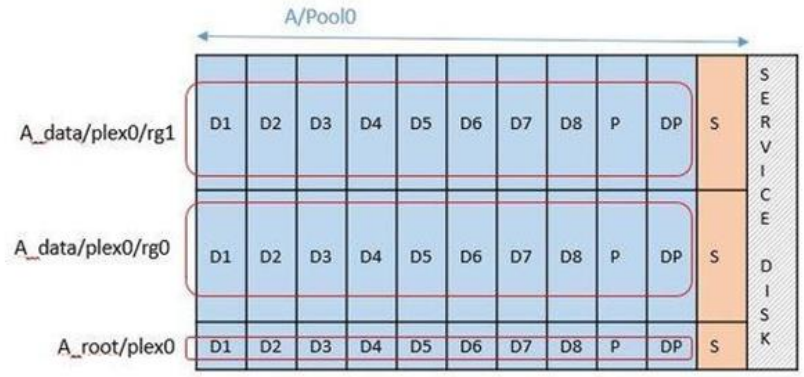
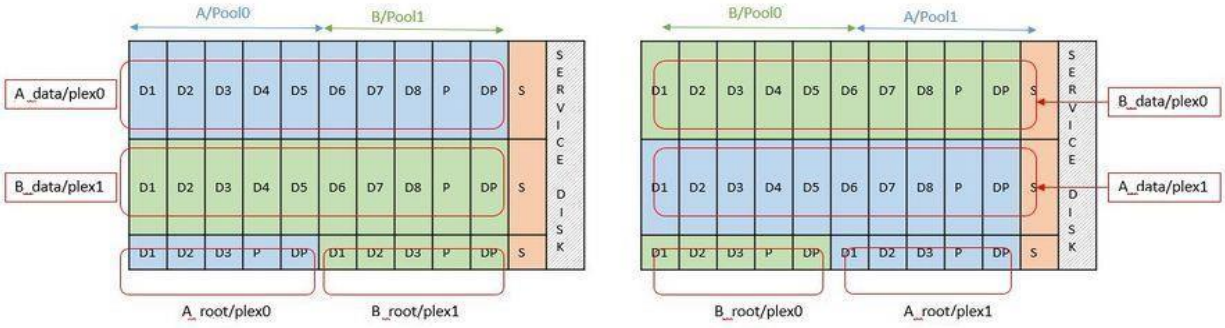


Figure 7) Sample layout on a 12-disk system for an HA pair.

f



Note: For the hardware RAID option in ONTAP 9.4 Deploy 2.8, the root and data disks are created with the same common storage pool as before.

In a multinode configuration, each disk's partition also stores the mirrored data from the other node. Disks are mirrored across the HA pair to provide the shared-nothing capability to provide availability and prevent against node failures. See the section "Storage Configuration with Software RAID."

Software RAID uses the virtio interface to gain direct access to disks. See Appendix VirtIO Interface for Raw Device Mapping.

Upgrade Options with ONTAP Select Deploy 2.8

Customers who want to upgrade from a previous ONTAP version to ONTAP 9.4 must continue using hardware RAID. The aggregates that were already created in a hardware RAID environment cannot be changed to be used in a software RAID configuration.

2.8 V3 Interface

Starting with ONTAP 9.4 Deploy 2.8, ONTAP Deploy features a new interface that that simplifies most cluster creation workflows. It provides more host, node, and cluster CLI and REST API operations that make it easy to manage and automate ONTAP Select clusters.

Table 7 summarizes a few of the main commands in a cluster creation workflow in the previous V2 CLI and the equivalent workflow in the new V3 CLI.

Table 7) High-level cluster creation workflow.

Step	V2 CLI	V3 CLI
Import host Define cluster	Host add	Host register
		Cluster create
		Cluster modify
Configure node • Show disks available on a host • Allocate disks to a node • Show disks allocated to a node	Host configure	Node modify • Host storage disk show • Node storage attach • Node storage disk show
Deploy cluster	Cluster create	Cluster deploy

Having a less integrated set of steps enables more flexibility in associating or disassociating nodes and corresponding storage disks from the cluster. It also facilitates newer software RAID operations while retaining similar functionality for hardware RAID-based operations.

Table 8 outlines some of the main commands with options in the previous V2 cluster creation workflow and shows the equivalent newer forms with the V3 CLI.

Table 8) Major commands for a cluster creation workflow with options.

Step	V2 CLI	V3 CLI
Import hosts	**host add** --host-id <HOST> --vcenter <VCENTER> --username <USER>	**host register** –name <HOST> -management_server <VCENTER> -hypervisor_type<ESX|KVM>
Define cluster		**cluster create** –name <CLUSTER> -num_nodes <1|2|4|6|8>
		cluster modify –name<CLUSTER> -mgmt-ip <IP> -netmask <NETMASK> -gateway <GATEWAY>
Configure node	**host configure** --host-id <HOST> --storage-pool <POOL> --capacity <SIZE GB> --internal-network <PORTGROUP> --data-network <PORTGROUP> --management-network <PORTGROUP> --instance-type <small|medium> --location <string>	**node modify** –cluster-name<CLUSTER> -name<NODE> -host-name<HOST> -instance-type <small|medium> -mgmt-ip <IP> -mgmt-network <PORTGROUP> -data-network <PORTGROUP> -internal-network <PORTGROUP>
Add storage		**node storage attach** --cluster-name <CLUSTER> -name <NODE> -capacity-limit <SIZE GB> -pool <POOL> -mirror-pool <POOL>
Deploy cluster	**cluster create** --name <CLUSTER> --cluster-mgmt-ip <IP> --netmask <NETMASK> --gateway	**cluster deploy** –name <CLUSTER>

Step	V2 CLI	V3 CLI
	<GATEWAY> --node-mgmt-ips <IP1> <IP2> --node-hosts <HOST1> HOST2>	

Starting with version 2.8, Deploy enables a more dynamic clustering environment for ONTAP Select, including support for multiple nodes on the same hosts and possible movement of ONTAP Select VMs across hosts.

Previous ONTAP Select Deploy versions stored the cluster configuration within the hypervisor host. Starting with Deploy 2.8, the configuration will be stored within the ONTAP Select node itself, making the workflows and APIs for cluster creation independent of the host. Furthermore, the V3 REST API conforms to the general ONTAP 9.4 REST API guidelines making it standardized across the NetApp portfolio.

Note: The Deploy V3 API and CLI is not backward compatible with the Deploy V2 API and CLI. However, V3 represents a more stable version that is not subject to any fundamental changes. Thus, it is required to properly plan for the migration to the newer V3 format.

2.9 High Availability Architecture

Customers are starting to move application workloads from enterprise-class storage appliances to software-based solutions running on commodity hardware. However, expectations for resiliency and fault tolerance have not changed. A high-availability solution providing a zero recovery point objective (RPO) is required, Customer data must be protected from loss due to a failure in any component in the infrastructure stack.

A large portion of the software-defined storage (SDS) market uses 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 synchronous replication features (RAID SyncMirror) provided by ONTAP to store an additional 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.

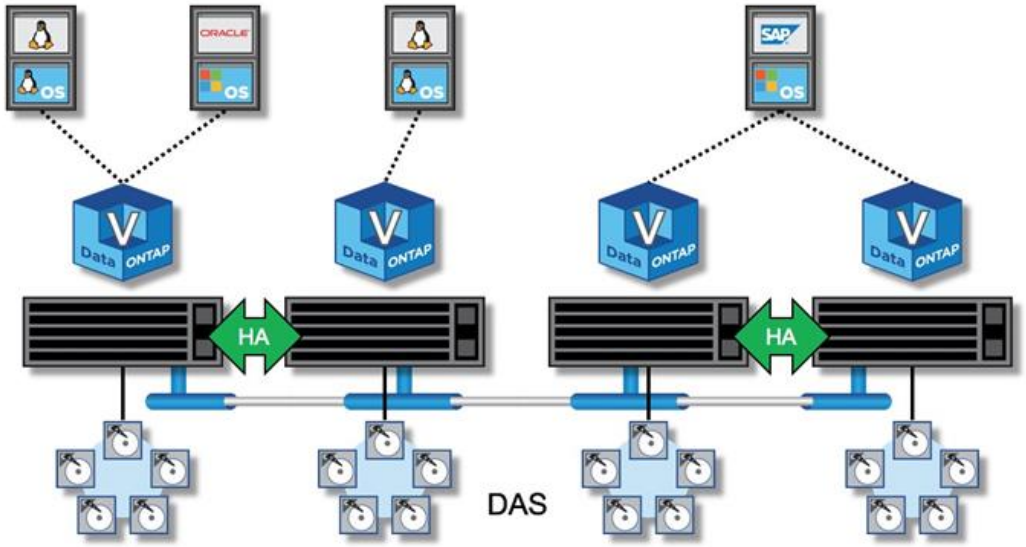
Note: In an ONTAP Select cluster, synchronous replication functionality is a function of the HA implementation, not a replacement for the asynchronous SnapMirror or SnapVault replication engines. Synchronous replication cannot be used independently from HA.

ONTAP Select high availability is provided through a multinode cluster.

Apart from the single-node cluster, there are four multinode ONTAP Select clustering options: the four-node cluster and the two-node cluster introduced with ONTAP Select 9.3 and ONTAP Deploy 2.6. Starting with ONTAP Deploy 2.7, six-node and eight-node clusters are available.

A four-node architecture is represented in Figure 8.

Figure 8) Four-node ONTAP Select cluster using local attached storage.

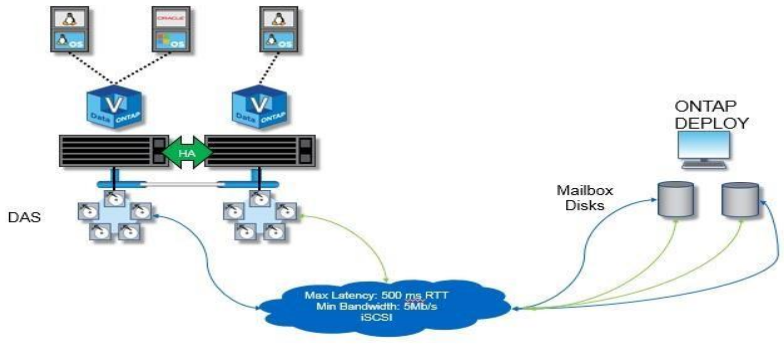


Note: The four-node ONTAP Select cluster is composed of two HA pairs. Within each HA pair, data aggregates on each cluster node are synchronously mirrored, and in the event of a failover there is no loss of data.

Note: A conversion between cluster types is not possible. For example, a four-node cluster cannot be converted to a six-node cluster, and a two-node cluster cannot be converted to a single node cluster.

A two-node Select cluster requires 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, in addition to maintaining the cluster database and providing the orchestration service (Figure 9).

Figure 9) Two-node ONTAP Select cluster using local attached storage.



Note: Only one ONTAP Select instance can be present on a physical server. That instance is tied to the server, meaning that the VM cannot be migrated to another server. 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.

Cluster Quorum in a Multinode Cluster

A multinode cluster, in this context, refers to two-node, six-node, or eight-node clusters. Unlike FAS arrays, ONTAP Select nodes in an HA pair communicate exclusively over the IP network. Therefore, the IP network is a single point of failure, and protecting against network partitions and split-brain scenarios becomes an important design aspect. For example, the four-node cluster can sustain single-node failures because the cluster quorum can be established by the three surviving nodes. This is the function of the mediator that runs transparently on each of the nodes of the cluster. The heartbeat network traffic between the ONTAP Select nodes and the ONTAP Deploy mediator service is minimal. The mediator configuration is automatically performed during setup, and no further administrative action is required.

Epsilon Transfer

An epsilon provides additional voting power for a tie, for example when two equal sides of a four-node cluster try to determine which one is the active pair. Here are some scenarios during a failure:

- When a node fails, the surviving node of a HA pair tries to transfer the epsilon to a healthy pair. If the transfer succeeds, the cluster survives.
- If the second node of a HA pair dies before it can transfer the epsilon to a healthy pair.
  - For a four-node cluster, the cluster goes down, and there will be an outage.
  - For a six-node or an eight-node cluster, a cluster algorithm running re-establishes a new epsilon and determine which one of the other two or three pairs will hold the new epsilon.

Note: The mediator for one pair runs on the other pair and conversely. This also applies to six or eight-node clusters.

Cluster Quorum in a Two-Node Cluster

A two-node cluster relies on the mediator service hosted by the ONTAP Deploy VM to achieve cluster quorum. At a minimum, ONTAP Deploy VM version 2.6 is required to support a two-node cluster with the mediator service. The heart-beat network traffic between the ONTAP Select nodes and the ONTAP Deploy mediator service is minimal and resilient. Therefore, the ONTAP Deploy VM can be hosted in a different data center than the ONTAP Select two-node cluster. Please note that 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 unavailable, then 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 latency of 125ms round-trip time (RTT) are required to provide proper functioning of the cluster quorum.

If an ONTAP Deploy VM acting as a mediator is temporarily or permanently unavailable, a secondary ONTAP Deploy VM (minimum version 2.6) 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. However, it can participate in the cluster quorum algorithm.

The iSCSI protocol is used for communication between the ONTAP Select nodes and the ONTAP Deploy VM. The ONTAP Select node management IP address is the initiator, and the ONTAP Deploy VM IP address is the target. The ONTAP Deploy-hosted mailbox disks are automatically created and marked to the proper ONTAP Select node management IP addresses at the time of two-node cluster creation. Therefore, the complete configuration is performed automatically during setup and no further administrative action is required. The ONTAP Deploy instance creating the cluster is the default mediator for that cluster.

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

Synchronous Replication

The ONTAP HA model is built on the concept of HA partners. ONTAP Select extends this architecture into the nonshared commodity server world by using the RAID SyncMirror functionality in ONTAP. RAID SyncMirror replicates data blocks between cluster nodes, providing two copies of user data spread across an HA pair.

Note: This product is not intended to be a Metro Cluster Configuration (MCC)-style disaster recovery replacement and cannot be used as a stretch cluster. Cluster network and replication traffic occurs using link local IP addresses and requires a low-latency, high-throughput network. Therefore, locating cluster nodes across long distances is not supported.

Mirrored Aggregates

An ONTAP Select cluster is composed of four nodes and contains two copies of user data synchronously mirrored across HA pairs over an IP network. This mirroring is transparent to the user and is a property of the aggregate assigned at the time of creation.

Note: All aggregates in an ONTAP Select cluster must be mirrored to make sure of data availability if there is a node failover. Mirroring also helps avoid a single point of failure 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:

- A local set of disks, contributed by the current ONTAP Select node
- A mirror set of disks, contributed by the HA partner of the current node

Note: Both the local and mirror disks used to build a mirrored aggregate must be of the same size. We refer to these aggregates as plex 0 and plex 1 to indicate the local and remote mirror pairs, respectively. The actual plex numbers might be different in your installation.

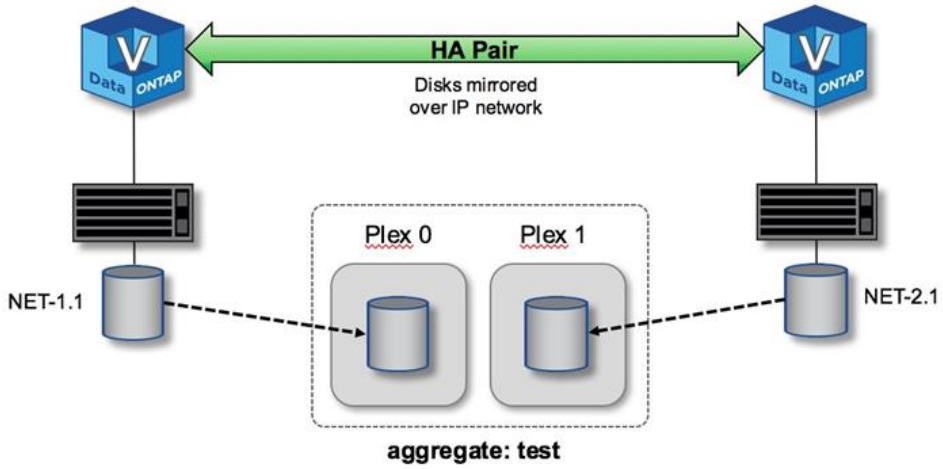
This is an important point and fundamentally different from the way standard ONTAP clusters work. This applies to all root and data disks within the ONTAP Select cluster. Because the aggregate contains both local and mirrored copies of data, an aggregate that contains N virtual disks offers N/2 disks'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

Figure 10 depicts 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 10, we have 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 plex 1. 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 10) ONTAP Select mirrored aggregate.



Note: When an ONTAP Select cluster is deployed, all virtual disks present on the system are autoassigned to the correct plex, requiring no additional action from the user with respect to disk assignment. This prevents the accidental assignment of disks to the incorrect plex and makes sure of optimal mirror-disk configuration.

Best Practice

The existence of the mirrored aggregate is needed to provide an up-to-date (RPO = 0) copy of the primary aggregate. However, care should be taken that the primary aggregate does not run low on free space. A low-space condition in the primary aggregate might cause ONTAP to delete the common NetApp Snapshot™ copy used as the baseline for storage giveback. Although the deletion of common snapshot works as designed to accommodate client writes, 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 Explained

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

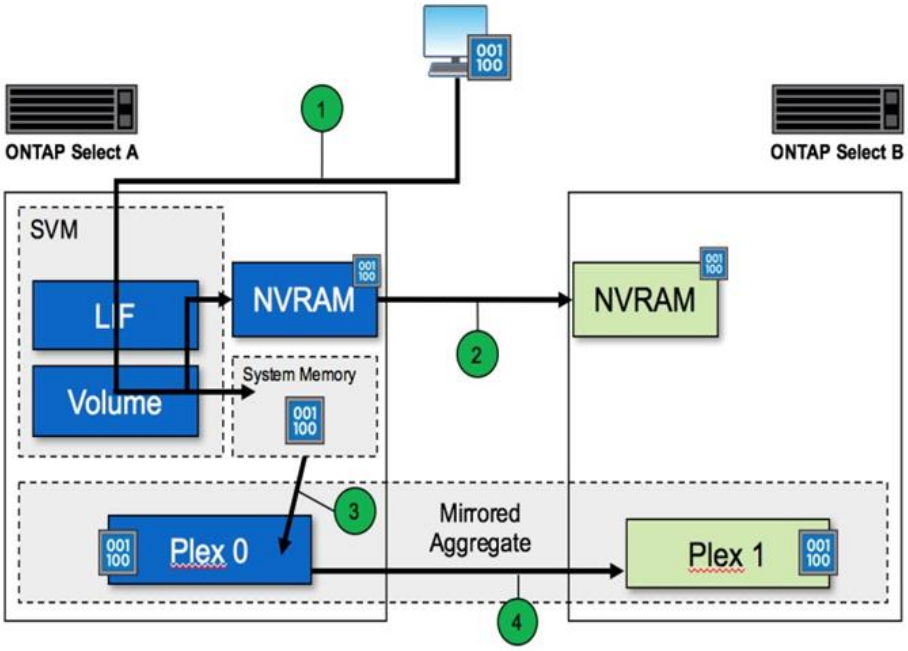
- Acknowledgment
- Destaging

Writes to a target volume occur over a data logical interface (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 step preserves file system consistency on the HA partner node in case of 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 11 shows the write path of an incoming write request to an ONTAP Select node.

Figure 11) ONTAP Select write path workflow.



An incoming write acknowledgment follows these steps:

1. Writes enter the system through a logical interface owned by 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, it is then acknowledged back to the client.

ONTAP Select destaging from NVRAM to the data aggregate (ONTAP CP) follows these steps:

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

Disk Heartbeat

Although the ONTAP Select HA architecture uses 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 heart beating, a nonnetwork-based method of communication used by cluster nodes to prevent network isolation from causing split-brain behavior. Split brain 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, and ONTAP does this through a customized disk-based method of heart beating. 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 that use a shared-storage HA architecture, ONTAP resolves split-brain issues through with the following features:

- SCSI persistent reservations
- Persistent HA metadata
- HA states sent over the 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 across an iSCSI network, so intelligent failover decisions can be made by a cluster node based on access to these disks. If a node is able to access the mailbox disks of all cluster nodes outside of its HA partner, it is likely to be up and healthy.

Note: The mailbox architecture and disk-based heart-beating methods for resolving cluster quorum and split-brain issues are some of the reasons that multinode variants of ONTAP Select require four, six, or eight separate nodes.

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 stating that the node is up and running. Within a healthy cluster at any given 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 node C. Also, cluster node D cannot post to the mailbox of node C. If so, then it's likely that node C is either down or network isolated and should be taken over.

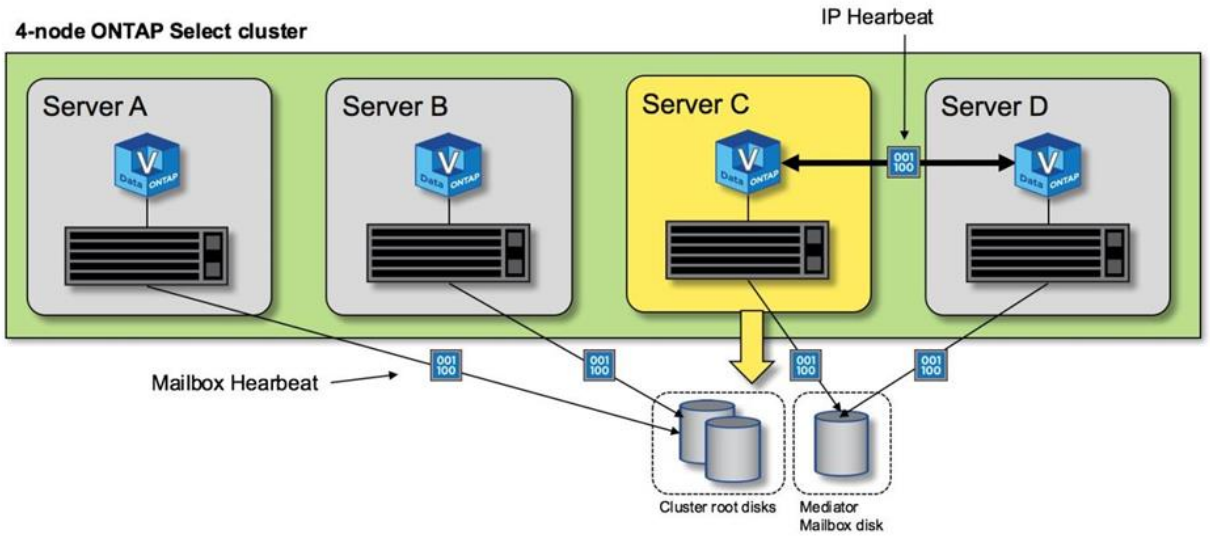
HA Heart Beating

Like NetApp FAS systems, ONTAP Select periodically sends HA heartbeat messages over the HA interconnect. Within the ONTAP Select cluster, this is done 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 messages are sent and received allows the ONTAP Select cluster to detect HA failure events within approximately 15 seconds, the same window available on FAS systems. When heartbeat messages are no longer being read, a failover event is triggered.

Figure 12 illustrates the process of sending and receiving heartbeat messages over the HA interconnect and 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 12) HA heart beating: steady state.



2.10 Disaster Recovery with ONTAP MetroCluster SDS

Starting with ONTAP 9.3 Deploy 2.7, ONTAP Select provides a simple disaster recovery solution by creating a campus or a stretched cluster called MetroCluster SDS. This lower-cost synchronous replication option is part of the MetroCluster business continuity solution from NetApp. It is available only with ONTAP Select, unlike NetApp MetroCluster, which is available on FAS Hybrid Flash, NetApp AFF arrays, NetApp Private Storage for Cloud, and NetApp FlexArray®.

MetroCluster SDS extends the functionality of a two-node ONTAP Select cluster by stretching the distance requirements between the two nodes of a single HA pair.

MetroCluster SDS provides an active-active solution and automated failure handling, and thus provides continuity for business applications by creating high-availability across two different sites. See Table 9 for a list of key differences between MetroCluster SDS and other NetApp replication solutions.

Table 9) NetApp replication and recovery solutions.

Description	HW	Distance	Sys Perf	Conn	Sync/ Async	Capacity	Storage	Bus. Critical	Protection Type
MetroCluster	NetApp	Up to 300KM	High	FC and FCP	Sync	Petabytes	NetApp	Yes	Aggr
MetroCluster SDS	Comm odity	-10KM	Medium	Ether net	Sync	400TB /node	DAS	Yes	Node
SnapMirror / SnapValult	NetApp	Long distance	High	Ether net	Async	Petabytes	NetApp	No	Vol

Here are some key points to consider if you implement a MetroCluster SDS system:

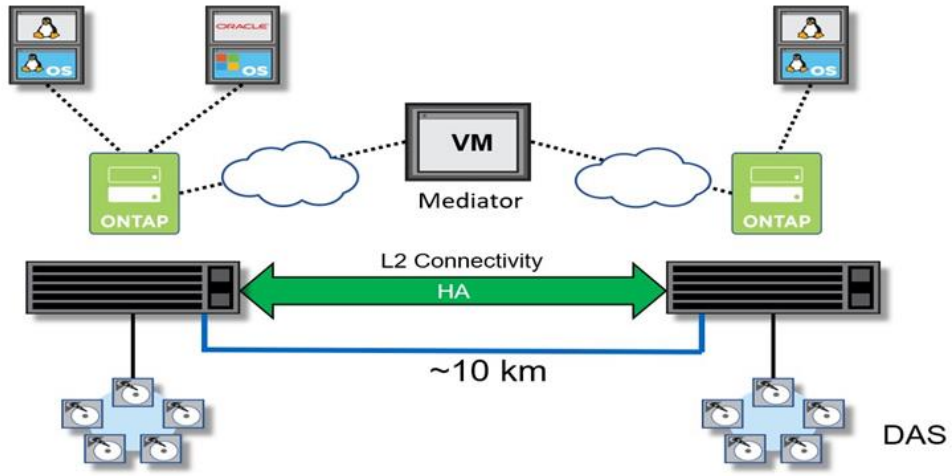
- Latency must be within 5ms RTT with an allowance for up to 5ms in jitter.
- Suitable for distances from 200m to approximately 10km.
- Might require a tunneling protocol such as VXLAN to traverse L3 networks in case a direct L2 network is not available.

- Requires a Premium license, is DAS-only, and works for both SSDs and HDDs
- Not supported for software RAID

A MetroCluster SDS solution should not require any special software or hardware equipment and can be supported by existing networking infrastructure. However, when the internode network traverses multiple switches, special tunneling schemes might be required (for example, L2 over L3 VXLAN networks or hop L3 networks).

A mediator is required as with any two-node cluster. However, the mediator should be located on a third location to avoid the effect of a catastrophic site failure if it were to be co-located on one of the two sites. The mediator should have a maximum latency of up to 250ms between each node of the HA pair. The data-serving ports and interconnect ports must connect to the same first switch. Performance is expected to be proportional to the latency between the nodes (Figure 13).

Figure 13) MetroCluster SDS.



Two-Node Stretched HA (MetroCluster SDS) Best Practices

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

1. After installing ONTAP Deploy, define two hosts (one in each data center) that will be 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.

The connectivity checker does not mark the test as failed if the latency exceeds 10ms. In other words, check the value of the latency instead of the status of the connectivity checker test run.

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 allow for any network overlay protocol overhead. The internal network MTU can be configured between 7,500 and 9,000.

This is a requirement for all HA traffic, whether the ONTAP Select cluster consists 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 EST or 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 or VST is supported for the entire two-node cluster.

3 Deployment and Management

This section covers the deployment and management aspects of an ONTAP Select cluster.

3.1 ONTAP Select Deploy

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

Deploy puts the configuration information into the ONTAP Select template and checks if the host is compatible and has adequate hardware capabilities. During host configuration, each host is assigned a unique ID (UID), and Deploy verifies the management IP addresses, licenses, and serial numbers. Deploy starts scanning the network for the other Select VM instances advertising their UIDs, and, as it finds them, Deploy pulls them into the cluster. This process happens when the VMs boot up as a part of cluster configuration.

When Deploy creates the Select VM instances in the cluster, it gets all the node (host)-specific information such as IP addresses for internal interfaces, HA partners, and so on. It also sets the virtual NICs and the corresponding IP addresses to be assigned to the individual VMs and finally makes the cluster available for configuration using ONTAP management tools.

Note: All nodes in a cluster are configured at the same time during the initial setup, and the nodes boot up simultaneously. Conversion of an existing single-node configuration into a clustered configuration is not allowed. Expansion by adding more Select VMs into an existing cluster is not supported.

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

- CLI
- REST API
- GUI

The ONTAP Deploy CLI is shell-based and immediately accessible upon connecting to the installation VM using Secure Shell (SSH).

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 by using Open API.

Specification (originally known as the Swagger Specification) and can be accessed at https://{IPaddress of Deploy}/api/v3/ui.

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. You can perform standalone upgrades from prior Deploy

versions directly to ONTAP Deploy 2.8. Earlier Deploy versions required an upgrade to ONTAP Select first.

Server Preparation

Although ONTAP Deploy provides the user with functionality for the 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 steps:

- For local storage, the RAID controller (Not for the software RAID option) and attached local storage are configured. If you are using ONTAP software RAID, verify that the correct drive type and number of drives is available.
- Verify physical network connectivity to the server. Network resiliency, speed, and throughput are critical to the performance of the ONTAP Select VM.
- You must install the hypervisor and all the associated dependent packages.
- Virtual networking constructs (OVS, bridges, and bonds) are configured.

Note: After the ONTAP Select cluster has been deployed, use the appropriate ONTAP management tools to configure 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 made available as a raw file. The components are available from the NetApp Support site.

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

- 2 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. The Deploy VM is used to verify hypervisor minimum requirements, deploy Select clusters, apply the license, and optionally troubleshoot network connectivity between Select nodes during setup.

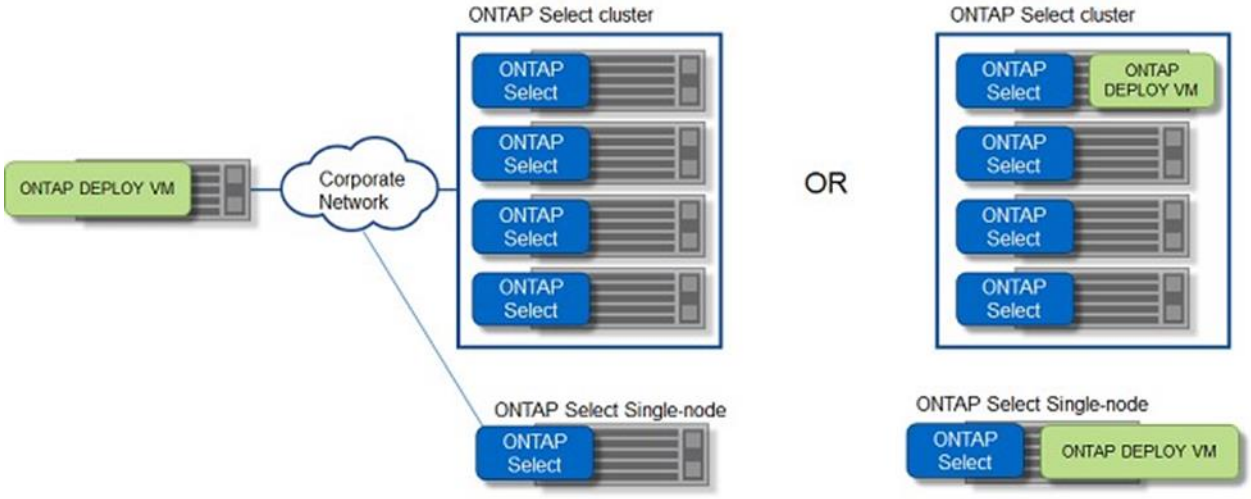
VM Placement

You can place the ONTAP Select installation (Deploy) VM on any host in your environment. It can be collocated on the same host as an ONTAP Select instance or on a separate host. You should install the Deploy VM in the same data center as the ONTAP Select cluster. In addition, there should be network connectivity between the ONTAP Select Deploy VM, the targeted KVM hosts, and the ONTAP Select cluster management IP addresses. Note that creating an ONTAP Select cluster requires copying and configuring the ONTAP Select binary files into individual nodes of the cluster.

Note: For a two-node cluster, the Deploy/mediator could be located externally on the WAN as long as the minimum bandwidth is 5Mbps and the maximum latency is 125ms RTT.

Figure 14 shows these deployment options.

Figure 14) ONTAP Select installation VM placement.



ONTAP Deploy can be installed like any regular VM in a Linux-KVM hypervisor environment.

The following example shows how the `virt-install` command can be used to install the Deploy tool.

```
virt-install --name=deploy-kvm --vcpus=2 --ram=4096 --os-type=linux  -controller=scsi,
model=virtio-scsi --disk
path=/home/deploypath/ONTAPdeploy.raw,device=disk,bus=scsi,format=raw -network
"type=bridge,source=ontapbr,model=virtio,virtualport_type=openvswitch" --console=pty --import --
wait 0
```

Multiple ONTAP Select Deploy Instances

Depending on the complexity of the environment, it might be beneficial to have more than one ONTAP instance.

You might want the Deploy instance to manage the ONTAP Select environment. If so, 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 NetApp does not recommend managing an ONTAP Select cluster using multiple ONTAP Deploy instances.

When deciding whether to use multiple installation VMs, keep in mind that, although 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, you should only use one Deploy instance per L2 network to manage existing or create new Select clusters or nodes.

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.

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 pool 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 the ONTAP Select licensing have continued to evolve. As mentioned previously, 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. For example, the ONTAP Select root aggregate no longer counts against the capacity license.

Also, the cluster create workflow in the web UI 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 to then 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 involved the license enforcement mechanism. With earlier versions of ONTAP Select, the VMs 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 user data (using a hardware RAID controller as opposed to ONTAP software RAID), the user can 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. Alternatively, it 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 and 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 virtual SAN (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 Failure to Tolerate (FTT) / Failure to Tolerance method (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 UI 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.

Capacity Tiers Licensing

A capacity tiers license 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 license files you must administer is equal to or exceeds the number of nodes you have created or plan to create.

Summary of the 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.
- **License lock**. Each license is locked to a specific ONTAP Select node.
- **License duration**. The license is perpetual, and renewal is not required.
- **Node serial number**. The node serial number is nine digits long and is the same as the license serial number.

Capacity Pool Licensing

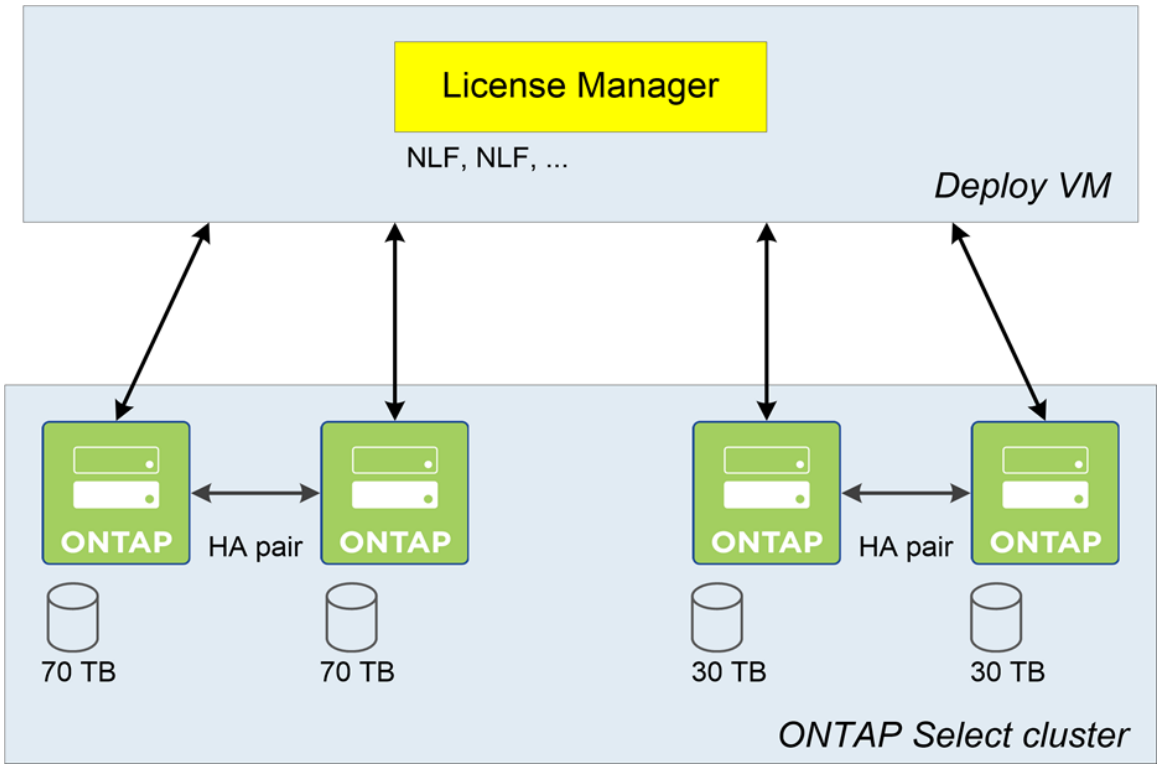
A capacity pool license is a new licensing model provided beginning with ONTAP Select 9.5 using Deploy 2.10. It provides an alternative to the capacity tiers model. The capacity pools licensing model provides several benefits, including the following:

- 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—a new software component introduced with ONTAP Select Deploy 2.10—manages the capacity pool licenses and leases. License Manager 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 License Manager. If a valid lease cannot be acquired, the data aggregate operation fails. The lease duration for each pool can be configured 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 License Manager. This is done through the License Lock ID (LLID), a unique 128-bit number identifying each License Manager instance (and therefore each Deploy instance). You can locate the LLID for your Deploy instance on the web UI 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 the 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 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.
- **License lock**. Each license is locked to a specific License Manager instance.
- **License duration**. The license is valid for a limited term (such as one year) and must be renewed.
- **Node serial number**. The node serial number is 20 digits long and is generated by the License Manager.

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 similarly results 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 Network Time Protocol [NTP])
- ONTAP Select cluster and 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. The older ONTAP Deploy version must have knowledge of and access to the ONTAP Select node so that the ONTAP Deploy upgrade process can append identifying information to that VM's metadata. This unique identifier is stored in the VM's metadata and the ONTAP Deploy database. Then future changes to the ONTAP Select cluster/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.

3.3 ONTAP Management

Because ONTAP Select runs ONTAP, it supports all common NetApp management tools. Therefore, 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 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, various 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 Deploy. ONTAP Deploy communicates with the ONTAP Select clusters it creates using information configured at the time of deployment. This information includes the host name or IP address and also the ONTAP Select cluster management IP address.

Changing the ONTAP Select cluster management IP address from inside ONTAP Select after deployment does not affect clients accessing the ONTAP Select data LIFs. However, it does prevent ONTAP Deploy from managing that ONTAP Select cluster. Changing the KVM host name or IP address is not supported.

After the installation, ONTAP Deploy can also 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. This 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 enables you to generate a NetApp AutoSupport® bundle containing the ONTAP Deploy logs. The GUI also allows you to download the bundle for immediate inspection.

If the cluster setup was successful but there are issues postdeployment, the Deploy command line is the only interface for connecting to the ONTAP Select node serial console using the `node console connect` command. Additional node-level commands available in the Deploy CLI include `node start/stop` and `node coredump/savecore`. These commands are disruptive and should not be invoked without guidance from Technical Support.

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

Given that Deploy plays an important role in troubleshooting the environment, the ONTAP Deploy database should be backed up regularly and after every change in the environment. It is not currently possible to rediscover an ONTAP Select cluster that was created by a different instance of ONTAP Deploy. Also, if you have an unmanaged cluster, you lose some important troubleshooting functionality. The ONTAP Deploy configuration database can be backed up using the ONTAP Deploy CLI and issuing the '`configuration backup`' command. With Deploy 2.8, this command has been changed to `deploy backup create`.

3.4 ONTAP Deploy Cluster Refresh

Starting with ONTAP 9.3 Deploy 2.6, ONTAP Deploy allows you to change to the ONTAP Select cluster or node and the host's VM properties. It also provides mechanisms to subsequently update these parameters on-demand.

Initiate a cluster refresh operation if the state of the Deploy database, which stores cluster and node information, becomes stale when the host or the Select VM node information change outside of Deploy. Changes can occur due to the direct actions of a Linux administrator directly on the host outside of ONTAP or through ONTAP administration. If these changes are not updated, Deploy might not be able to perform further operations on the cluster, such as a storage add operation or a more serious loss of quorum.

Deploy communicates with ONTAP to receive updated information about items such as the following:

- Cluster identity (name, universal unique identifier [UUID])
- Node identity (name, UUID)
- Networking information (the cluster and node management IP addresses, netmask, and gateway)
- ONTAP version
- Deploy can also communicate with the host to receive updated information about the VM and host information (VM state and name, storage pool name, and network names).

Specific use cases for a cluster refresh operation from Deploy included the following:

- If the ONTAP Select node names or cluster names are changed with ONTAP interfaces
- If the node or cluster management IP addresses are changed, for example, as part of corporate policies or security
- If the Select VM is migrated to a new host. Because Deploy is not involved in the live migration, it will carry stale information. The Deploy database could be updated with the new host's information.
- If the VM is shut down or restarted using virsh and Deploy is not aware. Deploy can contact the host for any VM and host information and update its database.

Deploy maintains unique entries for nodes and corresponding clusters using UUIDs generated by Deploy during the ONTAP Select node and cluster creation process. These UUIDs are persistent for the lifetime of the nodes and clusters and can be used to identify ONTAP Select nodes, clusters, and the Select VMs as seen from the Linux host.

The UUIDs form the metadata information and are stored as part of the VM's information maintained by the hypervisor host and not within ONTAP.

The following is the representation of the metadata:

```
{'cluster_uuid': uuid_from_ontap, 'node_uuid': uuid_from_ontap}
```

Cluster Refresh Examples

1. The VM state changed from online to offline or conversely with the virsh command, but Deploy still shows a stale state of the VM.

   In the Deploy GUI, hit the refresh button on the cluster page, or, in the CLI, enter the following command:

```
cluster refresh --cluster-name <cluster-name> --username admin
```

2. If the names for the Select node and the VM are changed and need to be updated, then, in the Deploy GUI, hit the refresh button on the cluster page, or, in the CLI, enter the following command:

```
cluster refresh --cluster-name <cluster-name> --username admin
```

3. When VMs are renamed, supply the node names (from Deploy) whose VM names need to be updated.

   In the Deploy GUI, hit the refresh button on the cluster page, or, in the CLI, enter the following command:

```
cluster refresh --cluster-name <cluster_name> --username admin --node-names <Node name>
```

4. When ONTAP Select VMs are migrated to a different host, supply node host IDs (FQDN) and node names (from Deploy) for the hosts that need to be updated.

   In the Deploy GUI, hit the refresh button on the cluster page, or, in the CLI, enter the following command:

```
cluster refresh --cluster-name <cluster_name> --username admin --node-names <Node name> -node-
hosts <Node Host ID>
```

5. When the node management IP addresses (Select VM IP addresses) or the cluster management IP addresses need to be updated.

   In the Deploy GUI, hit the refresh button on the cluster page. The new cluster management IP address can be provided in the GUI dialog box. In the CLI, enter the following command:

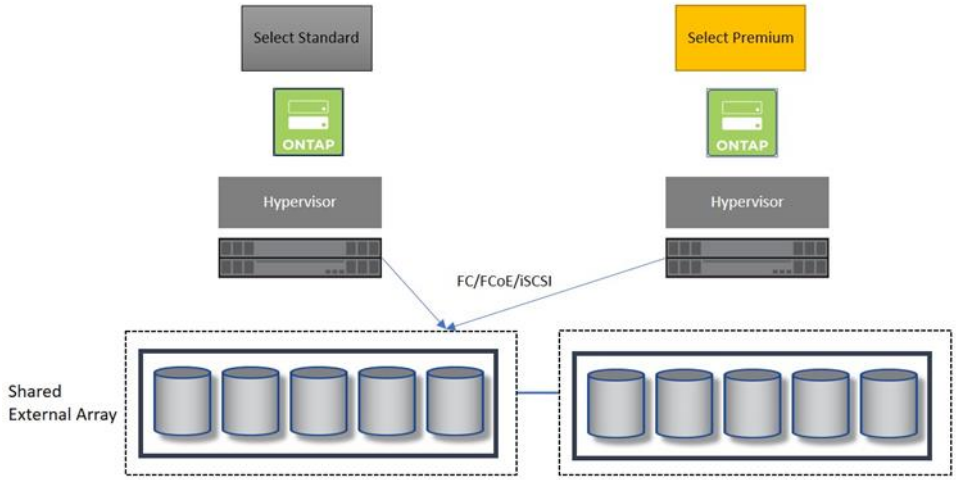
```
cluster refresh --cluster-name <cluster_name> --username admin --cluster-mgmt-ip <newcluster-
mgmt-ip>
```

3.5 Shared External Storage Management

The external storage can be exposed to the host as FC/FCoE LUNs or as iSCSI storage pools. LVM storage pools can be created out of these LUNs or pools that can then be used within ONTAP Select as VMDISKs.

You might need to configure shared external storage so that multiple hosts can use the same storage and thus enable activities such as ONTAP Select VM migration from one host to another. To do so, you must set up a Linux host high-availability and clustering infrastructure outside of ONTAP Select or ONTAP Deploy. For this, you must install more components on the underlying hosts such that the hosts and thus the ONTAP Select VMs have the same synchronized view of the back-end storage (Figure 16).

Figure 16) ONTAP Select shared storage configuration with single-node clusters.



Host Clustering using Clustered Logical Volume Manager

Clustering on the underlying host is required when the ONTAP Select VM on a single-node cluster on host A must migrate to a different host, say host B. Note that Host B is not already running an ONTAP Select VM instance.

CLVM, a set of clustering extensions to logical volume management (LVM), is one of the primary components needed to manage storage from a logical pool on storage provisioned from external array with shared storage. CLVM allows a cluster of hosts to manage shared storage (for example, on a SAN). Logical volumes created with CLVM on shared storage are visible to all hosts that access that shared storage.

An administrative action is required to create an HA cluster using the pacemaker configuration system, distributed lock manager (DLM), and CLVM resources. Clusters are managed with Pacemaker. CLVM logical volumes are supported only along with Pacemaker clusters and must be configured as cluster resources. After the resources are set up, creating and managing logical pools and volumes on storage provisioned by an external storage array are no different than creating storage provisioned from directly attached storage.

Pacemaker on Linux provides cluster infrastructure functionality through cluster management, lock management, fencing, and cluster configuration management. It performs all cluster-related activities, such as monitoring cluster membership, managing the services and resources, and fencing cluster members.

Perhaps only one node of your system requires access to the storage you are configuring as logical volumes. Then you can use LVM and create storage pools without CLVM extensions, and the logical volumes created with that node are all local to the node. If more than one node of your cluster requires access to your storage, which is then shared among the active nodes, then you must use CLVM. CLVM allows a user to configure logical volumes on shared storage by locking access to physical storage while a logical volume is being configured. LVM uses clustered-locking services to manage the shared storage.

The CLVMD daemon is the key clustering extension to LVM. The CLVMD daemon runs on each cluster host and distributes LVM metadata updates in a cluster, presenting each cluster computer with the same view of the logical volumes.

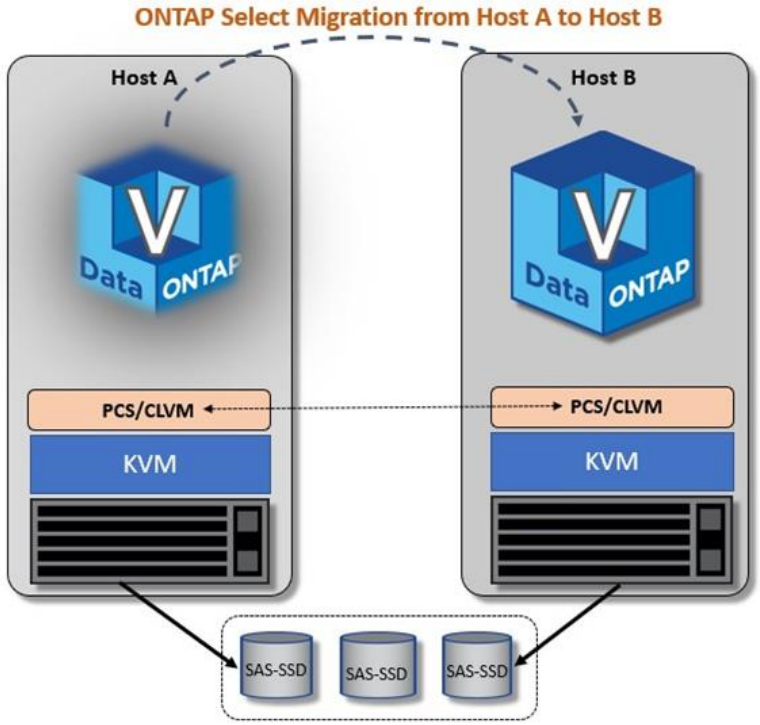
Note: See Appendix C, “Linux External Shared Storage and Host Clustering” for more details on external storage, host clustering setup, and VM migration in Linux.

3.6 ONTAP Select VM Migration

Migration is the process of moving a guest VM from one physical host machine to another. It is possible to migrate ONTAP Select VMs that form a single-node cluster on one host onto other hosts that are not already running ONTAP Select. ONTAP Select must be able to access the same storage after it has migrated to a new host. This is possible only if both of the underlying hosts share access to the same back-end external storage. Both live migration of online VMs and migration of offline VMs is possible. For more details on migration, refer to Appendix D, Guest VM Migration on Linux (Figure 17).

Note: Migration is not supported for software RAID.

Figure 17) ONTAP Select migration.



Migration of VMs can be used for the following use cases:

- **Load balancing. G**uest VMs can be moved to host physical machines with lower usage when their host physical machine becomes overloaded.
- **Hardware independence.** When we need to upgrade, add, or remove hardware devices on the host physical machine, we can safely relocate guest VMs to other host physical machines. This means that guest VMs do not experience any downtime for hardware improvements.
- **Energy saving.** Guest VMs can be redistributed to other host physical machines and can thus be powered off to save energy and cut costs in low usage periods.
- **Geographic migration.** Guest VMs can be moved to another location for lower latency or potentially in emergency scenarios.

Requirements for ONTAP Select VM Migration

- ONTAP Select must be installed on shared storage using one of the following protocols:
  - FC/FCoE-based LUNs
  - iSCSI pools
- Appropriate TCP/IP ports should be open if a firewall is in use.

  - Specifically, ports 49152:49261 should be open. For example: `iptables -I INPUT -p tcp –dport 49152:49261 -j ACCEPT`.

- A separate (external) system exporting the shared storage medium must be available. Storage should not reside (direct-attached) on either of the two physical host machines being used for migration.
- Shared storage must mount at the same location on the source and destination systems. The mounted directory names must be identical. Although it is possible to keep the images using different paths, NetApp does not recommend it.

  Note: If you use virt-manager to perform migration, the path names must be identical. If, however, you use virsh to perform the migration, different network configurations and mount directories can be used with the help of the –xml option or prehooks when performing the migrations.

- When you attempt migration on an existing guest VM in a public bridge+tap network, the source and destination physical host machines must be located in the same network. Otherwise, the Select VM network cannot operate after migration.
- The network and storage resources must be identical to the source host where the Select VM was running previously.
- The same virtual LANs (VLANs) must be set up for the vnet (tap) interfaces associated with the Deploy VM virtual port.
- Allow some time for the VM IP (node management IP) to be reachable again.
- The libvirtd service must be enabled and running.

Note: Refer to Appendix D, “Guest VM Migration on Linux” for more details on guest VM migration in Linux.

3.7 Multinode vNAS and Multiple Instances (Nodes) per Host

Multinode vNAS

Multinode vNAS is a feature that is specific to external arrays. Starting with ONTAP 9.4 Deploy 2.8, a multiple node cluster can be deployed on external arrays. Before this release, only single-node ONTAP Select cluster instances were allowed with external arrays. Multinode vNAS enables a lot of flexibility for using external arrays. However, the IT or the storage administrator has to carefully consider how storage LUNs can be provisioned along with CLVM to enable private access for each ONTAP Select node.

Note that even though the underlying external storage might be shared, the ONTAP Select architecture with respect to an HA pair is still based on mirroring and shared-nothing.

Multinode vNAS can be used in cases in which each ONTAP Select node in the cluster has the following properties:

- Has separate LUNs on the same external array of which the administrator is fully aware. Each node must have its own private LUN. This configuration does not require CLVM.
- Uses the same external storage. In this case, the administrator does not need to worry about individual LUNs. However, the administrator needs CLVM for the creation of a consistent pool that can be seen across all the hosts and for the synchronization of access. For example, four nodes can participate and share the same LVM pool.

Both of the nodes, each on different hosts, use the same external array, and thus they need to use CLVM to have a consistent view of the LVMs. CLVM also enables live migration in case the host is brought down for maintenance. Note that both nodes on the same cluster can be on the same host temporarily. Event messages show warnings that this is not a permitted configuration.

```
[root@sdot-b200-013 ~]# virsh list --all
 Id    Name                           State
----------------------------------------------------
```

```
 1     KVM_sharedPool-01             running                  ---> This is the node from 2node
cluster created on Host sdot-b200-013 with shared external pool(CLVM) root@sdot-b200-013 ~]#
lvscan
  WARNING: Not using lvmetad because config setting use_lvmetad=0.
  WARNING: To avoid corruption, rescan devices to make changes visible (pvscan --cache).
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-02_DataONTAPv.raw' [10.74 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-01_DataONTAPv.raw' [10.74 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-02_coredisk' [120.00 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-01_coredisk' [120.00 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-02_sdotconfig.iso' [4.00 MiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-01_sdotconfig.iso' [4.00 MiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-02_root_1' [68.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-01_root_1' [68.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-02_root_2' [68.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-01_root_2' [68.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-02_ontap-select-storagepool_1'
[241.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-01_ontap-select-storagepool_1'
[241.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-02_ontap-select-storagepool_2'
[241.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_sharedPool-01_ontap-select-storagepool_2'
[241.00 GiB] inherit
```

The second host runs another instance (the second node) of an ONTAP Select instance belonging to the same ONTAP Select cluster as the node above. In this case, it uses a CLVM pool because both of the nodes are sharing the same external array.

```
[root@sdot-b200-014 ~]# virsh list --all
 Id    Name                           State
----------------------------------------------------
 1     KVM_sharedPool-02              running         ---> This is the node from 2 node cluster
created on Host sdot-b200-014 with shared external pool(CLVM)
[root@sdot-b200-014 ~]# virsh pool-list --all
 Name                 State      Autostart -----------------------------------------  ontap-
select-storage-pool active     yes
```

Note: Use best practices specified by the external array vendor. Configure the external array with CLVM in a fashion that allows a consistent view from multiple hosts.

Multiple Instances (Multiple Nodes) per Host

Multiple-instances support enables multiple ONTAP Selects nodes, each belonging to different ONTAP Select clusters, to reside on the same physical host. This results in better utilization of host resources.

Multiple instances can be used with ONTAP Select nodes in the cluster sharing the same storage pool. This storage pool can be from an external array, which requires that you configure CLVM so that nodes on different hosts can access the same shared array.

Note: Multiple instances are not supported for software RAID. Rather, multiple instances are supported primarily for external arrays. A warning event is generated at the time of creation when a second instance is found using the same DAS.

The following example shows a single host running two ONTAP Select nodes, each from a different cluster using a DAS-based storage pool.

LVMs for multiple Select nodes can be carved out from a single large volume group, for example on a host with DAS (Figure 18, Figure 19, and Figure 20).

```
[root@sdot-b200-015 ~]# virsh list --all
 Id    Name                           State
```

```
-----------------------------------------------------
1       KVM_NonShared-01               running        --> This is the node from a 2-Node cluster
created on Host sdot-b200-015 with a non-shared pool
2       KVM_Non_shared_SN-01           running        --> This is the single node cluster created
on Host sdot-b200-015 with a non-shared pool
[root@sdot-b200-015 ~]# lvscan
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-01_DataONTAPv.raw' [10.74 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-01_coredisk' [120.00 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-01_sdotconfig.iso' [4.00 MiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-01_root_1' [68.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-01_root_2' [68.00 GiB] inherit
ACTIVE           '/dev/ontap-select-storage-pool/KVM_NonShared-01_ontap-select-storage-pool_1'
[241.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-01_ontap-select-storage-pool_2'
[241.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_Non_shared_SN-01_DataONTAPv.raw' [10.74
GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_Non_shared_SN-01_coredisk' [120.00 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_Non_shared_SN-01_sdotconfig.iso' [4.00
MiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_Non_shared_SN-01_root_1' [68.00 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_Non_shared_SN-01_ontap-select-
storagepool_1' [301.00 GiB] inherit
  ACTIVE            '/dev/rhel/swap' [1.00 GiB] inherit
  ACTIVE            '/dev/rhel/root' [<8.00 GiB] inherit
```

The other node of one of the clusters in the above example, for the two-node cluster, resides on a different host:

```
[root@sdot-b200-016 ~]# lvscan
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-02_DataONTAPv.raw' [10.74 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-02_coredisk' [120.00 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-02_sdotconfig.iso' [4.00 MiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-02_root_1' [68.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-02_root_2' [68.00 GiB] inherit
ACTIVE           '/dev/ontap-select-storage-pool/KVM_NonShared-02_ontap-select-storage-pool_1'
[241.00 GiB] inherit
  ACTIVE            '/dev/ontap-select-storage-pool/KVM_NonShared-02_ontap-select-storage-pool_2'
[241.00 GiB] inherit
  ACTIVE            '/dev/rhel/swap' [1.00 GiB] inherit
  ACTIVE            '/dev/rhel/root' [<8.00 GiB] inherit
```

Figure 18) Configuration showing multinode vNAS and multiple nodes (instances) per host.

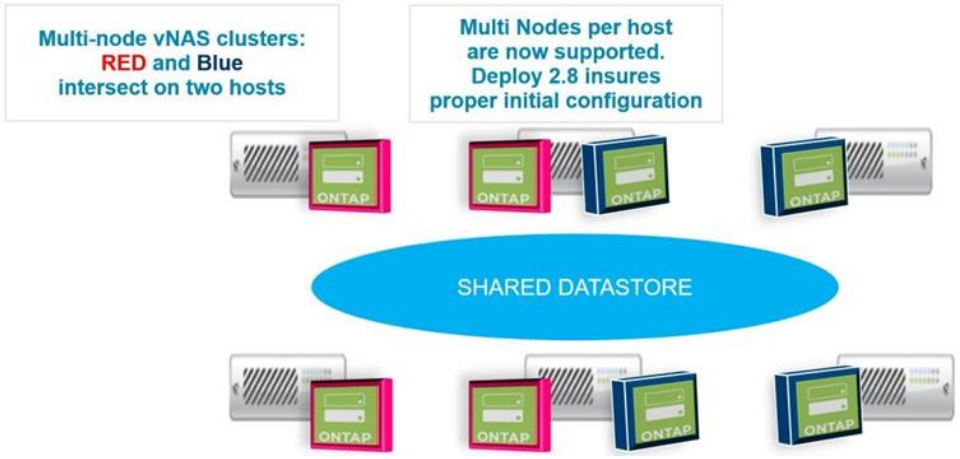


Figure 19) Migration and cluster refresh in a multinode configuration to a different host.

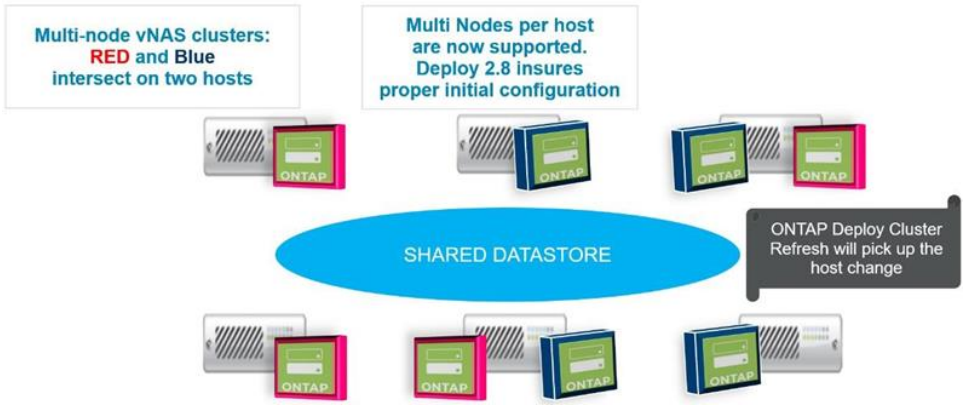
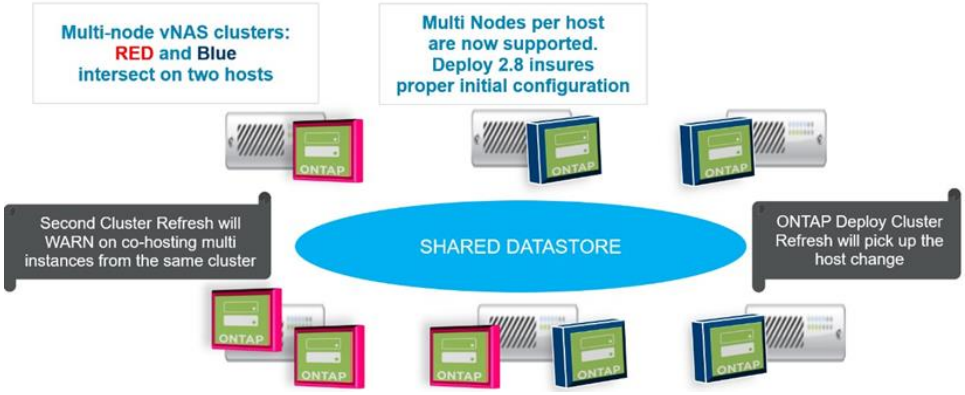


Figure 20) Two nodes of the same cluster ending up on the same host.



4 Network Design Considerations and Supported Configurations

This section covers the various network configurations and best practices that you should consider when building an ONTAP Select cluster in a KVM environment. Like the design and implementation of the underlying storage, you should take care when making network design decisions. These choices have a significant effect on both the performance and resiliency of the ONTAP Select cluster.

Note: Refer to Appendix A, “KVM Networking” for general information about various terminologies, brief concepts, and some of the external industry design practices.

In traditional FAS systems, interface groups 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 uses the underlying hypervisor’s virtualization of multiple physical network interfaces to achieve the same goals of throughput aggregation and resiliency. The network interface cards that ONTAP Select manages are therefore logical constructs, and configuring additional interface groups does not achieve the goals of throughput aggregation or recovering from hardware failures.

4.1 Host Network Configuration

In this section, some of the preinstallation steps to prepare the host to run an ONTAP Select VM instance are highlighted using examples. These steps can be part of a configuration script.

Using standard configuration files to store network information helps keep the host in a consistent state across host reboots.

1. Check if the host has two or four physical NIC ports. Suppose that this configuration found four physical ports (eno0, eno1, ens2f0, and ens2f1), all of which support 10Gb speeds. For a single-node configuration, the NICs could support 1Gb speeds.
2. Use all of these NIC ports to create an OVS bridge.

```
ovs-vsctl add-br ontap-br
```

3. Create an OVS bond within this bridge consisting of the physical ports.

```
ovs-vsctl add-bond ontap-br bond-ontap eno0 eno1 ens2f0 ens2f1 bond_mode=balance-slb lacp=active
other_config:lacp-time=fast
```

4. Set MTU for each port to 9000.

```
ip link set eno0 mtu 9000 up
ip link set eno1 mtu 9000 up
ip link set ens2f0 mtu 9000 up
ip link set ens2f1 mtu 9000 up
```

5. Set up the network configuration for persistency across host reboots. Under the `/etc/sysconfig/network-scripts/` directory, set up individual `ifcfg-<port>` files for each of the physical ports, the bridge, and the bond. The files should look like the following examples that have these entries:

   a. **Physical Ports.** File name `ifcfg-eno0`

```
DEVICE=eno0
TYPE=Ethernet
MTU=9000
ONBOOT=yes
```

   b. **Bridge.** File name `ifcfg-ontap-br`

```
DEVICE=ontap-br
DEVICETYPE=ovs
TYPE=OVSBridge
MTU=9000
ONBOOT=yes
BOOTPROTO=static
IPADDR=192.168.10.10 (optional)
PREFIX=24 (optional)
GATEWAY=192.168.10.1 (optional)
```

   c. **Bond.** File name `ifcfg-bond-ontap`

```
DEVICE=bond-ontap
DEVICETYPE=ovs TYPE=OVSBond
MTU=9000
OVS_BRIDGE=ontap-br
```

```
BOND_IFACES=bond-ontap
OVS_OPTIONS=bond_mode=balance-slb lacp=active other_config:lacp-time=fast
ONBOOT=yes
```

6. Set rules to set the MTU for the vnet (tap) devices to 9000. This is required to persist the vnet port MTU to 9000 across host reboots.

   Add the following to the file `/etc/udev/rules.d/70-persistent-net.rules`.

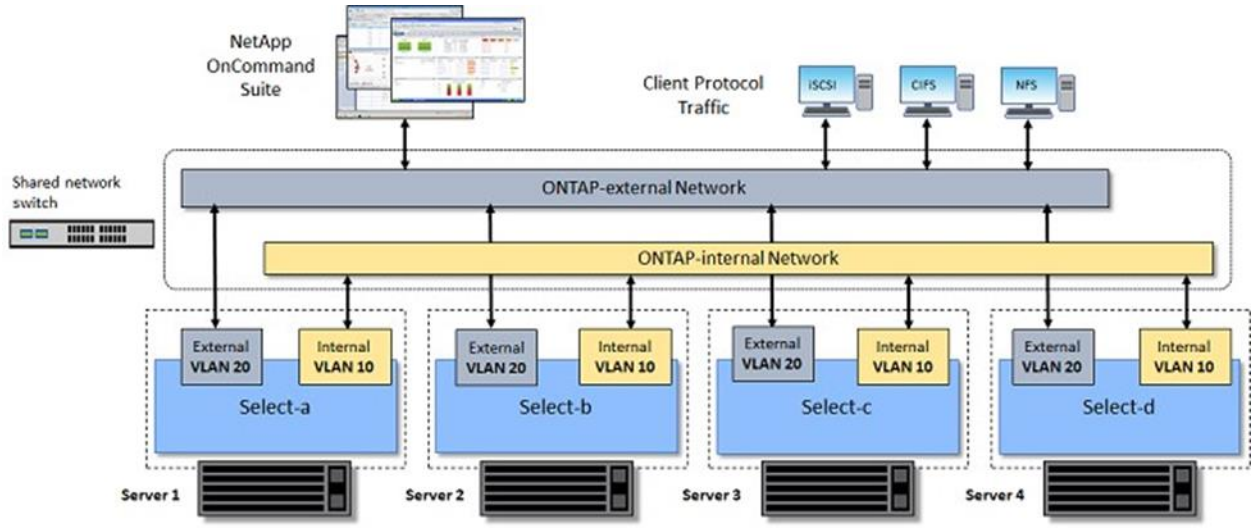
```
SUBSYSTEM=”net”,ACTION=”add”,KERNEL==”ontapn-*”,ATTR{mtu}=”9000”
```

4.2 Network Configuration: Multinode

A multinode ONTAP Select network configuration consists of two networks: 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 important for cluster resiliency.

These networks are represented in Figure 21, which shows a four-node ONTAP Select cluster running on a KVM hypervisor system. Note that each ONTAP Select instance resides on a separate physical server. In addition, internal and external traffic is isolated by using VLANs, which allow the cluster nodes to share the same physical switch infrastructure.

Figure 21) ONTAP Select multinode network configuration.



ontapn- (previously called vnet or tap) devices on ONTAP Select VMs are created during host configuration. Each ONTAP Select VM contains six virtual network ports (tap devices: ontapn0 through ontapn-5) that are presented to ONTAP as a set of six network ports, e0a through e0f. Although ONTAP treats these tap devices as physical NICs, they are virtual and map to a set of physical interfaces through a virtualized network layer. Therefore, 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:

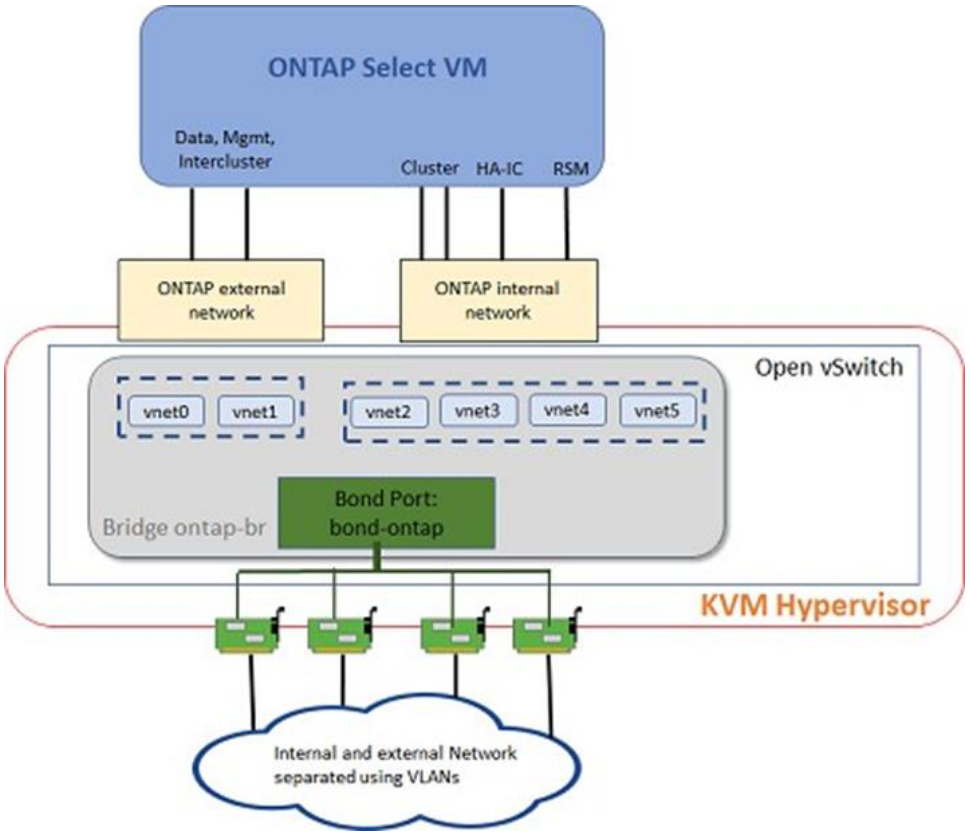
- **e0a, e0b, and e0g.** Management and data LIFs
- **e0c and e0d.** Cluster network LIFs
- **e0e.** RAID SyncMirror
- **e0f.** HA interconnect

Ports e0a and e0b 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 L2 network. You do not need to separate these virtual adapters across different networks. You must associate the external network ports with routable physical interfaces to serve NAS or iSCSI traffic as shown in Figure 14.

Network separation is achieved by tagging ontapn- devices for internal and external networks using different VLAN IDs. OVS bonds aggregate links into a single group to provide increased throughput as well as to achieve redundancy when one of the links fails.

The relationship between these ports and the underlying physical adapters can be seen in Figure 22, which depicts one ONTAP Select cluster node on the KVM hypervisor. This configuration uses a single bond (port group) to group all four physical interfaces. Because there is a single bridge having a single bond, segregation between internal and external traffic occurs by using different VLAN tags.

Figure 22) Network configuration of a multinode ONTAP Select VM.



Segregating internal and external traffic across the physical NICs using VLANs doesn't introduce unintended latencies into the system. However, it does provide clean access to network resources. Additionally, aggregation through bonding makes sure that failure of a single network adapter does not prevent the ONTAP Select cluster node from accessing the respective network.

Network configuration is based on the custom setting on the physical switch and the tagged packets that flow through switches. If the network uses the native tags of the physical switch, then you do not need to tag the network.

Prior to Deploy 2.8, the `host configure` command allowed you to specify the bridge (for example, ontap-br) to be used for the internal, data, and management traffic, with corresponding VLANs. However, starting with Deploy 2.10, the `node modify` command allows you perform this function:

```
node modify --host-id 10.xx.xx.xx --internal-network ontap-br --internal-vlan 10 -data-network
ontap-br --management-network ontap-br --management-vlan 20
```

Note: In this example, the management VLAN setting can be optional and can use the native VLAN. The external network (management and data) traffic can be left untagged, and the internal network can be separated by using a valid VLAN tag. Alternatively, the cluster network (internal network) could always use the native tag of the switch and the data and management network (external network) could use valid VLAN ID tags.

LIF Assignment for a Multinode Cluster

ONTAP Select clusters contain a default and a cluster IPspace. You could place network ports e0a and e0b into the default IPspace and ports e0c and e0d into the cluster IPspace. The remaining ports (e0e and e0f) within the ONTAP Select cluster node are assigned automatically to private IPspaces providing internal services. The RAID SyncMirror and HA interconnect LIFs hosted on ports e0e and e0f are not exposed through the ONTAP shell.

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

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

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

ONTAP ports e0a, e0b, and e0g are candidate ports for logical interfaces that carry the following types of traffic:

- SAN/NAS protocol traffic (CIFS, NFS, and iSCSI)
- Cluster management, node management, 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 must be created post deployment.

Cluster Network LIFs (e0c, and e0d)

ONTAP ports e0c and e0d are home ports for cluster LIFs. 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 nodes must be physically located within close proximity (for example, in a multipack, single data center). Building a stretch cluster configuration by separating nodes across a WAN or across significant geographical distances is not supported.

Note: To create maximum throughput for cluster network traffic, the cluster network port is configured to use jumbo frames (9000 MTU). To provide 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 LIF residing on network port e0e. This LIF is configured by ONTAP during cluster setup and requires no configuration by the administrator. Because this port is reserved by ONTAP for internal replication traffic, neither the port nor the hosted LIF is visible in the ONTAP CLI or management tooling. This LIF is configured to use an

automatically generated-link local IP address, and the reassignment of an alternate IP address is not supported.

Note: This network port requires the use of jumbo frames (9000 MTU).

Throughput and latency requirements that are critical to the proper behavior of the replication network dictate that ONTAP Select nodes be located within close physical proximity. Therefore, building a hot disaster recovery solution is not supported.

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 typically have this type of equipment available (for example, InfiniBand or iWARP devices), so an alternate solution is used. Within an ONTAP Select cluster, the functionality of the HA interconnect (typically provided by hardware) has been designed into the OS and uses Ethernet as a transport mechanism.

Each ONTAP Select node is configured with an HA interconnect port (e0f). This port hosts the HA interconnect LIF, 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 RDMA frames within Ethernet packets. Similar to RAID SyncMirror, neither the physical port nor the hosted ONTAP LIF is visible to users from either the ONTAP CLI or management tooling. As a result, the IP address of this LIF cannot be modified, and the state of the port cannot be changed.

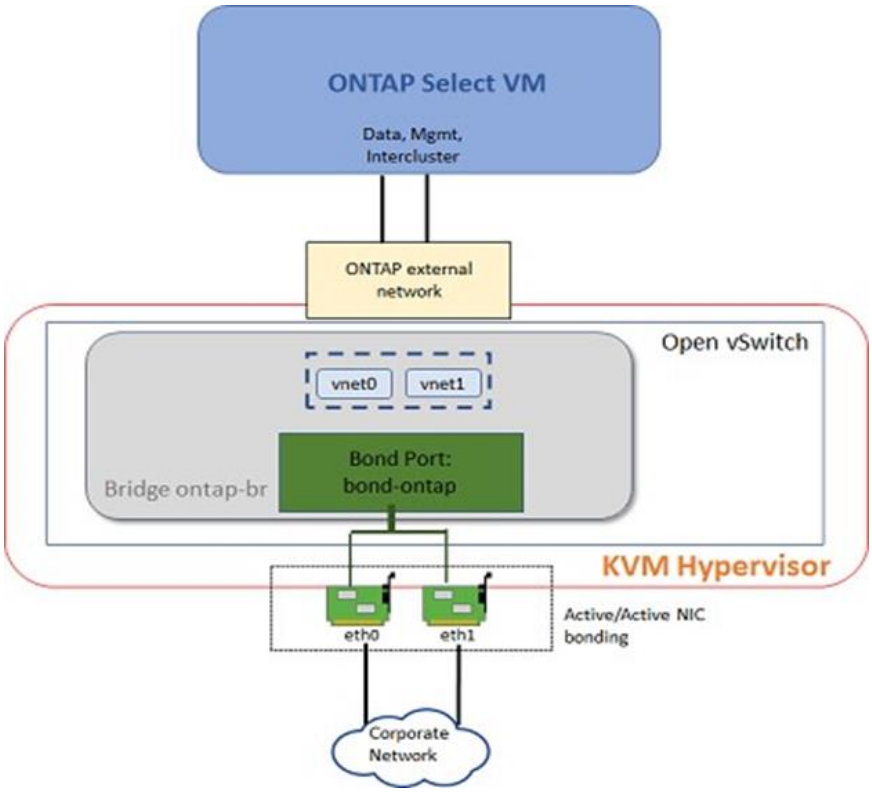
Note: This network port requires the use of jumbo frames (9000 MTU).

4.3 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. The multinode version of ONTAP Select, contains six virtual network ports. However, each ONTAP Select VM contains two virtual network ports (tap devices: ontapn-0 and ontapn-1) that are presented to ONTAP network ports e0a, e0b, and e0g, as shown in Figure 16. These ports are used to provide data and management services.

The relationship between these ports and the underlying physical adapters can be seen in Figure 23, which depicts one ONTAP Select cluster node on the KVM hypervisor.

Figure 23) Network configuration of single-node ONTAP Select VM.



Note that physical NIC bonding is still required, although two adapters are sufficient for a single-node cluster.

LIF Assignment for Single-Node Cluster

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 system 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.

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

ONTAP ports e0a, e0b, and e0g are 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)

4.4 Networking: Internal and External

ONTAP Select Internal Network

The internal ONTAP Select network (only available 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 and e0d. Hosting cluster network LIFs
- e0e. Hosting the RAID SyncMirror 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 aside from intracluster 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 one of the following configurations:

- Assigning a VLAN tag to the internal virtual port (ontapn-2 through ontapn-5)
- Using the native VLAN provided by the upstream switch so that the native VLAN is not used for any other traffic

Internal Network Validation and Troubleshooting

Using the ONTAP Select Deploy tool, the internal network in a multinode cluster can be validated using the network connectivity checker functionality, which can be invoked from the Deploy CLI with the `network connectivity-check start` command. This tool should be run before deploying the cluster.

Run the `network connectivity`-check show `--run-id X` (where 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. The default value must 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 for both single-node and multinode configurations. This network does not have the tightly defined throughput requirements of the internal network. However, the administrator should be careful not to create network bottlenecks between the external network client and the ONTAP VM. Doing so could cause performance issues to be mischaracterized as ONTAP Select problems.

Internal Versus External Network

Table 10 highlights the major differences between the ONTAP Select internal and external networks.

Table 10) Internal versus external network quick reference.

Description	Internal Network	External Network
Network services	Cluster HA/IC RAID SyncMirror	Data Management Intercluster (SnapMirror and SnapVault)
Network isolation	Required	Optional
Frame size (MTU)	7,500 to 9,000[2]	1,500 (default)
NIC aggregation[1]	Required before ONTAP Select 9.3	9,000 (supported)
IP address assignment	Autogenerated	Required before ONTAP Select 9.3
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 Aggregation

To make sure that internal and external networks have the necessary bandwidth and resiliency characteristics required to provide high performance and fault tolerance, physical network adapter port aggregation (NIC bonding) is used. This is a requirement on both the internal and external networks of the ONTAP Select cluster and provides the ONTAP Select cluster with two major benefits:

- Isolation from a single physical port failure
- Increased throughput

NIC port aggregation allows the KVM hypervisor instance to balance network traffic across two physical ports.

Best Practice

If a NIC has multiple ASICs, select one network port from each ASIC when building network aggregation constructs through bonds 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 with a system-specific, organizationally unique identifier (OUI) specific to NetApp. 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.

Note: At no point should the administrator modify the assigned MAC address of a network port.

Best Practice

If you are planning to use multiple NICs, select different NIC vendors to avoid failures on both of the NIC cards at the same time due to driver misbehavior.

4.5 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, offering 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 effect on the overall performance of the ONTAP Select cluster.

Link aggregation is a core construct used to provide sufficient bandwidth to both the external and internal ONTAP Select networks. The Link Aggregation Control Protocol (LACP) is a vendor-neutral standard providing an open protocol for network endpoints used to bundle groupings of physical network ports into a single logical channel

Table 11 shows the various supported configurations. The use of LACP is called out because environmental and hypervisor-specific dependencies prevent all combinations from being supported.

Table 11) Network configuration support matrix.

Client Environment	Select Configuration	Best Practices
• Two or more 10Gb physical ports • Standard OVS • Physical uplink switch supports LACP and large MTU sizes on all ports	• Single LACP channel with all ports. • Both internal and external networks use the same link aggregation group (bond). • For the internal network, the ontapn-interfaces on the host use either no tagging or switch (OVS) VLAN tagging. • For the external network, the ontapn-interfaces use either no tagging or can be tagged at the OVS layer (switch VLAN tagging) by using access ports. • For the external network, the ontapn-interfaces on the host can also be tagged as trunk ports to enable guest VLAN tagging.	• The load-balancing policy at the bond level is “route based on source MAC address hash and output VLAN tag (balance-slb)” on the link aggregation group (bond). • LACP mode is set to active on both the OVS bond and the physical switchports connected to the OVS Bond’s physical interfaces. • The LACP timer should be set to fast (1 second) on the OVS bond and the physical switchports connected to the OVS bond’s physical interfaces. • The Spanning Tree Protocol (STP) should be set to Portfast on the physical switch ports connected to the OVS bond’s physical interfaces.
• Single 10Gb physical port • Standard OVS • Physical uplink switch support • 9,000 MTU size	• Both internal and external networks use the same physical port. Therefore, you do not need to configure an OVS bond. • For the internal network, the ontapn-interfaces on the host use either no tagging or switch (OVS) VLAN tagging. • For the external network, the ontapn-interfaces use either no tagging or can be tagged at the OVS layer (switch VLAN tagging) by using access ports. • For the external network, the ontapn-interfaces on the host can also be tagged as trunk ports to enable guest VLAN tagging.	• Load balancing cannot be performed because there is only a single port. • There is no redundancy. Therefore, LACP cannot be configured. • STP can be configured.

When choosing an ONTAP Select network configuration, the use of LACP, which requires specialized hardware support, might be a primary consideration. Although LACP requires support from both the OVS

switch and the upstream physical switch, it can provide a significant throughput benefit to incoming client protocol traffic.

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

The performance of the ONTAP Select VM is tied directly to the characteristics of the underlying hardware. Therefore, 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 you from designing a system with four 10Gb NIC ports, two 10Gb NIC ports can be used.

Best Practice
To make sure of optimal load balancing across the physical NICs, the load-balancing policy of balance-slb, which is “Route based on originating MAC address hash and the output VLAN,” should be used on the link aggregation group (bond).

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

Best Practice
Only configurations that have LACP configured, both on the physical switch and the OVS switch, are currently supported. Non-LACP configurations are not supported. Only a single bond–single bridge configuration is supported. Multiple bonds are not supported.

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 L2 VLANs.

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

Physical switch ports should be part of the same port channel and should be configured as trunk ports with or without a native VLAN, depending on the VLAN configuration of the internal and external ONTAP Select networks. ONTAP Select external traffic can be separated across multiple layer 2 networks by using ONTAP VLAN-tagged virtual ports to management port e0a, data port e0b, and port e0g.

ONTAP Select internal network traffic separation 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.

Best Practice
NetApp recommends that the STP be set to portfast on the switch ports connected to the KVM hosts. Not setting the STP to portfast on the switch ports might affect ONTAP Select's ability to tolerate uplink failures.

Best Practice

NetApp recommends setting the LACP mode to ACTIVE on both the KVM host and the physical switches. Furthermore, the LACP timer should be set to FAST (1 second) on the port channel interfaces and the NIC bonds.

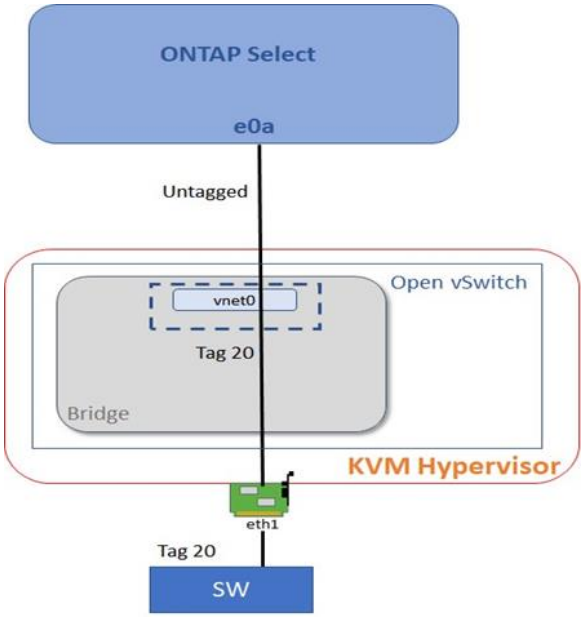
Shared Physical Switch

Figure 24 and Figure 25 depict possible configurations for an ONTAP Select node when sharing a physical switch. In these examples, the physical NICs used by the open switch (bond) hosting the two separate networks are cabled to the same physical upstream switch. Switch traffic is kept isolated by using separate VLANs.

The example in Figure 24 uses untagged frames at the VM that are tagged at the OVS layer. These VLAN-tagged frames are passed out through the eth1 interface to the physical switch.

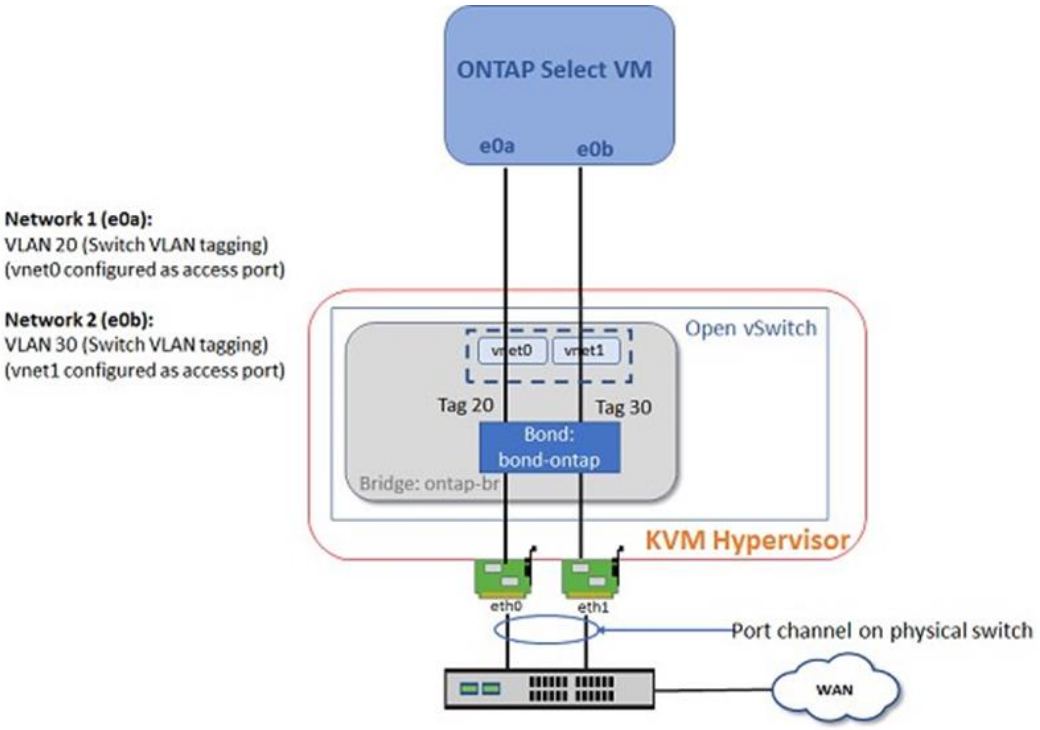
Note: eth1 acts like a trunk and is capable of passing out tagged frames. The physical switch needs no configuration because, by default, most switches support trunking. Therefore, switches can pass tagged frames to ONTAP Select VMs on to the other hosts, for example in between the HA pair of ONTAP Select nodes.

Figure 24) Network configuration showing native VLAN using shared physical switch.



The example in Figure 25 uses switch VLAN tagging at the OVS layer by configuring the vnet ports in access mode with the respective VLAN tags for each network. The physical switch ports in this case are configured as trunk ports without a native VLAN. By default, most physical switches support trunks and pass VLAN tags between Select nodes on separate hosts.

Figure 25) Network configuration using shared physical switch.



Note: In this configuration, the shared physical Ethernet switch becomes a single point of failure. If possible, you should use multiple switches so that a physical hardware failure does not cause a cluster network outage.

Best Practice

By default, all OVS ports are VLAN trunks, so the physical host interfaces pass all VLANs.

The physical switch must be capable of forwarding VLAN-tagged traffic and the physical switch ports should operate as VLAN trunks. Usually this is the default behavior.

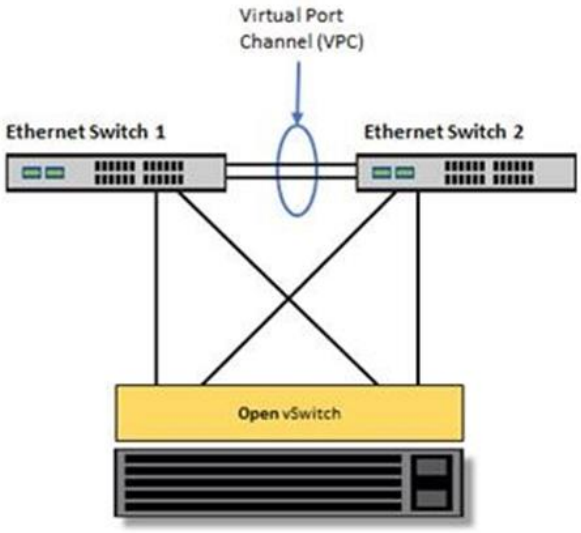
Multiple Physical Switches

When redundancy is needed, you should use multiple physical network switches. Figure 26 depicts a recommended configuration used by one node in a multinode ONTAP Select cluster. NICs from both the internal and external bonds are cabled to 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 following multiswitch configuration because of to the added protection this configuration provides against physical switch failures.

Figure 26) Network configuration using multiple physical switches.



4.7 Data and Management Separation

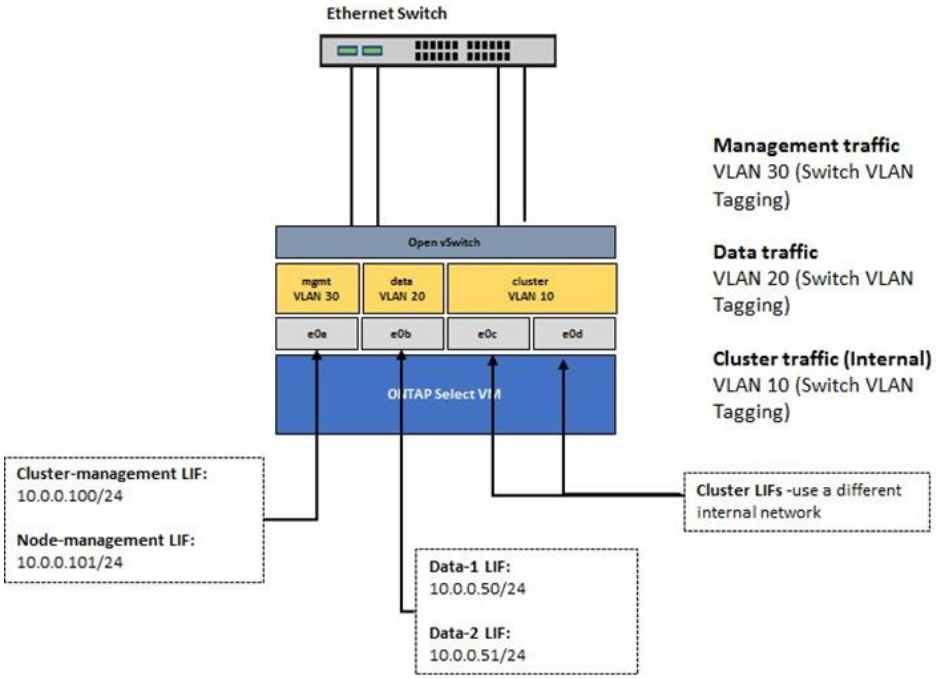
ONTAP Select external network traffic is defined as data (CIFS, NFS, or iSCSI), management, and replication (SnapMirror) traffic. Within an ONTAP cluster, each style of traffic uses a separate logical interface that must be hosted on a virtual network port. For the multinode version of ONTAP Select,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

NetApp recommends isolating data traffic and management traffic into separate L2 networks. In the ONTAP Select environment, this is done with VLAN tags. You can achieve isolation by assigning a VLAN tag to the tap devices ontapn-0 (port e0a) for management traffic and ontapn-1 (port e0b) and port e0g for data traffic.

Note: Data and management network separation through guest VLAN tagging is not available when using the ONTAP Deploy utility. This process must be performed after cluster setup has completed.

Figure 27 shows a scenario using switch VLAN tagging in which traffic is tagged at the OVS layer by assigning different VLAN IDs to the tap devices for data, cluster, and management traffic. In this configuration, node and cluster management LIFs are assigned to ONTAP port e0a and tagged with VLAN ID 30 through the associated ontapn- (vnet) port. The data LIF is assigned to port e0b and given VLAN ID 20, while cluster LIFs use ontapn- port (vnets) associated with cluster ports e0c and e0d on VLAN ID 10.

Figure 27) Data and management separation using switch VLAN tagging.



Best Practice

If data traffic spans VLANs and the use of VLAN ports is required, guest VLAN tagging should be used. This configuration is performed outside of the Deploy tool, after the Select VMs are up, and from within the VM.

4.8 Four-Port, Two-Port, and One-Port NIC Configurations

As described previously, supported network configurations involve permutations based on two and four physical NIC ports. For optimum performance and resiliency, NetApp strongly recommends that the ONTAP Select instance reside on a physical server with four 10Gb NIC ports. NIC bonding is a requirement on both two-port and four-port configurations and having four NIC ports present on the system reduces the potential for network-based bottlenecks between the internal and external networks.

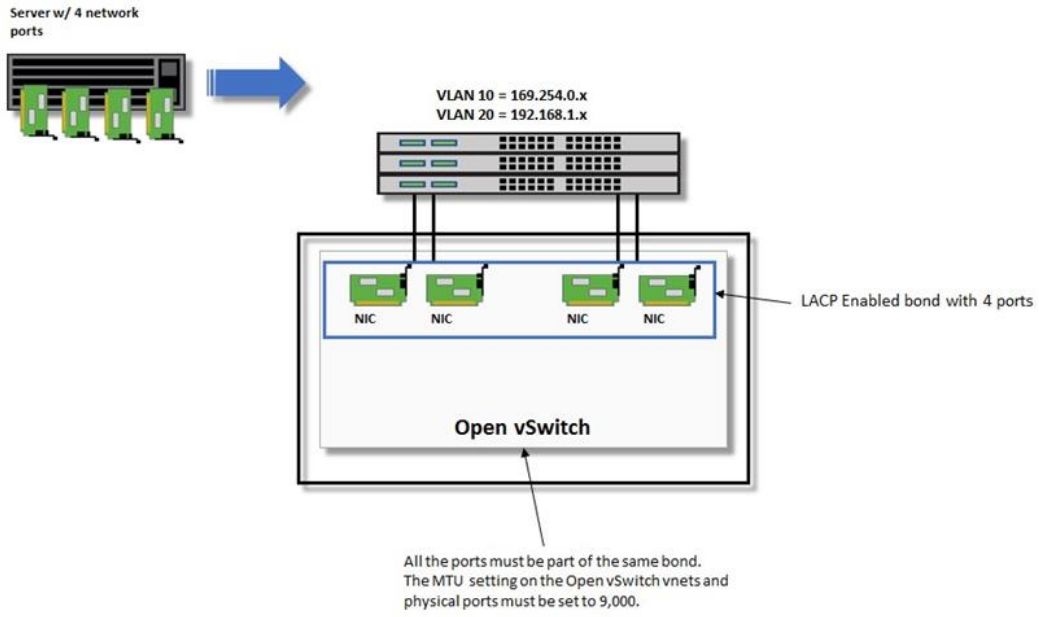
Therefore, for a multinode cluster, the internal ONTAP network requires 10Gb connectivity, and 1Gb NICs are not supported. Trade-offs can be made to the external network, however, because limiting the flow of incoming data to an ONTAP Select cluster does not affect its ability to operate reliably. NIC bonding provides the cluster with increased throughput and resiliency in the event of a NIC failure.

Best Practice

Currently, bonds should consist of NICs with the same speeds. In a heterogeneous configuration with two 1Gb NIC ports and two 10Gb NIC ports, only a single bond consisting of the two physical 10Gb NIC ports should be used. Multinode clusters only support bonds with 10Gb ports, while single-node clusters also support bonds with 1Gb ports.

Figure 28 depicts a way in which to configure the network on a physical server with four physical NIC ports.

Figure 28) Four-port NIC homogeneous network configuration with LACP on OVS.



Note that, in all cases, VLAN tagging for internal network traffic is performed by the ontapn- (vnets) ports (VLAN 10). External traffic, however, is untagged at the OVS layer and instead is tagged by the upstream switch using the native VLAN tag (VLAN 20). This is only intended to highlight one possible way of implementing layer 2 tagging within an ONTAP Select cluster.

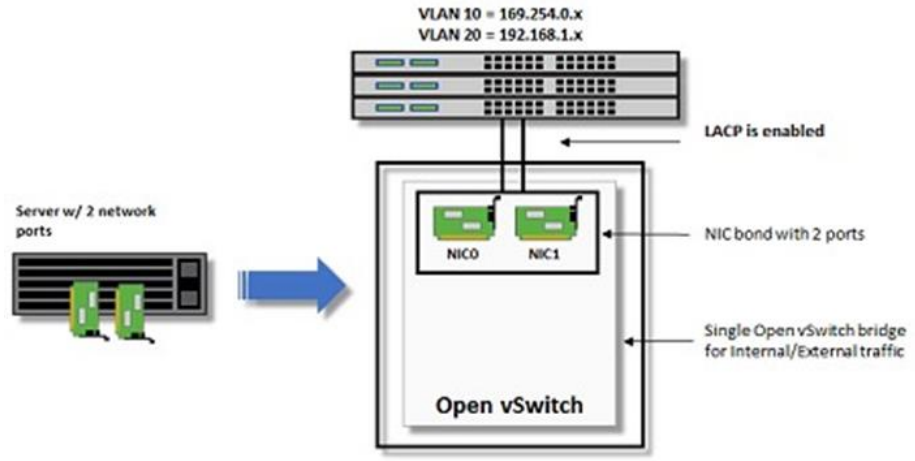
Like the ONTAP internal bond, a static VLAN ID can also be assigned to the external network instead of relying on native VLAN tagging. Implementing tagging at the VM layer and not at the OVS level does have one added benefit, however. In a manner similar to FAS systems, ONTAP Select allows the use of multiple IPspaces and VLAN tagging in its support for multitenancy implementations. To make this functionality available to an ONTAP Select administrator, VLAN tagging should be performed at the ONTAP Select VM level.

Implementing tagging within a VM is known as guest VLAN tagging. Using guest VLAN tagging with ONTAP Select, rather than implementing VLAN tagging with physical switches (switch tagging), allows data, management, and replication traffic to be further split across multiple layer 2 networks.

Two-port NIC configurations require the use of 10Gbps NICs. Running ONTAP Select on a system with only two 1Gbps NIC ports is only supported for single-node Select clusters.

Figure 29 shows a two-port NIC configuration.

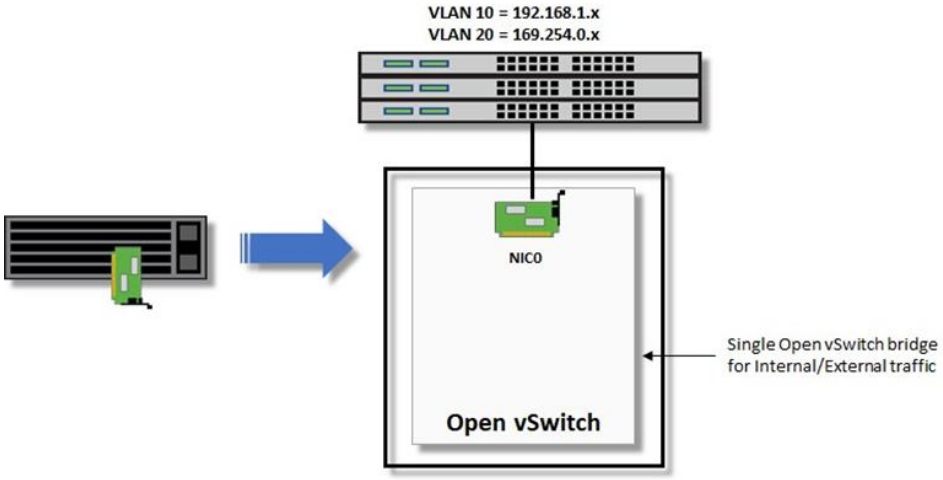
Figure 29) Two-port NIC network configuration.



A single-port NIC configuration requires the use of a 10Gbps NIC. In this configuration, a bond is not required, and an LACP channel cannot be configured for redundancy or improved bandwidth. A bridge should still be used and contains the virtual (or vnet) interfaces for the VMs.

Figure 30 shows a single-port NIC configuration.

Figure 30) Single-port 10GbE NIC network configuration.



Best Practice

The single 10Gbps NIC port configuration is supported for both single-node and two-node configurations. The recommended practice is to use at least two 10Gb ports or two 1Gb and two 10Gb NIC ports for redundancy and improved bandwidth.

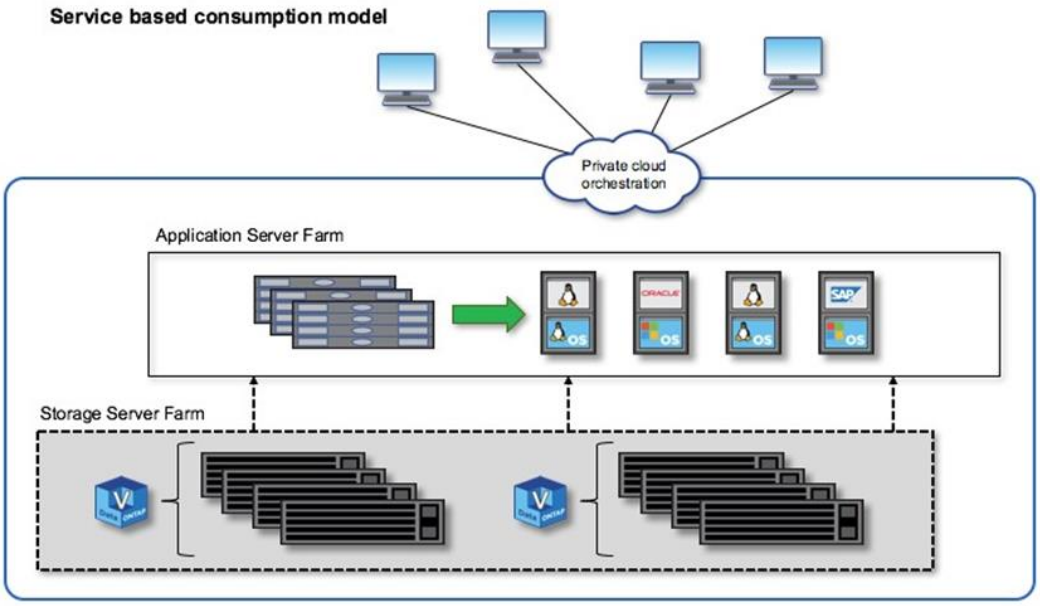
5 Use Cases

5.1 Private Cloud (Data Center)

A common use case for ONTAP Select is providing storage services for private clouds built on commodity servers. In Figure 31, 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 storage virtual machines (SVMs) to the deployment and configuration of application VMs, is automated through a private cloud orchestration framework.

This is the service-oriented private cloud model, and using the HA version of ONTAP Select creates the same ONTAP experience one 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.

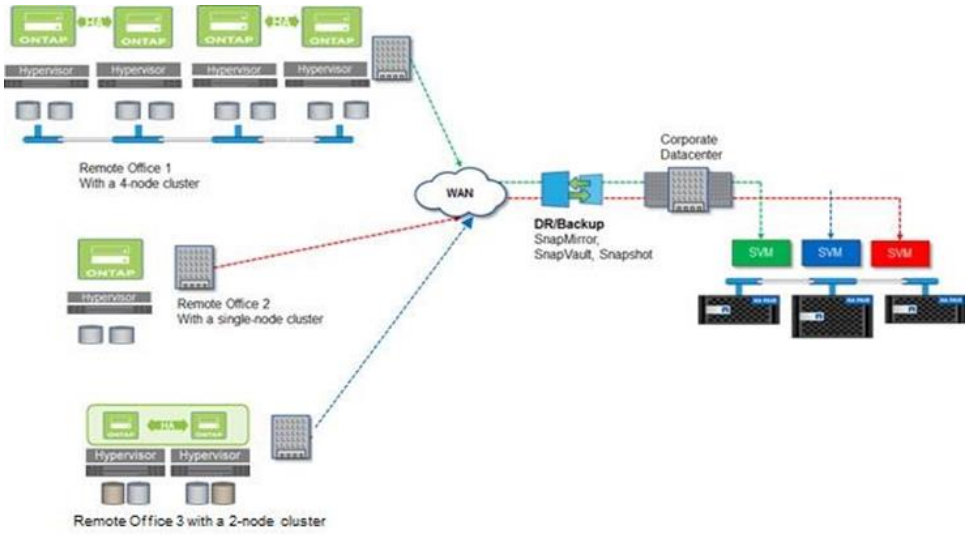
Figure 31) Private cloud built on direct-attached storage.



5.2 Data Replication and Backup

Other use cases for ONTAP Select are data replication and backup. 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).

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



5.3 Extreme Edge

ONTAP Select is suited for extreme edge environments. They can be deployed in ruggedized, small-form-factor hardware that can withstand extreme conditions of shock and temperature with limited space or connectivity.

ONTAP Select enables data collection and aggregation at the data center with dedicated ONTAP FAS-hosted solutions, long-term data archiving at the extreme edge with ONTAP Select NetApp E-Series, and cloud computing solutions with NetApp SolidFire with cloud tiers or NetApp Cloud Volumes ONTAP (Figure 33).

Figure 33) ONTAP Select can be deployed on small form-factor devices.

ONTAP Select provides the following benefits:

- Simple and cost-effective data management at disconnected and constrained edge environments
- Data security features such as volume encryption and SnapLock that can help prevent data from being compromised

- Edge processing with enterprise data protection
- Integration into the NetApp Data Fabric, which can enable valuable insights into data

ONTAP Select completes the NetApp Data Fabric with on-site data acquisition and ONTAP features of data processing, data protection, replication, security, and storage from rugged environments to on-premises and cloud environments.

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 12 describes the support matrix for ONTAP Select and ONTAP Deploy.

Table 12) ONTAP Deploy versus ONTAP Select support matrix.

	Select 9.0	Select 9.1	Select 9.2	Select 9.3	Select 9.4	Select 9.5
Deploy 2.2.2	Supported	Supported	Not supported	Not supported	Not supported	Not supported
Deploy 2.3	Supported	Supported	Not supported	Not supported	Not supported	Not supported
Deploy 2.7 (limited support)	Not supported	Supported	Supported	Supported	Supported	Not supported
Deploy 2.8	Not supported	Supported	Supported	Supported	Supported	Not supported
Deploy 2.9	Not supported	Supported	Supported	Supported	Supported	Not supported
Deploy 2.10	Not supported	Supported	Supported	Supported	Supported	Supported

ONTAP Deploy can only manage the Select clusters that it has deployed. One instance of ONTAP Deploy cannot currently discover ONTAP Select clusters installed by another Deploy instance. 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.

Best Practice

NetApp recommends backing up the Deploy database on a regular schedule, every time a configuration change is made, and before any upgrade.

7 Configuring Storage

Appendix B provides a quick overview of some of the commands used on the host to configure storage using the Linux LVM to prepare storage for use on the VMs.

7.1 Creating Volumes and Storage Pools

The different layers of storage are as follows:

- LVM logical volumes
- LVM volume groups (storage pools)
- Physical volumes (LUNs)
- Hardware RAID controller (RAID 5/6)
- Hard drives

Storage at the back end is aggregated and exposed as a block device. By using a tool such as `lsblk` that can list all available block devices, the target device on which the storage pool should be created can be determined. In the following examples, the hardware RAID controller exposes a single LUN called `/dev/sdc` from the direct-attached storage (DAS) to the host. This LUN can then be used to configure storage pools to use within ONTAP Select.

At the lowest layer are the device files (for example, `/dev/sda`, `/dev/sdb`, `/dev/sdc`, and so on) seen by the underlying host. One of the device files (typically with the largest capacity) represents the direct-attached storage.

The `virsh` command enables creation of a pool dedicated for the use of the two running ONTAP Select instances. The storage pool (`ontap_select` in the following example) is created by specifying the source as an available block device representing the back-end DAS storage.

The `virsh pool-define-as` command can be used to define a new LVM-based storage pool. In the following example, a new pool `ontap_select` is defined and uses the device `/dev/sdc`. The pool is built using the `virsh pool-build` command and initialized using `virsh pool-start` command.

```
virsh pool-define-as ontap_select logical --source-dev /dev/sdc --target=/dev/ontap_select virsh
pool-build ontap_select virsh pool-start ontap_select
```

This pool name can be specified when you configure a new host, as shown in the section “Deployment and Management.”

Using the Linux command `lsblk`, you can list all available block devices including LVMs. The command `vgscan` can be used to display volume groups and the command `lvscan` can be used to display logical volumes.

LSBLK on the host for a single-node cluster

```
NAME                                          MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT sda
8:0    0    60G  0 disk
+-ontap_select-volume_deploy                  8:1    0    60G 0 lvm  /mnt/deploy sdb
8:16   0    20G  0 disk +-sdb1                                        8:17   0   500M  0 part
/boot
+-sdb2                                        8:18   0  19.5G  0 part
  +-rhel-root                                253:0    0  17.5G  0 lvm  /   +-rhel-swap
253:1    0     2G  0 lvm  [SWAP] sdc                                          8:32   0     3T  0
disk +-ontap_select-onenode--sdot--8_DataONTAPv.raw 253:2    0  10.8G  0 lvm
+-ontap_select-onenode--sdot--8_coredisk       253:3    0   120G  0 lvm
+-ontap_select-onenode--sdot--8_sdotconfig.iso 253:4    0     4M  0 lvm
+-ontap_select-onenode--sdot--8_root_1         253:5    0    68G  0 lvm +-ontap_select-onenode--
sdot--8_ontap_select_1 253:6    0   1.6T  0 lvm
```

LSBLK on the host for two nodes of a four-node cluster

```
sdc                                                    8:32   0  3.3T  0 disk ├─ontap_select-
fournode--node1_DataONTAPv.raw           253:2    0 10.8G  0 lvm
├─ontap_select-fournode--node1_coredisk                253:3    0  120G  0 lvm
├─ontap_select-fournode--node1_DataONTAPv_mediator.raw  253:4    0  556M  0 lvm
├─ontap_select-fournode--node1_sdotconfig.iso           253:5    0    4M  0 lvm
├─ontap_select-fournode--node1_root_1                   253:6    0   68G  0 lvm
├─ontap_select-fournode--node1_root_2                   253:7    0   68G  0 lvm
├─ontap_select-fournode--node1_ontap_select_1           253:8    0  822G  0 lvm
└─ontap_select-fournode--node1_ontap_select_2           253:9    0  822G  0 lvm [root@sdot-node-1
~]#
```

```
sdc                                                     8:32   0     3T  0 disk ├─ontap_select-
fournode--node4_DataONTAPv.raw          253:3    0  10.8G  0 lvm
```

```
├─ontap_select-fournode--node4_coredisk                253:4    0   120G  0 lvm
├─ontap_select-fournode--node4_DataONTAPv_mediator.raw 253:5    0   556M  0 lvm
├─ontap_select-fournode--node4_sdotconfig.iso          253:6    0     4M  0 lvm
├─ontap_select-fournode--node4_root_1                  253:7    0    68G  0 lvm
├─ontap_select-fournode--node4_root_2                  253:8    0    68G  0 lvm
├─ontap_select-fournode--node4_ontap_select_1          253:9    0   431G  0 lvm
└─ontap_select-fournode--node4_ontap_select_2          253:10   0   431G  0 lvm [root@sdot-node-4
~]#
```

Note: The command `virsh vol-list --details ontap_select` can also be used to display the details for a storage pool, which in the preceding instance is `ontap_select`. The storage pool `ontap_select` was created by the `virsh pool-define-as` command.

In the case of a multinode cluster, the storage pools for HA pair nodes 1 and 2 look identical, and so do the storage pools for HA pair nodes 3 and 4.

Deploy creates volume groups from the storage by default with assigned partitions for the boot, core, mediator, configuration, root aggregate, and data. These partitions (LVMs or logical volumes) are typically of a fixed sized and dedicated for ONTAP use.

The following details are about each of the partitions:

- `DataONTAPv.raw` is the boot disk. The NVRAM is part of the boot disk and occupies 512MB of space within the partition containing `DataONTAPv.raw`.
- The `coredisk` partition is used to save core files.
- `sdotconfig.iso` has the saved configuration properties of the Select node.
- `root_1` is for this node's root aggregate, and `root_2` is the corresponding HA partner node's root aggregate.
- `ontap_select_1` and `ontap_select_2` are data disks that are available as spares and can be provisioned later to increase the capacity of existing aggregates or to create new aggregates.
- The `DataONTAPv_mediator.raw` partition is for the mediator, and it is used to broker and resolve split-brain scenarios. In a multinode cluster, the mediator on one of nodes of a two-node HA pair brokers the other two HA nodes. For example, the mediator on node 1 and 2 acts as the broker for HA pairs 3 and 4, and vice versa.

Each host in the HA pair sees the same volume groups.

```
[root@sdot-node-4 ~]# virsh pool-list --details
 Name          State    Autostart  Persistent   Capacity  Allocation  Available
-------------------------------------------------------------------------------  deploy
running  yes        yes         39.25 GiB   16.51 GiB  22.74 GiB  ontap_select  running  no
yes          3.00 TiB    1.14 TiB   1.85 TiB
[root@sdot-node-4 ~]# virsh vol-list ontap_select --details
Name                          Path                                                Type Capacity Allocation
-----------------------------------------------------------------------------------------------------------
fournode-node4_coredisk       /dev/ontap_select/fournode-node4_coredisk       block 120.00 GiB
120.00 GiB
fournode-node4_DataONTAPv.raw /dev/ontap_select/fournode-node4_DataONTAPv.raw block 10.74 GiB
10.74 GiB
fournode-node4_DataONTAPv_mediator.raw  /dev/ontap_select/fournode-node4_DataONTAPv_mediator.raw
block  556.00 MiB  556.00 MiB
fournode-node4_ontap_select_1 /dev/ontap_select/fournode-node4_ontap_select_1 block 431.00 GiB
431.00 GiB
fournode-node4_ontap_select_2 /dev/ontap_select/fournode-node4_ontap_select_2 block 431.00 GiB
431.00 GiB
fournode-node4_root_1         /dev/ontap_select/fournode-node4_root_1         block  68.00 GiB
```

```
68.00 GiB
fournode-node4_root_2         /dev/ontap_select/fournode-node4_root_2         block  68.00 GiB
68.00 GiB
fournode-node4_sdotconfig.iso /dev/ontap_select/fournode-node4_sdotconfig.iso block  4.00 MiB
4.00 MiB
```

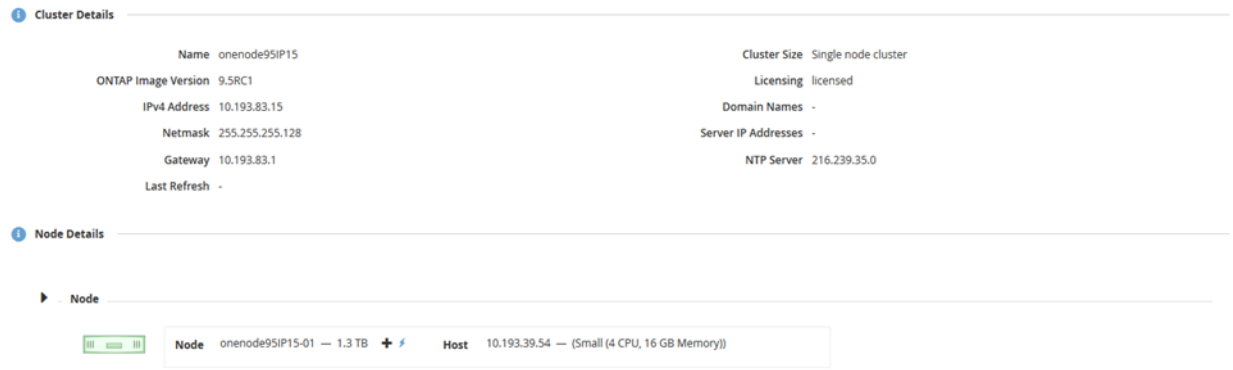
To summarize this section, LVMs are either boot, core, root volumes, configuration, or data. Data volumes are seen as VMDISKs from within ONTAP Select. A new storage pool created from a storage-add command creates more LVMs that show up as additional VMDISKs on ONTAP.

7.2 Increasing Capacity for Configurations 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. **Error! Reference source not found.** shows the "+" icon.

Figure 34) Storage-add functionality in Deploy VM.



The following considerations are important for the success of a capacity-expansion operation. Adding capacity requires that an existing license covers existing plus newly created space. A storage-add operation that results in the node exceeding its licensed capacity fails. You must first install a new license with sufficient capacity.

The Deploy tool supports creating single-node Select clusters using local storage (DAS) for its storage pool. If the extra capacity is to be added to the existing ONTAP Select aggregate, then the new storage pool should have a performance profile similar to that of the existing storage pool.

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

You can add the new space to same volume group as an extent, as long as the total size of the volume group does not exceed the KVM-supported maximum volume group size. You can dynamically add an extent to the volume group where ONTAP Select is already installed. Doing so does not affect the operations of the ONTAP Select node.

If the ONTAP Select node is part of an HA pair, some additional considerations are in order. Increasing capacity in an HA pair requires adding local storage to both nodes in the pair. 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) to make sure that all the data from node 1 is replicated to node 2. That is, 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 to make sure that the node 1 data is fully protected during an HA event.

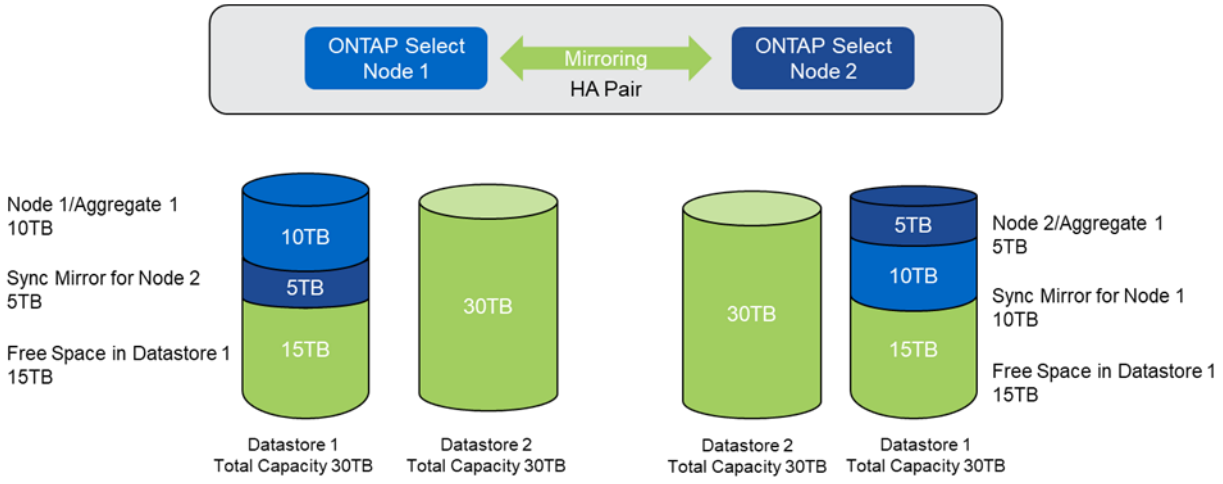
There is an additional consideration concerning performance. The data on node 1 is synchronously replicated to node 2. Therefore, the performance of the new space on node 1 must match the performance of the new space on node 2. In other words, adding space on both nodes but using different drive technologies or different RAID group sizes can lead to performance issues. These issues arise because RAID SyncMirror maintains 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, each node consuming 10TB of space from datastore 1. ONTAP Deploy configures each node with 5TB of active space per node.

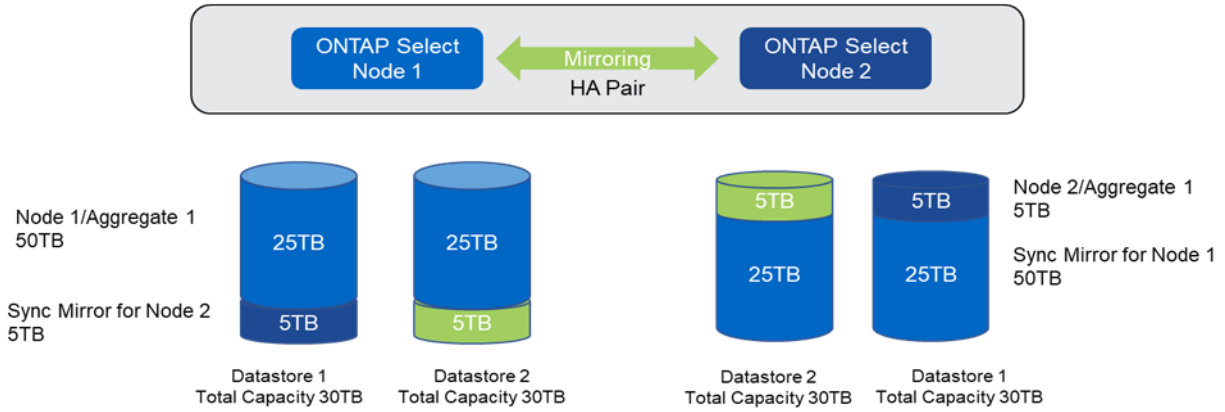
Figure 35 shows the results of a single storage-add operation for node 1. ONTAP Select still uses an equal amount of storage (15TB) on each node. However, node 1 has more active storage (10TB) than node 2 (5TB). Both nodes are fully protected because each node hosts a copy of the other node's data. There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

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



Best Practice

NetApp recommends that, whenever capacity must be expanded by adding new disks on the hardware RAID controller, you should create a new LUN. You should then create a new storage pool using this LUN (block device) and perform a storage-add operation using the new pool.

Best Practice

NetApp recommends adding a new aggregate instead of expanding an existing aggregate. If you must expand an existing aggregate, take care to add exactly the same type of storage. For example, you can create a new RAID group with the same number of disks with capacity and performance characteristics similar to existing drives. These disks are the allocated to each of the four nodes of the cluster. This new RAID group can be used to create a new aggregate or added to an existing aggregate.

Best Practice

Thin provisioning of LVMs on the host is not recommended. However, it is the administrator's responsibility to satisfy the actual available storage constraints on the ONTAP Select nodes if the storage system is thin-provisioned. In addition, Deploy would not recognize the thin-provisioned storage.

7.3 Creating New Aggregates on ONTAP Select VM

NetApp recommends letting the Deploy utility perform disk assignments automatically through the `storage add` command. This section only covers the process of manually creating a new aggregate.

After the storage is added to the host, a new RAID group is created, and the new LUN is imported as an extent. Virtual disks must be created and attached to the ONTAP Select VM. This process must be done using LVM commands on the host.

Note: Nodes in an HA pair must have the same total capacity. Increasing capacity for node 1 by 32TB implies a similar and simultaneous capacity expansion on its HA partner (node 2).

Within each ONTAP Select node, the newly assigned storage should be split into a number of equal-sized virtual disks, with no virtual disk exceeding 8TB.

For example, if 32TB of storage is added to the ONTAP Select cluster node, configure four 8TB virtual disks. If 7TB of storage is added to the ONTAP Select node, configure one 7TB virtual disk.

After the virtual disks have been provisioned, use the following mirrored aggregate creation workflow to assign and configure newly attached storage:

1. Assign the disks to the proper cluster node and plex if they haven't already been assigned. To do so, From the ONTAP CLI, run the `disk show` command. In this example, we're using a newly installed ONTAP Select cluster with two 822GB data disks per node.

```
fournode::> disk show
                    Usable           Disk    Container   Container
Disk                  Size Shelf Bay Type    Type        Name                      Owner
---------------- ---------- ----- --- ------- ----------- ---------                 --------
NET-1.1             822.0GB     -   - VMDISK  spare       Pool0                     fournode-node1
NET-1.2             66.93GB     -   - VMDISK  aggregate   aggr0_fournode_node2_0    fournode-node2
NET-1.3             66.93GB     -   - VMDISK  aggregate   aggr0                     fournode-node1
NET-1.4             822.0GB     -   - VMDISK  spare       Pool1                     fournode-node2
NET-2.1             66.93GB     -   - VMDISK  aggregate   aggr0_fournode_node2_0    fournode-node2
NET-2.2             66.93GB     -   - VMDISK  aggregate   aggr0                     fournode-node1
NET-2.3             822.0GB     -   - VMDISK  spare       Pool0                     fournode-node2
NET-2.4             822.0GB     -   - VMDISK  spare       Pool1                     fournode-node1
NET-3.9             431.0GB     -   - VMDISK  spare       Pool0                     fournode-node3
NET-3.10            431.0GB     -   - VMDISK  spare       Pool1                     fournode-node4
NET-3.11            66.93GB     -   - VMDISK  aggregate   aggr0_fournode_node4_0    fournode-node4
NET-3.12            66.93GB     -   - VMDISK  aggregate   aggr0_fournode_node3_0    fournode-node3
NET-4.9             431.0GB     -   - VMDISK  spare       Pool0                     fournode-node4
NET-4.10            431.0GB     -   - VMDISK  spare       Pool1                     fournode-node3
NET-4.11            66.93GB     -   - VMDISK  aggregate   aggr0_fournode_node3_0    fournode-node3
NET-4.12            66.93GB     -   - VMDISK  aggregate   aggr0_fournode_node4_0    fournode-node4
```

   The owner field lists the ONTAP Select cluster node that has a physical connection to the backing storage disk. This is the owning node.

2. To create an aggregate on the node fournode-node1 if it is not already assigned, assign a local disk to storage pool 0 (another term for plex) and a mirror disk to storage pool 1. Remember that the mirror disk must be contributed by the HA partner, in this case node fournode-node2.

   Our aggregate uses the disks NET-1.1 and NET-2.4. Although both disks have been assigned to ONTAP Select node fournode-node1, NET-1.1 is physically connected to the ONTAP Select VM fournode-node1. NET-2.4 is physically connected to the ONTAP Select VM fournode-node2 (hence the pool 1 designation).

3. Now that disks have been assigned to the correct plex (pool), you can now create the aggregate by issuing the following command:

   Note: This step can also be performed using System Manager.

```
fournode::> aggregate create -aggregate data_aggr1 -diskcount 2 -mirror true -node fournode-node1
(storage aggregate create)

Info: The layout for aggregate "data_aggr1" on node "fournode-node1" would be:

      First Plex

        RAID Group rg0, 1 disks (advanced_zoned checksum, raid0)
          Position   Disk                      Type                  Size
          ---------- ------------------------- ---------- ----------------
          data       NET-1.1                   VMDISK             809.1GB
      Second Plex

        RAID Group rg0, 1 disks (advanced_zoned checksum, raid0)
          Position   Disk                      Type                  Size           ---------- --
---------------------- ---------- ---------------           data       NET-2.4
VMDISK             809.1GB
```

```
    Aggregate capacity available for volume use would be 691.8GB.
```

Note: From this point, SVM, volume, LIF, and protocol configuration can be performed with ONTAP System Manager (or the ONTAP CLI) using the same procedures you would use to configure these parameters on FAS.

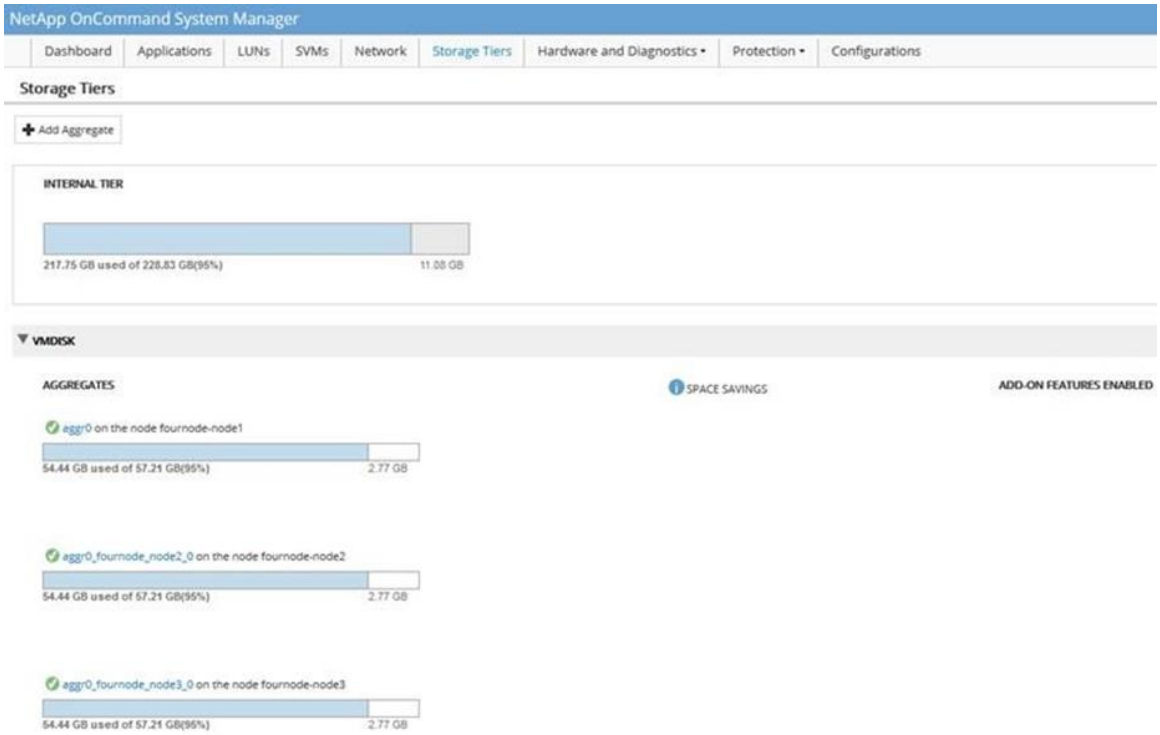
7.4 Example of New VMDISKs Using the ONTAP System Manager GUI

ONTAP System Manager can be invoked using the cluster IP address and the admin credentials created with the cluster-create operation on Deploy. Adding more capacity to an aggregate can be done by choosing one of the two ontap_select local data disks (VMDISKs) from the HA pair of the multinode cluster. This capacity can be added as part of the existing RAID group or as part of a new RAID group in an aggregate (aggregates consist of RAID groups).

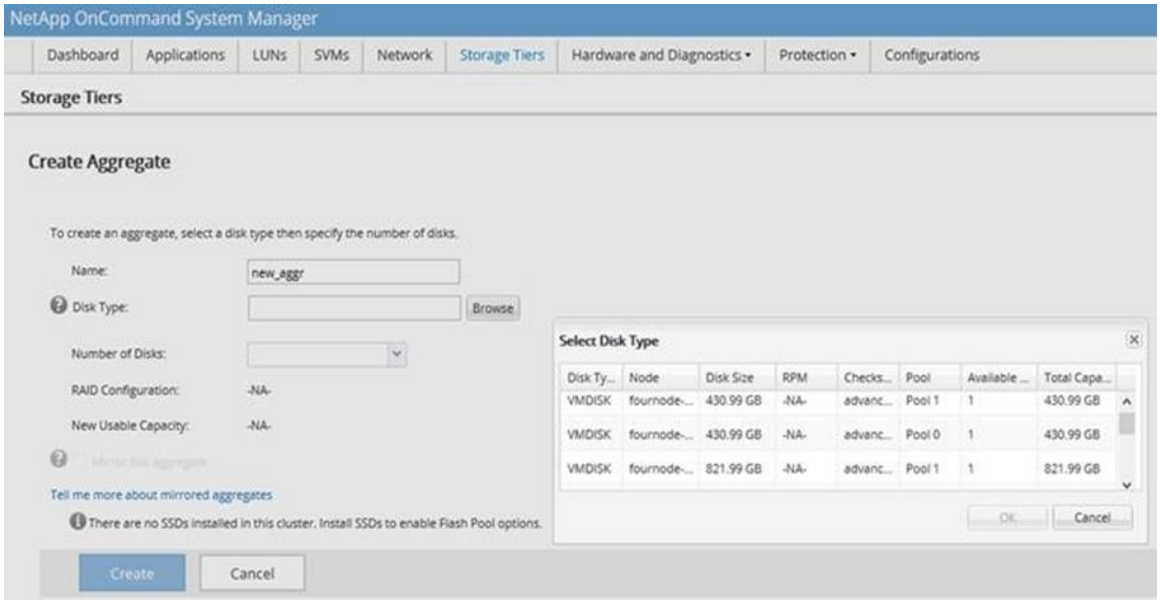
New aggregates can be created out of any of the ontap_select data disks (eight disks distributed across the four nodes of the cluster), which are seen as VMDISKs in the OnCommand System Manager.

Be sure to create similar RAID groups with the same number of disks, similar capacities, and performance to have consistent performance across the aggregate. To create new aggregates, complete the following steps:

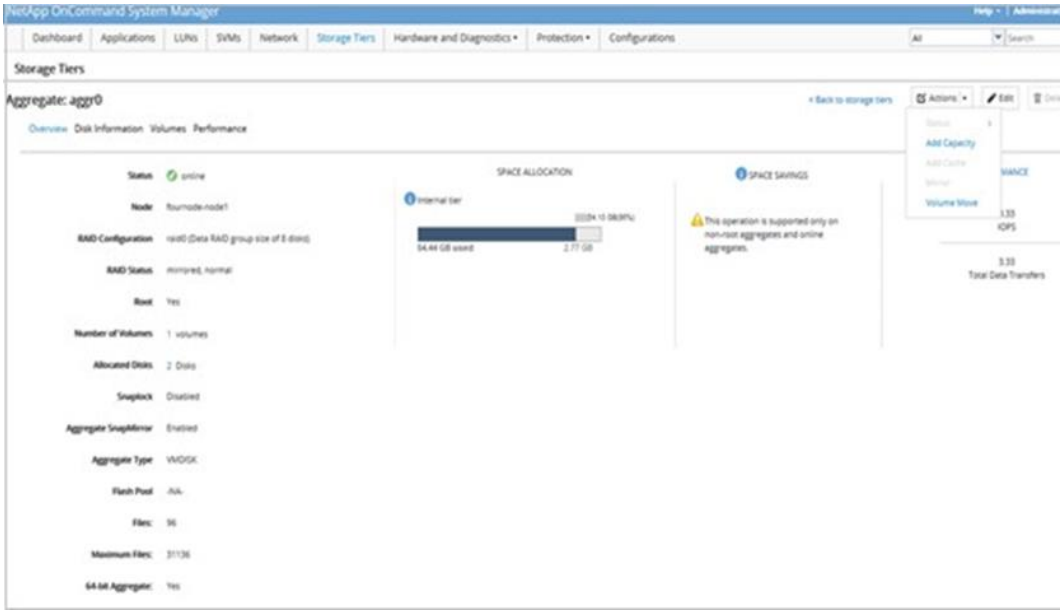
1. Display all the aggregates in the cluster.



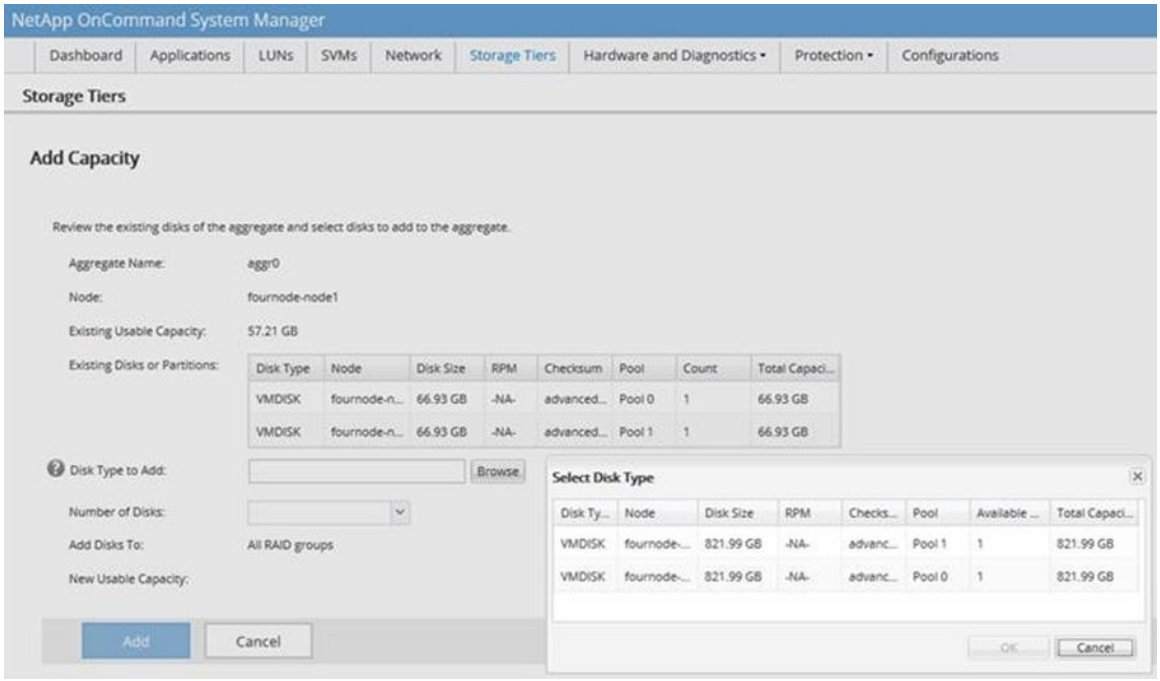
2. Create a new aggregate and display all the VMDISKs to use in the cluster.



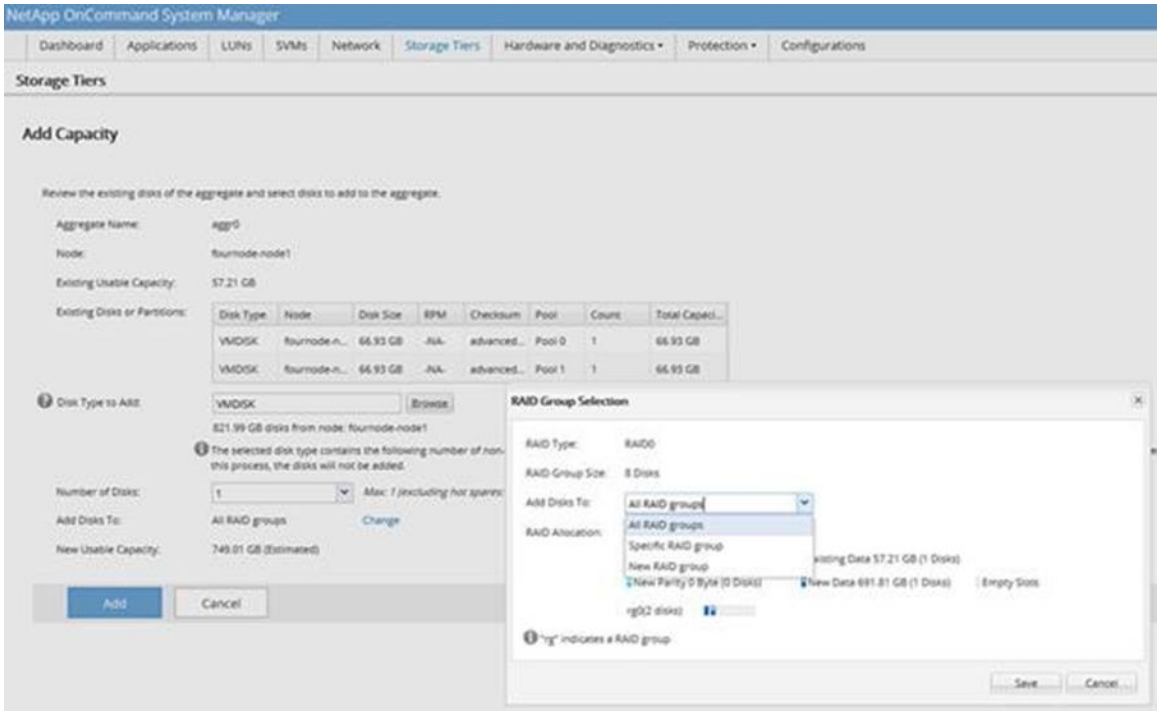
3. Add capacity to an existing aggregate from the top left corner drop-down list.



4. Add capacity to an aggregate by choosing VMDISKs.



5. Add capacity by adding to an existing RAID group or by creating a new RAID group.



7.5 Storage Configuration with Software RAID

With software RAID, the physical disks are directly exposed on the hosts by the host HBA device driver. The virtio-scsi passthrough interface allows the storage LUN to be directly connected to a VM instance. (See more in Appendix E: VirtIO Interface for Raw Device Mapping).

To enable SCSI passthrough and to use host disks as bare-metal SCSI devices inside the ONTAP Select VM, the `<disk>` element's `device` attribute must be set to `lun`. The following domain XML snippet (`virsh dumpxml <domain>`) shows the `device` attribute's values for the specific domain (guest). This snippet shows devices mapped into ONTAP Select VM as raw LUNs. Note the directsync option.

```
<disk type='block' device='lun'>
      <driver name='qemu' type='raw' cache='directsync'/>       <source dev='/dev/disk/by-id/ata-
SAMSUNG_MZ7WD480HAGM-00003_S16MNYAF334190'/>       <backingStore/>
      <target dev='sdc' bus='scsi'/>
      <alias name='scsi0-0-0-2'/>
      <address type='drive' controller='0' bus='0' target='0' unit='2'/>     </disk>
    <disk type='block' device='lun'>
      <driver name='qemu' type='raw' cache='directsync'/>       <source dev='/dev/disk/by-id/ata-
SAMSUNG_MZ7WD480HAGM-00003_S16MNYAD800879'/>       <backingStore/>
      <target dev='sdd' bus='scsi'/>
      <alias name='scsi0-0-0-3'/>
      <address type='drive' controller='0' bus='0' target='0' unit='3'/>     </disk>
    <disk type='block' device='lun'>
      <driver name='qemu' type='raw' cache='directsync'/>       <source dev='/dev/disk/by-id/ata-
SAMSUNG_MZ7WD480HAGM-00003_S16MNYAD809905'/>       <backingStore/>
      <target dev='sde' bus='scsi'/>
      <alias name='scsi0-0-0-4'/>
      <address type='drive' controller='0' bus='0' target='0' unit='4'/>     </disk>
    <disk type='block' device='lun'>
      <driver name='qemu' type='raw' cache='directsync'/>       <source dev='/dev/disk/by-id/ata-
SAMSUNG_MZ7WD480HAGM-00003_S16MNYAF334128'/>       <backingStore/>
      <target dev='sdf' bus='scsi'/>
      <alias name='scsi0-0-0-5'/>
      <address type='drive' controller='0' bus='0' target='0' unit='5'/>     </disk>
    <disk type='block' device='lun'>
      <driver name='qemu' type='raw' cache='directsync'/>       <source dev='/dev/disk/by-id/ata-
SAMSUNG_MZ7WD480HAGM-00003_S16MNYAD710371'/>       <backingStore/>
      <target dev='sdg' bus='scsi'/>
      <alias name='scsi0-0-0-6'/>
      <address type='drive' controller='0' bus='0' target='0' unit='6'/>     </disk>
```

On the host, the raw devices can be seen under the `/dev/disk/by-id/` directory.

```
[root@sti-c6220-0046 ]# ls -al /dev/disk/by-id/ total 0 ……
lrwxrwxrwx. 1 root root   9 May 17 09:43 ata-SAMSUNG_MZ7WD480HAGM-00003_S16MNYAD710371 ->
../../sdf
lrwxrwxrwx. 1 root root   9 May 17 09:43 ata-SAMSUNG_MZ7WD480HAGM-00003_S16MNYAD800879 ->
../../sdd
lrwxrwxrwx. 1 root root   9 May 17 09:43 ata-SAMSUNG_MZ7WD480HAGM-00003_S16MNYAD809905 ->
../../sdc
lrwxrwxrwx. 1 root root   9 May 17 09:43 ata-SAMSUNG_MZ7WD480HAGM-00003_S16MNYAD809991 ->
../../sde
lrwxrwxrwx. 1 root root   9 May 17 09:43 ata-SAMSUNG_MZ7WD480HAGM-00003_S16MNYAF334128 ->
../../sdb
lrwxrwxrwx. 1 root root   9 May 17 09:43 ata-SAMSUNG_MZ7WD480HAGM-00003_S16MNYAF334190 ->
../../sda
lrwxrwxrwx. 1 root root   9 May 17 09:43 wwn-0x50025385000686e8 -> ../../sdf lrwxrwxrwx. 1 root
root   9 May 17 09:43 wwn-0x5002538500069f0d -> ../../sdd lrwxrwxrwx. 1 root root   9 May 17
09:43 wwn-0x50025385000722b5 -> ../../sdc lrwxrwxrwx. 1 root root   9 May 17 09:43 wwn-
0x500253850007230b -> ../../sde lrwxrwxrwx. 1 root root   9 May 17 09:43 wwn-0x5002538500180f0f -
> ../../sdb lrwxrwxrwx. 1 root root   9 May 17 09:43 wwn-0x5002538500180f4d -> ../../sda
```

The following `aggr show` and `disk show` commands show typical hardware RAID disk usage.

```
HWRAID_3738::> aggr show aggr_37a -disk
Aggregate #disks Disks
--------- ------ ---------------------------
aggr_37a       2 NET-1.2 HWRAID_3738::> disk show
                     Usable           Disk    Container   Container
Disk                   Size Shelf Bay Type    Type        Name      Owner
---------------- ---------- ----- --- ------- ----------- --------- --------
NET-1.1             66.93GB     -   - VMDISK  aggregate   aggr0_HWRAID_3738_01
                                                                    HWRAID_3738-01
NET-1.2             114.2GB     -   - VMDISK  aggregate   aggr_37a  HWRAID_3738-01
```

```
NET-1.3             66.93GB     -   - VMDISK  aggregate   aggr0_HWRAID_3738_02
                                                          HWRAID_3738-02
NET-1.4             114.2GB     -   - VMDISK  aggregate   aggr_38a  HWRAID_3738-02 NET-3.1
66.93GB     -   - VMDISK  aggregate   aggr0_HWRAID_3738_01
                                                          HWRAID_3738-01
NET-3.2             114.2GB     -   - VMDISK  aggregate   aggr_37a  HWRAID_3738-01
NET-3.3             114.2GB     -   - VMDISK  aggregate   aggr_38a  HWRAID_3738-02 NET-3.4
66.93GB     -   - VMDISK  aggregate   aggr0_HWRAID_3738_02
HWRAID_3738-02
8 entries were displayed.
```

In contrast, the following commands show a Software RAID aggregate and disk usage. Using ONTAP Deploy, storage in the form of individual physical disks can be allocated to each ONTAP Select node based on the disks present in the corresponding host.

```
(ONTAPdeploy) host storage disk show -host-name sti-c6220-0045.ctl.gdl.englab.netapp.com
+-----------------------------------+---------+-----------+---------------------------+----------
---+
| Name                              | Adapter | Capacity  | Used By                   | Device
Type |
+-----------------------------------+---------+-----------+---------------------------+----------
---+
| /dev/disk/by-id/ata-              | scsi0   | 447.13 GB | qemu-kvm                  | SSD
|
| SAMSUNG_MZ7WD480HAGM-             |         |           |                           |
|
| 00003_S16MNYAD809934              |         |           |                           |
|
| /dev/disk/by-id/ata-              | scsi0   | 447.13 GB | qemu-kvm                  | SSD
|
| SAMSUNG_MZ7WD480HAGM-             |         |           |                           |
|
| 00003_S16MNYAF334130              |         |           |                           |
|
| /dev/disk/by-id/ata-              | scsi0   | 447.13 GB | qemu-kvm                  | SSD
|
| SAMSUNG_MZ7WD480HAGM-             |         |           |                           |
|
| 00003_S16MNYAD840985              |         |           |                           |
|
| /dev/disk/by-id/ata-              | scsi0   | 447.13 GB | qemu-kvm                  | SSD
|
| SAMSUNG_MZ7LM480HCHP-             |         |           |                           |
|
| 000G3_S2GRNYAG600315              |         |           |                           |
|
| /dev/disk/by-id/ata-              | scsi0   | 447.13 GB | qemu-kvm                  | SSD
|
| SAMSUNG_MZ7WD480HAGM-             |         |           |                           |
|
| 00003_S16MNYAF804324              |         |           |                           |
|
| /dev/disk/by-id/ata-              | scsi0   | 447.13 GB | ontap-select-storage-pool | SSD
|
| SAMSUNG_MZ7WD480HAGM-             |         |           |                           |
|
| 00003_S16MNYAF605589              |         |           |                           |
|
| /dev/sdg                          | scsi7   | 30 GB     | multipath                 | Non-SSD
|
+-----------------------------------+---------+-----------+---------------------------+----------
---+
(ONTAPdeploy) node storage disk show -cluster-name KVM_SWRAID_N45N46 -node-name sti-c6220-0045a
+---------------------------------------+-----------+------+------------+
| Name                                  | Capacity  | Pool | Ontap Name |
+---------------------------------------+-----------+------+------------+
| /dev/disk/by-id/ata-                  | 447.13 GB | -    | NET-1.3    |
| SAMSUNG_MZ7WD480HAGM-                 |           |      |            |
| 00003_S16MNYAD809934                  |           |      |            |
| /dev/disk/by-id/ata-                  | 447.13 GB | -    | NET-1.4    |
```

```
| SAMSUNG_MZ7WD480HAGM-                    |           |      |            |
| 00003_S16MNYAF334130                     |           |      |            |
| /dev/disk/by-id/ata-                     | 447.13 GB | -    | NET-1.2    |
| SAMSUNG_MZ7WD480HAGM-                    |           |      |            |
| 00003_S16MNYAD840985                     |           |      |            |
| /dev/disk/by-id/ata-                     | 447.13 GB | -    | NET-1.1    |
| SAMSUNG_MZ7WD480HAGM-                    |           |      |            |
| 00003_S16MNYAF804324                     |           |      |            |
| /dev/disk/by-id/ata-                     | 447.13 GB | -    | NET-1.5    |
| SAMSUNG_MZ7LM480HCHP-                    |           |      |            |
| 000G3_S2GRNYAG600315                     |           |      |            |
+------------------------------------------+-----------+------+------------+
```

Within ONTAP Select, the root and data disks are seen as follows:

```
KVM_SWRAID_N45N46::> aggr show Aggr_45 -disk
Aggregate #disks Disks
--------- ------ --------------------------
Aggr_45        6 NET-1.3, NET-1.4, NET-1.1

KVM_SWRAID_N45N46::> disk show
                     Usable           Disk    Container   Container
Disk                   Size Shelf Bay Type    Type        Name      Owner
---------------- ---------- ----- --- ------- ----------- --------- --------

Info: This cluster has partitioned disks. To get a complete list of spare disk capacity use
"storage aggregate show-spare-disks".
NET-1.1             446.2GB     -   - SSD     shared      Aggr_45, Aggr_46, aggr0_sti_c6220_0046a
                                                          sti-c6220-0046a NET-1.2
446.2GB     -   - SSD     shared      Aggr_46, aggr0_sti_c6220_0046a
sti-c6220-0046a
NET-1.3             446.2GB     -   - SSD     shared      Aggr_45, Aggr_46, aggr0_sti_c6220_0045a
                                                          sti-c6220-0045a NET-1.4
446.2GB     -   - SSD     shared      Aggr_45, aggr0_sti_c6220_0045a
sti-c6220-0045a NET-1.5             446.2GB     -   - SSD     shared      -         sti-c6220-
0045a NET-3.1             446.2GB     -   - SSD     shared      Aggr_46, aggr0_sti_c6220_0046a
sti-c6220-0046a
NET-3.2             446.2GB     -   - SSD     shared      Aggr_45, Aggr_46, aggr0_sti_c6220_0045a
                                                          sti-c6220-0045a NET-3.3
446.2GB     -   - SSD     shared      Aggr_45, aggr0_sti_c6220_0045a
sti-c6220-0045a
NET-3.4             446.2GB     -   - SSD     shared      Aggr_45, Aggr_46, aggr0_sti_c6220_0046a
                                                          sti-c6220-0046a NET-3.5
446.2GB     -   - SSD     shared      -         sti-c6220-0046a 10 entries were displayed.
```

Compared to hardware RAID, more virtual drives are required with software RAID because the RAID groups are managed by ONTAP. With hardware RAID, a smaller set of drives can be used. The drives are already abstracted by the hardware RAID controller (providing a single RAID 5/6 option) and exposed as a single LUN.

In the hardware RAID configuration, Deploy automatically carves up the system disks, mainly for the internal mediator and mailbox disks, the boot, the vNVRAM, and the coredump disk. The /root contains all the system disks except the coredump disk and the boot disk. Data disks are created during the initial ONTAP Select configuration and more data disks can be created later using the "`storage add`" feature from within Deploy.

With hardware RAID, a virtual disk (VMDISK) seen within ONTAP Select is one of several logical volumes (LVs) carved from the single large storage pool. However, with software RAID, a storage pool is not used to carve out root disks and data disks.

A virtual disk (VMDISK) seen within the ONTAP Select VM in Software RAID would be one of the three pre-created partitions from a real physical disk. Software RAID involves real physical drives, and configuration involves requirements around the actual number and capacity of individual physical drives. Spare disks are required in the case where a physical disk failure occurs and the physical disk needs to be replaced.

Capacity Usage on the Storage Pool

The following `lvscan` command shows how the storage is laid-out for a Software RAID configuration.

This shows only the `DataONTAPv` (Boot, NVRAM), `coredisk` (Coredump), and `sdotconfig` (Config) LVs.

```
[root@sti-c6220-0046 dev]# lvscan
  ACTIVE            '/dev/ontap-select-storage-pool/sti-c6220-0046a_DataONTAPv.raw' [10.74 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/sti-c6220-0046a_coredisk' [120.00 GiB]
inherit
  ACTIVE            '/dev/ontap-select-storage-pool/sti-c6220-0046a_sdotconfig.iso' [4.00 MiB]
inherit ……
```

Note: About 131GB of raw disk capacity is used by the storage pool for the creation of the System Disks with Software RAID.

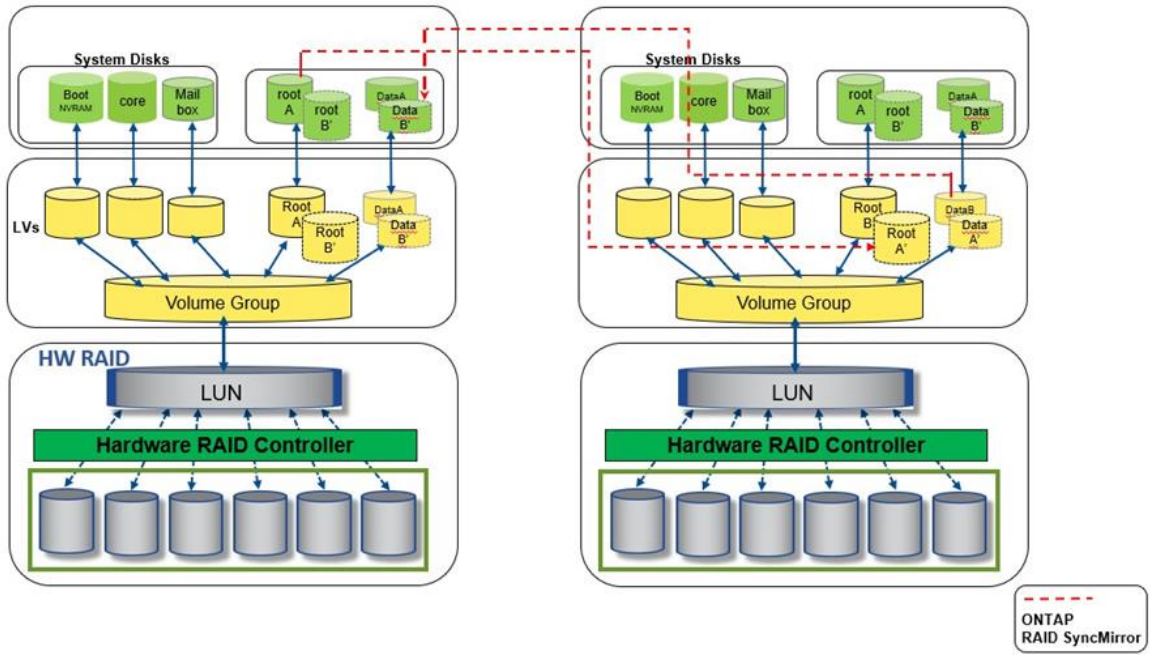
Note that the root disks and the data disks are not seen in this group of Logical Volumes (LVs). With Software RAID, the root and data disks are entirely managed by the physical raw disks allocated for the ONTAP Select nodes. Thus, the total Storage Pool size would be about 131GB.

With Hardware RAID, the root disks add an additional LV of 68GB for Single node and two LVs of 68GB per node (total of 156GB per node) for multinode configurations. Thus the total Storage pool size needed would be about 300GB.

In a multinode configuration, a small additional LV of about 556MB is created for the mediator disk. This LV is added in both the Hardware and Software RAID cases.

In a typical Hardware RAID configuration, the root and virtual data disks are Sync-mirrored across the HA pair as shown in Figure 37. The SyncMirror thus happens between the pairs of LVs for root and data disks.

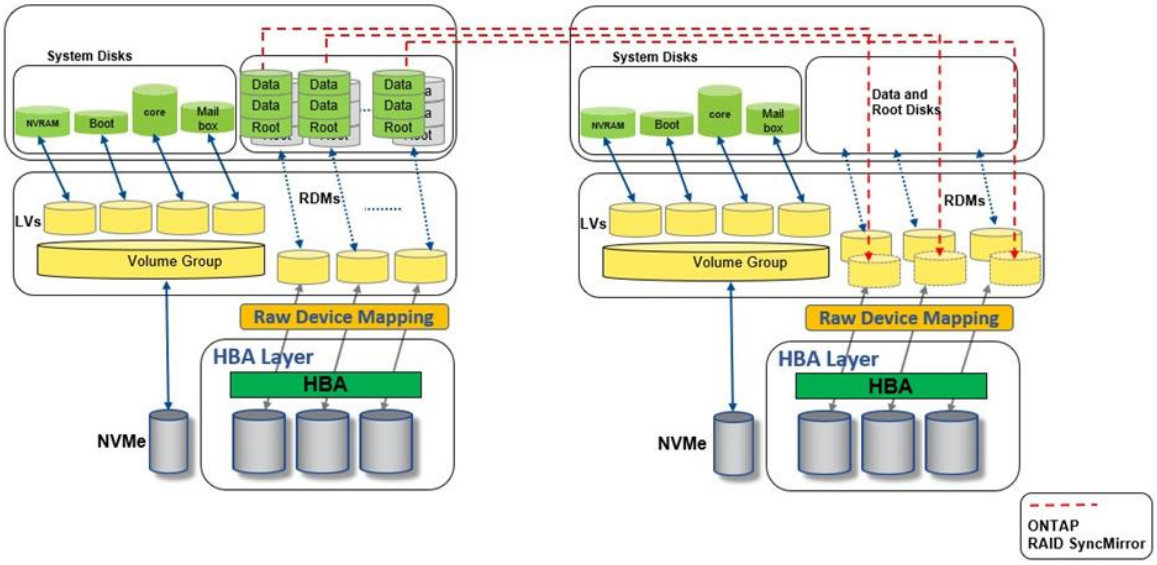
Figure 37) RAID SyncMirror flow in a Hardware RAID environment.



With Software RAID, an aggregate requires two sets of disks, one set for the local node and the other set for the partner node (Sync Mirrored shown shaded). For example, for the root aggregate, Figure 38 shows a local Root A disk and a Root B disk for the partner.

For a two-node setup, a third node called a mediator is needed to provide a mailbox disk as a tie breaker during takeover.

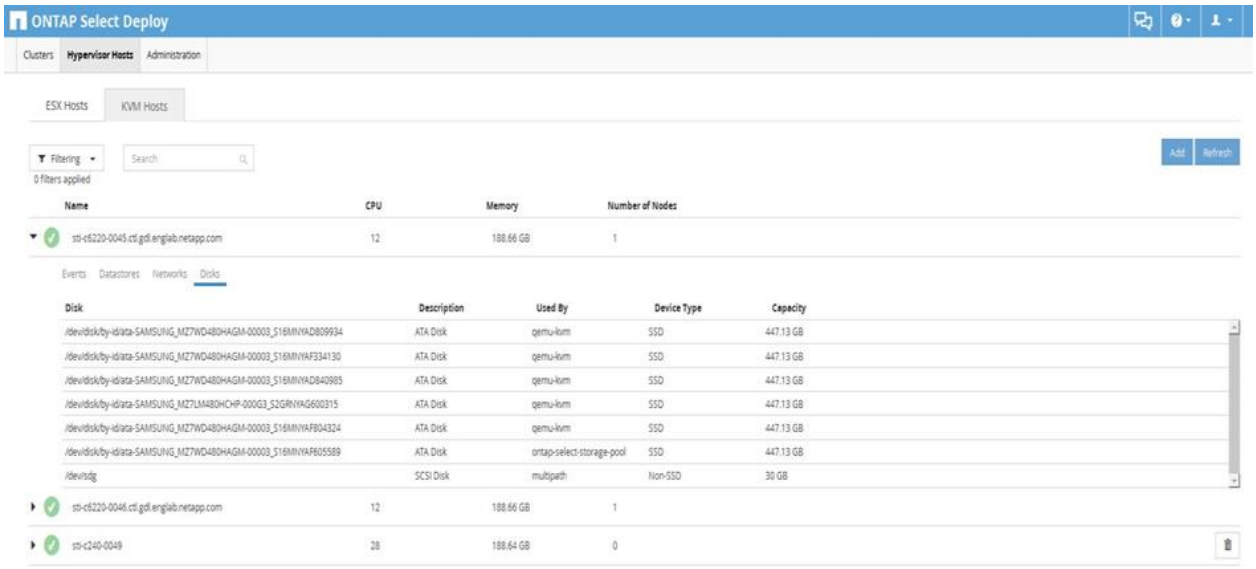
Figure 38) RAID SyncMirror Flow in a Software RAID environment.



Software RAID Disk Association GUI Steps

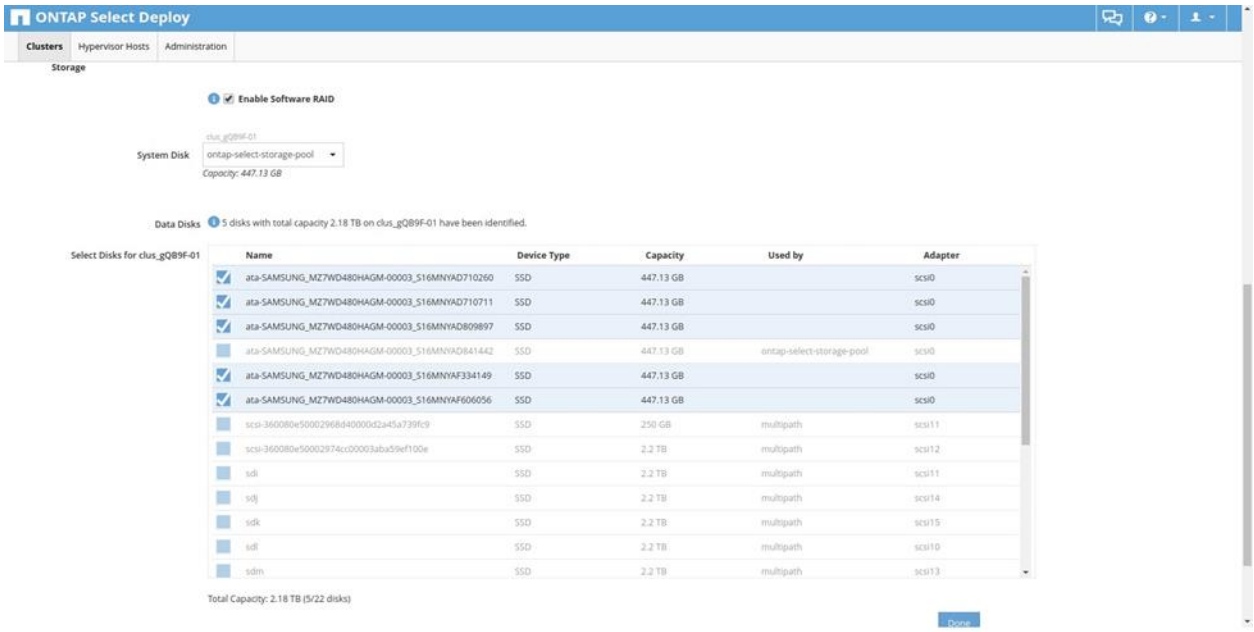
To associate raw disks to an ONTAP Select node, complete the following steps:

1. Show available disks on a host. A list of all available disks can be seen for a host as shown.

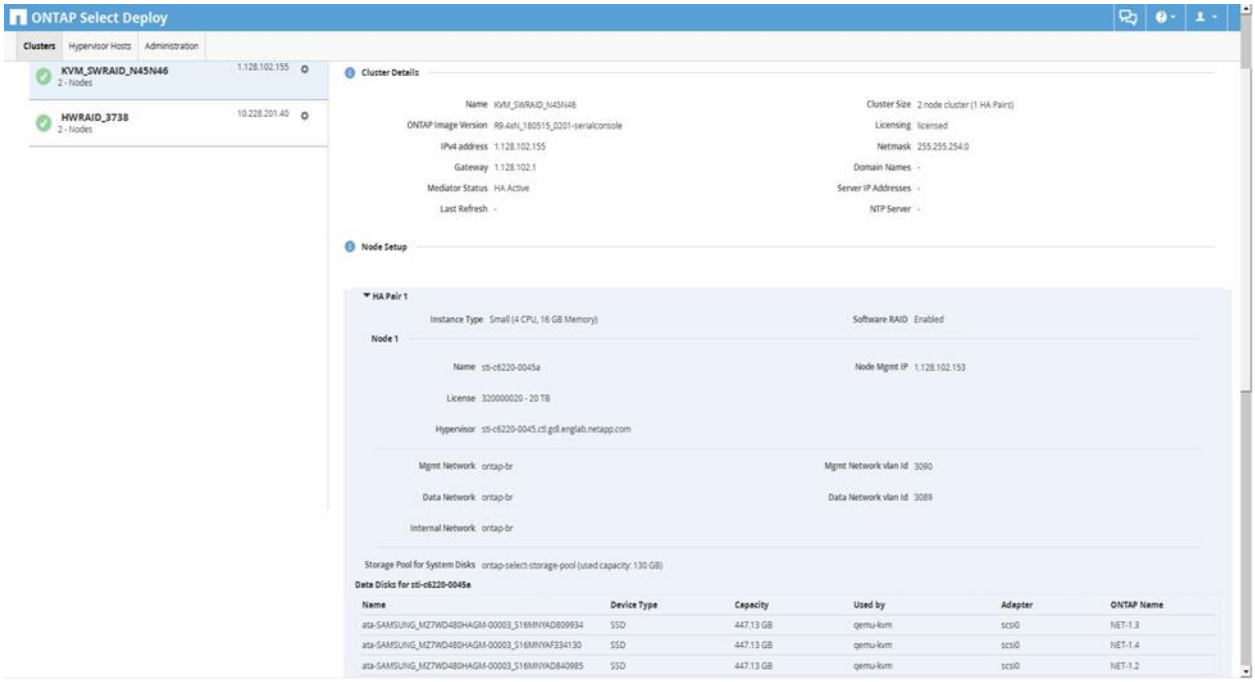


Note: You must check the Enable Software RAID option to choose software RAID.

2. Attach some or all available disks to the ONTAP Select node.



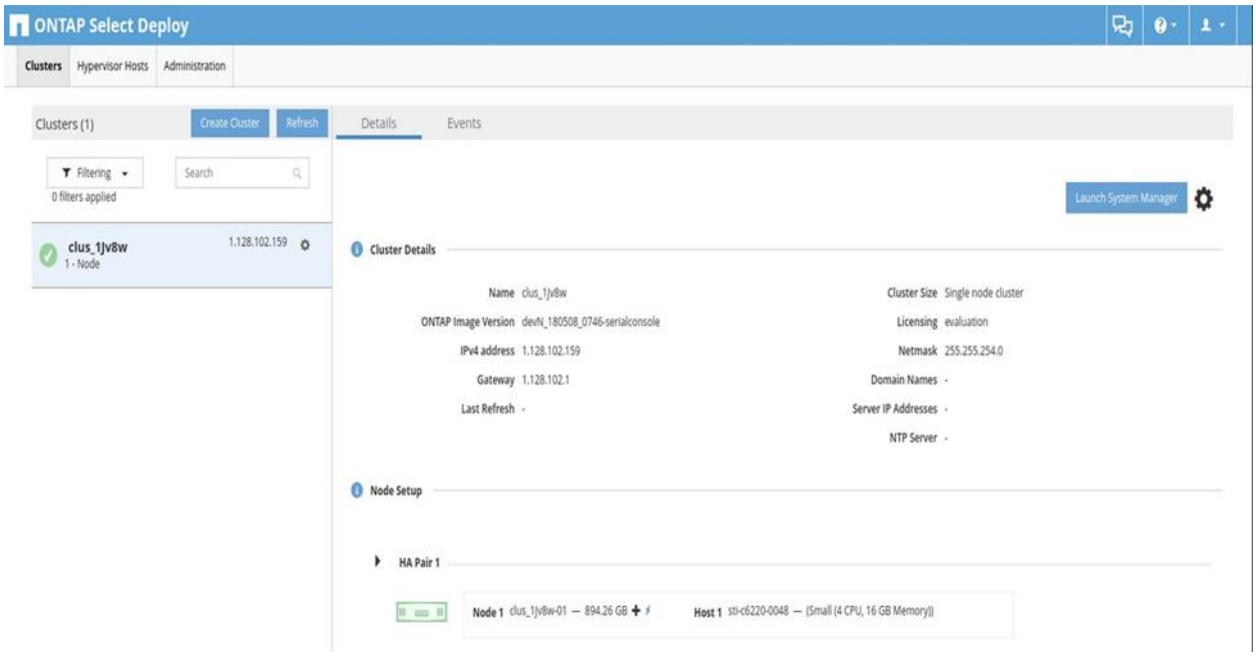
3. Show disks attached to the ONTAP Select node.



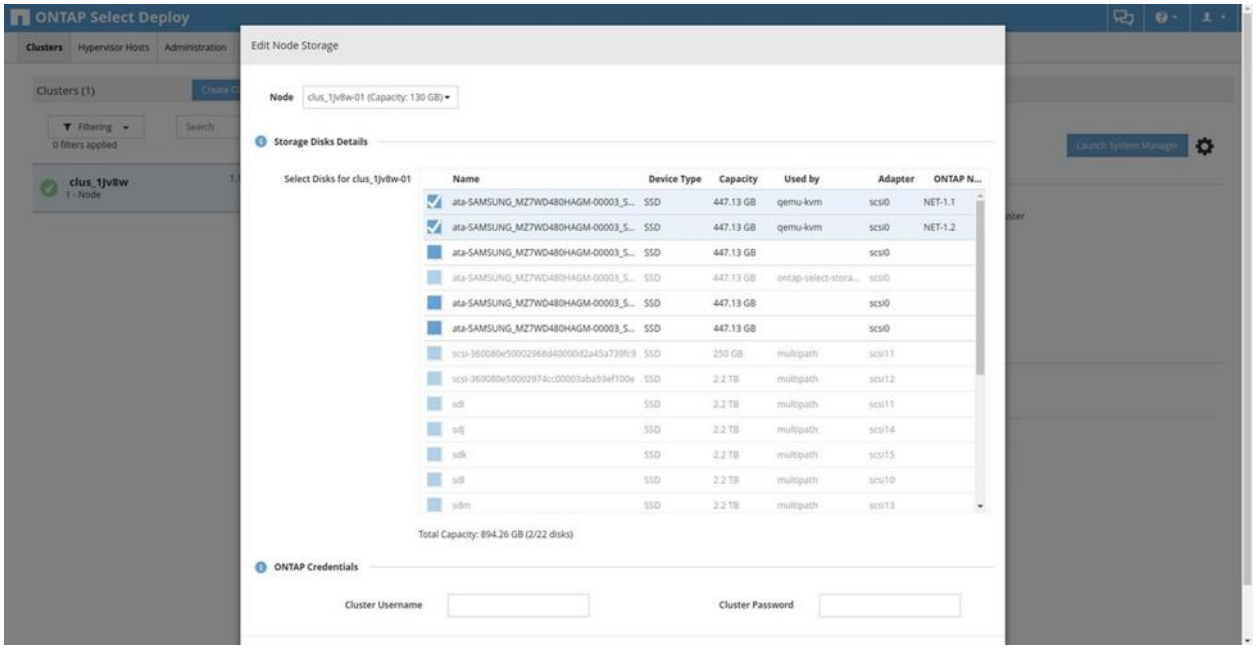
Software RAID Disk Addition Post-Deployment

More disks can be added post-cluster deployment into an ONTAP Select node. To select more disks, click the (+) symbol next to the node in the HA Pair 1 section as is described in the following steps:

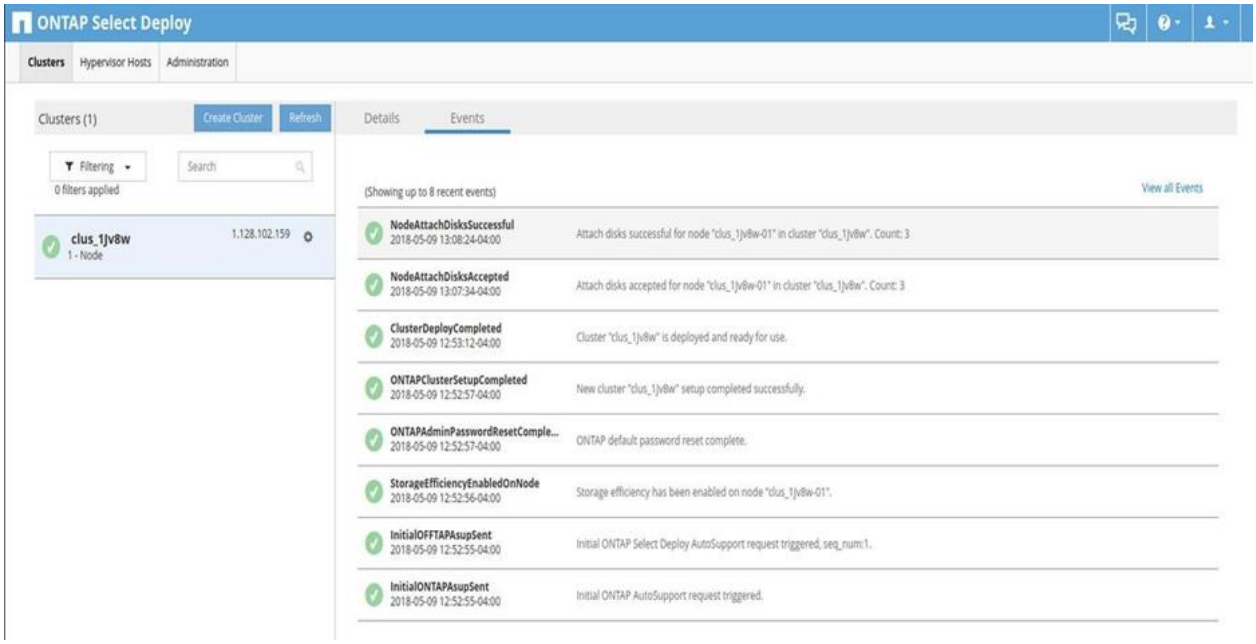
1. Choose the ONTAP Select node.



2. Edit storage for the ONTAP Select node.



3. Verify that the event logs show disks as being attached.

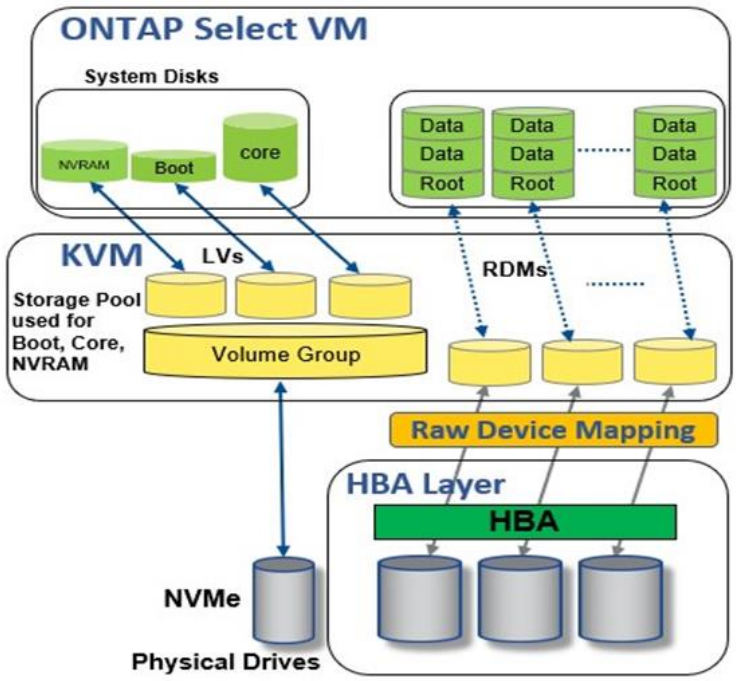


4. After this step, the ONTAP Select disk listing should show that the new set of disks belongs to the node. These disks can then be used to create new volumes.

Using an NVMe Drive as a Service Disk for High Performance

You can configure an NVMe drive as a service disk for use as a system disk. A separate storage pool or volume group is created. The boot, vNVRAM, and Coredump logical volumes are created for the ONTAP Select VM as shown in Figure 39.

Figure 39) NVMe disk as a service disk in a software RAID environment.



Replacement of a Failed Disk in a Software RAID Group

When ONTAP detects a disk failure, the assigned spare disk is automatically used to replace the failed disk. RAID reconstruction takes place that will rewrite the new spare disk and integrate it into the RAID group.

To replace the failed physical hard disk, the disk unique ID from ONTAP should be used to identify the back-end device. The device ID helps to identify the `/dev/disk/by-id/scsi-<unique-id>` that has unique HBA, target, and LUN numbers. Certain Linux tools such as ledctl can be used to perform LED control to physically change LEDs and visually identify the failed drive.

The new disk can then be rescanned and attached to an ONTAP node (RDM mapping), and, within ONTAP, the new disk can then be assigned to an ONTAP Select node. The new disk then becomes the spare disk.

Disks can also be identified by looking at the UUID field with the disk show command in ONTAP.

Adding Storage Capacity after Cluster Deployment

If new physical drives are added and then those disks are attached to an ONTAP Select node, each drive is divided into two to create equal-sized data partitions. These data partitions are of a different size than the data partitions on the pre-existing drives because they do not include the root partitions already created on the pre-existing drives.

Note: The first few physical drives added are always partitioned into three partitions: one small partition for the root stripe and two equal-sized partitions for the data drives.

Depending on the number of drives, existing data aggregates can be expanded or new aggregates created. This process follows generic ONTAP best practices for aggregates. When expanding existing aggregates, if the new data partitions are smaller, they must be added as a new RAID group to the aggregate. If the data partitions are the same size or larger, they can be added to pre-existing RAID groups that have not been maxed out.

In the absence of a separate system disk, forming a new aggregate requires a minimum of seven disks in RAID-DP (four disks in RAID 4). In an HA configuration, because all disks are mirrored, equivalent disks must be added to each node.

Handling Irregular Sized Disks

Although NetApp recommends that all disks be the same size, it is not required. If there are irregular disk sizes during deployment, all disks are still used for the root aggregate. The root partition size is the same across all disks, and the data partition size differs based on the physical disk size.

If the data partitions on the local node are of different sizes, data aggregate can be created from those partitions. However, the higher capacity partitions are downsized to the lowest capacity partition available in the RAID group. The handling of mixed-size partitions follows the same rules as for the handling of mixed-size whole disks in ONTAP and follows ONTAP best practices.

For example, if a four-data-disk plus two-parity-disk RAID group is created with three 400G partitions and three 800G partitions, the effective capacity of the 800G partitions is downsized to 400G. To mirror such a RAID group with different-sized partitions, all partitions from the target plex should be of at least the size of the lowest capacity disk in the source plex. Otherwise, aggregate-mirroring fails.

As another example, there are six 400G partitions and six 800G partitions per node. An aggregate with two RAID groups can be created in which the first RAID group contains five 400G partitions (three data and two parity). The second RAID group contains five 800G partitions (three data and two parity). One 400G partition and one 800G partition are then left as hot spares for the corresponding RAID groups.

7.6 Single-Node to Multinode Upgrade

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

8 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. Indeed, the hardware configuration of a system 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.

8.1 ONTAP Select 9.2 (Premium) Hardware

Workload and Hardware Details

- 64K sequential read
- 64K sequential write
- 4K and 8K random read
- 4K and 8K random write
- 4K and 8K random read/write 50/50
- I/O generator
- sio
- ONTAP Select Premium
- 8-core 64GB memory with SSDs
- Gen1
  - Server (Cisco UCS C240 M4S2)
  - CPU: Intel Xeon CPU E5-2697
  - 14 cores, 1 socket, 2.6 GHz
  - Memory: 128GB physical 2133MHz
- Gen2
  - Server (Cisco UCS C240 M4SX)
  - CPU: Intel Xeon CPU E5-2620
  - 16 cores, dual socket (8 x 2 sockets), 2.1GHz
  - Memory: 128GB physical 2400MHz
- Disks (SSD)

- 24 400GB SSDs per node, RAID 5
- Hypervisor
- KVM Ubuntu 14.04

Configuration Information

- 1500MTU for the data path between the clients and the Select cluster
- No storage efficiency features in use (compression, deduplication, Snapshot copies, SnapMirror, and so on)

Results

Select Premium: KVM Sequential Read and Write

Select Premium: Gen2 Cisco hardware (Figures 52, Figure 53, and Figure 54).

Figure 40) Comparison between different versions of ONTAP Select Premium and hypervisors: maximum throughput per HA pair.

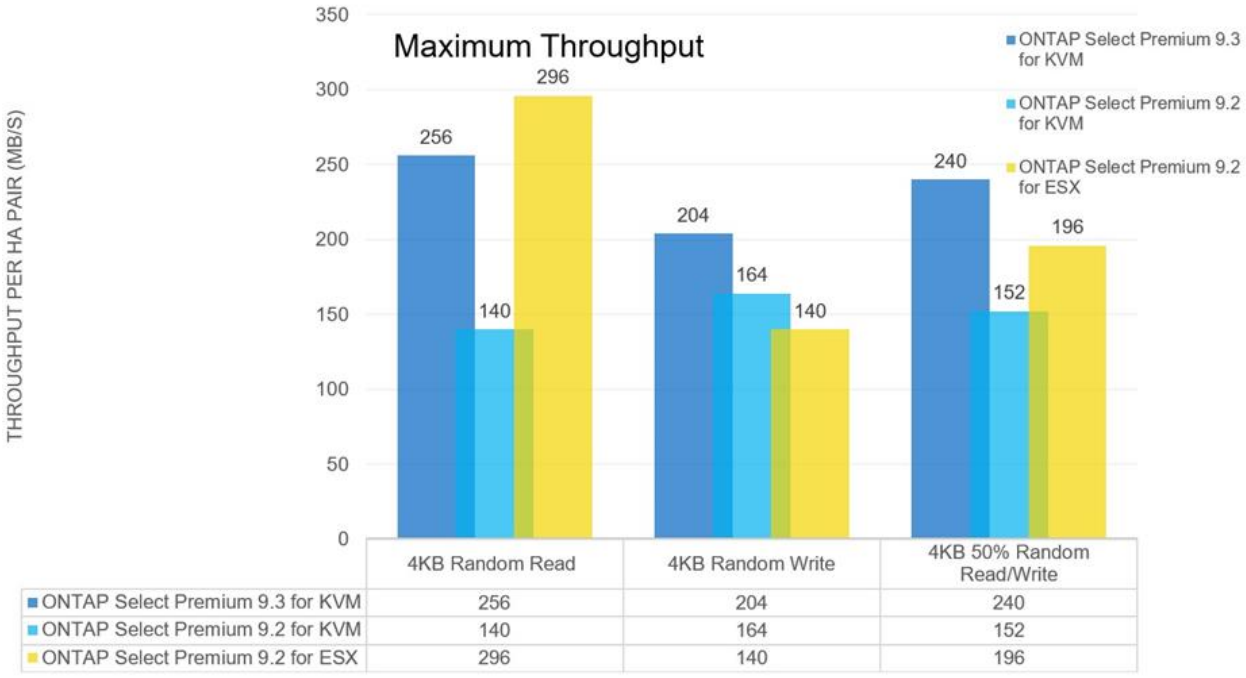


Figure 41) Comparison between different versions of ONTAP Select Premium and hypervisors: throughput at 1ms latency per HA pair.

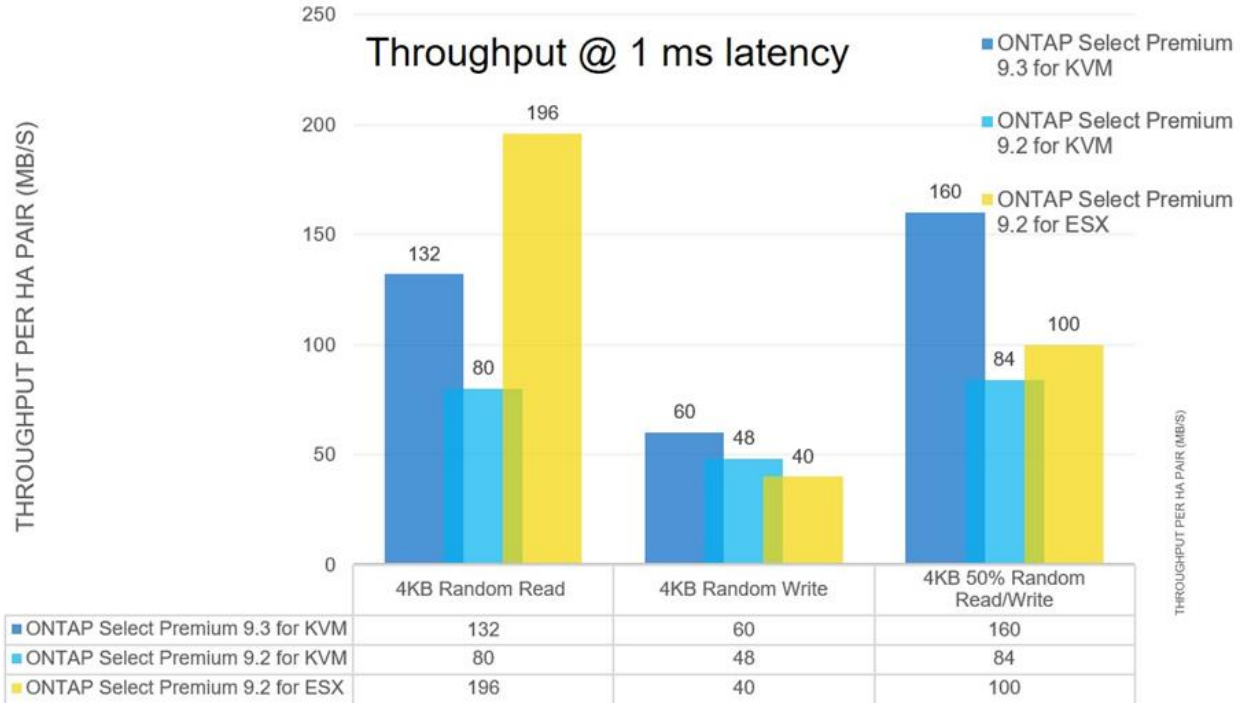
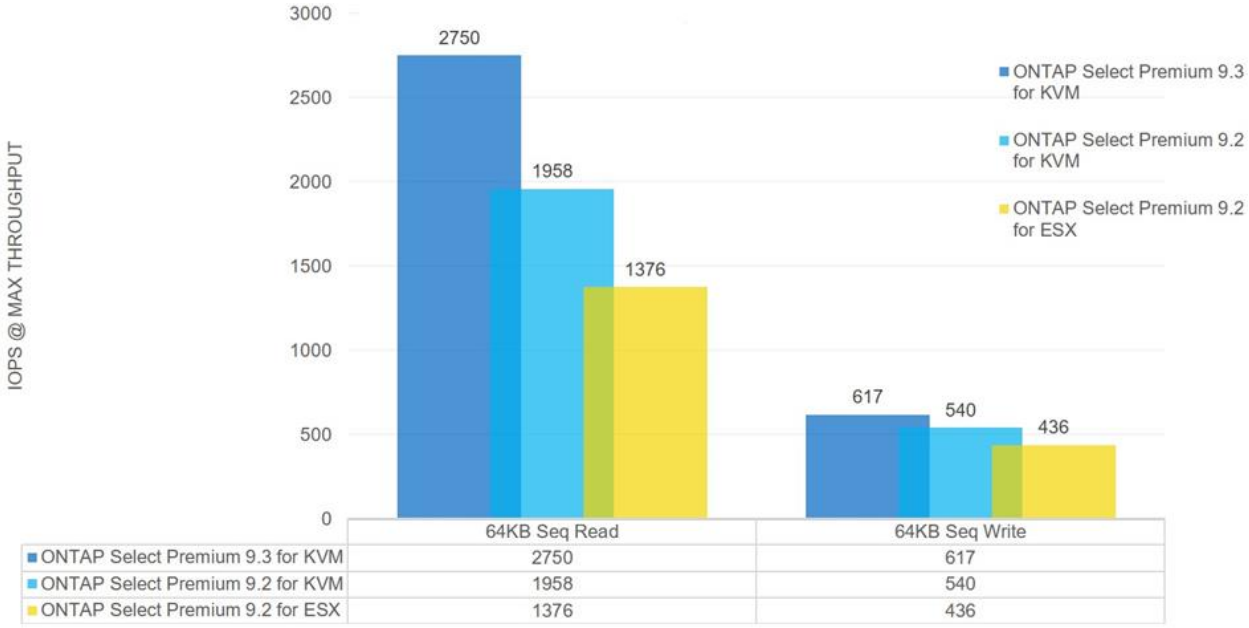


Figure 42) Comparison between different versions of ONTAP Select Premium and hypervisors: IOPS at maximum throughput.



Select Premium: Random Read and Write IOPS

Table 13) Random read and write IOPS, latency cutoff of 1ms.

Workload	Latency Cutoff	9.2 on ESX	9.2 on KVM	9.3 on KVM
4KB random read	1ms	49K	20K	33K
4KB random write	1ms	NA*	NA*	15K
4KB 50% random read/write	1ms	25K	21K	40K

Table 14) Random read and write IOPS, latency cutoff of peak tput.

Workload	Latency Cutoff	9.2 on ESX	9.2 on KVM	9.3 on KVM
4KB random read	Peak tput	74K	35K	64K
4KB random write	Peak tput	35K	41K	51K
4KB 50% random read/write	Peak tput	49K	38K	60K

Workload Details

Sequential Read

Details:

- SIO direct I/O enabled
- 1 data NIC
- 1 data aggregate (1TB)
  - 64 volumes, 64 SIO procs/threads
  - 32 volumes per node (64 total)
  - 1 SIO proc per volume, 1 SIO thread per file
  - 1 file per volume; files 12GB each
  - Files precreated using mkfile

Using 100% sequential 64KiB I/O, each thread reads through each file sequentially from beginning to end. Each measurement lasts for 300 seconds. Tests are purposefully sized so that the I/O never wraps within a given file. Performance measurements are designed to force I/O from disk.

Sequential Write

Details:

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

Using 100% sequential 64KiB I/O, each thread writes through each file sequentially from beginning to end. Each measurement lasts for 300 seconds. Tests are purposefully sized so that the I/O never wraps within a given file. Performance measurements are designed to force I/O to disk.

Random Read

Details:

- SIO direct I/O enabled
- 1 data NIC
- 1 data aggregate (1TB)
  - 64 volumes, 64 SIO procs, 512 threads
  - 32 volumes per node (64 total)
  - 64 SIO procs per volume, each with 8 threads
  - 1 SIO proc per volume, 8 threads per file
  - 1 file per volume; files are 8192MB each
  - Files precreated using mkfile

Using 100% random 4KiB I/Os, each thread randomly reads through each file. Each measurement lasts for 300 seconds. Performance measurements are designed to force I/O from disk.

Random Write

Details:

- SIO direct I/O enabled
- 1 data NIC
- 1 data aggregate (1TB)
  - 64 volumes, 128 SIO procs, 512 threads
  - 32 volumes per node (64 total)
  - 64 SIO procs, each with 8 threads
  - 1 SIO proc per volume, 8 threads per file
  - 1 file per volume; files are 8192MB each

Using 100% random 4KiB I/O, each thread randomly writes through each file. Each measurement lasts for 300 seconds. Performance measurements are designed to force I/O to disk.

8.2 SW RAID Performance on Select 9.4 Premium

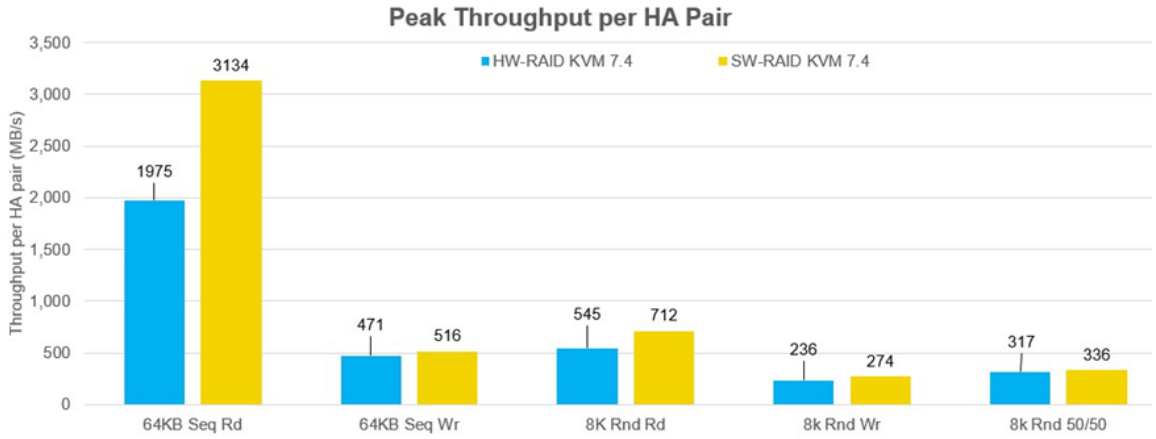
Workload and Hardware Details

- 64KB sequential read
- 64KB sequential write
- 8KB random read
- 8KB random write
- 8KB random read/write 50/50
- I/O generator
- sio
- ONTAP Select Premium
- 8-Core with SSDs
- 64GB Memory

- Gen1
  - Server (Cisco UCS C240 M4S2)
  - CPU: Intel Xeon CPU E5-2697
  - 14 cores dual socket, 2.6GHz, hyperthreading enabled
  - Memory: 128GB physical 2133 MHz
- Gen2
  - Server (Cisco UCS C240 M4SX)
  - CPU: Intel Xeon CPU E5-2620
  - 16 cores dual socket (8 x 2 sockets), 2.1GHz, hyperthreading enabled
  - Memory: 128GB physical 2400MHz
- Hypervisor - KVM Red Hat 7.4

Figure 43 compares the peak throughput per HA pair for both hardware RAID and software RAID in KVM 7.4.

Figure 43) Peak throughput per HA pair for hardware RAID versus software RAID.



8.3 SW RAID Performance on Select 9.5

Workload and Hardware Details

- 8-Core
- 64 GB Memory
- CPU: Intel Xeon CPU E5-2620
- 16 cores Dual Socket (8 x 2 sockets), 2.1 GHz
- Memory: 128 GB physical 2400 MHz
- 24 900GB SSDs per node present
- 1x 9+2 RAID-DP (11 drives)
- NVMe SSD for NVRAM
- HA-pair (2-node)
- Storage efficiency disabled
- RHEL 7.4 3.10.0-693
- Hypervisor KVM 7.4
- NFSv3

Table 15 lists the throughput (peak MiBps) measured against read/write workloads on an HA pair of ONTAP Select Premium nodes using software RAID. Performance measurements were taken using the SIO load-generating tool.

Table 15) Performance results (peak MiBps) for a single node (part of a four-node medium instance) ONTAP Select 9.5 cluster on DAS (SSD) with software 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) SW RAID	1,582	348	411	197	154

Appendix A: KVM Networking

This section provides a deeper explanation of KVM networking in general and can help with the preinstallation and preparation of the system. However, this section by no means provides best practices or otherwise serves as a sole guide to using the system. Therefore, the figures in this section should be treated as unsupported configurations and are provided to help with understanding important concepts.

General Overview and Terminology

Open vSwitch

OVS is a virtual software switch targeted at multihost virtualization deployments characterized by dynamic endpoints that forward traffic between VMs on same or different physical hosts. It also helps maintain logical interface abstractions and groupings. OVS doesn't provide a fully distributed switching capability but rather runs on individual physical hosts and provides remote management that enables distributed capabilities.

VLAN Tags

VLAN IDs help separate traffic between two different networks. They are especially useful when there are not enough NIC ports on the host to separate traffic physically. Therefore, there is a need to multiplex traffic from different networks on the same NIC port (or group of NIC ports).

To enable VLAN tags to be propagated on OVS and the physical switch, a port might need to be configured for a VLAN using two modes:

- **Access port.** The port carries traffic on exactly one VLAN specified by the tag. Any other packet on that port with an unconfigured VLAN tag is dropped. To configure an access port with VLAN ID 10, use the following command:

```
ovs-vsctl set port vnet0 vlan_mode=access tag=10
```

- **Trunk port.** The port carries traffic on one or more VLANs. The VLAN tag is retained when a packet egresses through a trunk port with the configured VLAN ID. A packet with an unconfigured VLAN tag is dropped. A packet with no VLAN tag is also forwarded. To configure a trunk port with VLAN ID 20 and 30, use the following command:

```
ovs-vsctl set port vnet0 vlan_mode=trunk trunks=20
```

Open vSwitch Bridge

An OVS bridge is a collection of ports. Ports within an OVS bridge are logical constructs and can refer to a bond (collection of ports), a physical interface, or a virtual interface (vnet port). Vnet ports are also called tap devices. Vnet ports are created during installation of the VMs and have a one-to-one correspondence with the VM ports on the VM (Figure 44).

Figure 44) Single-bridge OVS (native VLAN, untagged; unsupported ONTAP Select configuration).

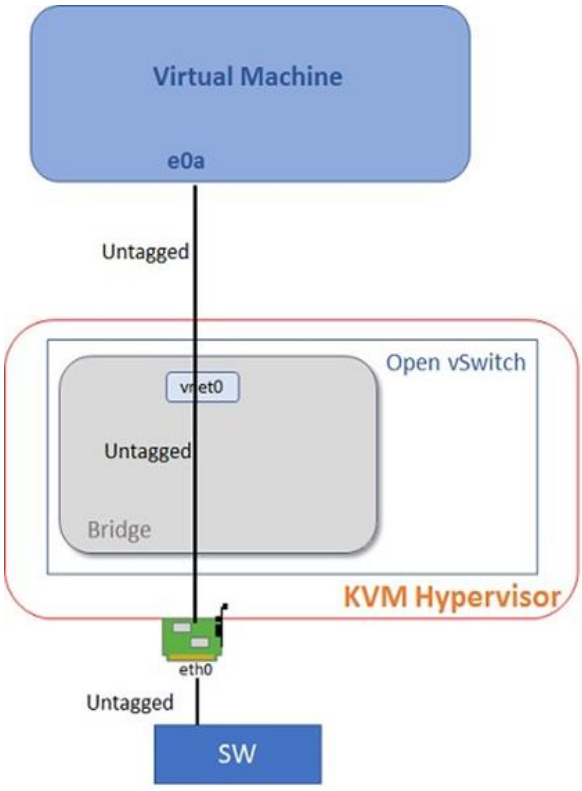
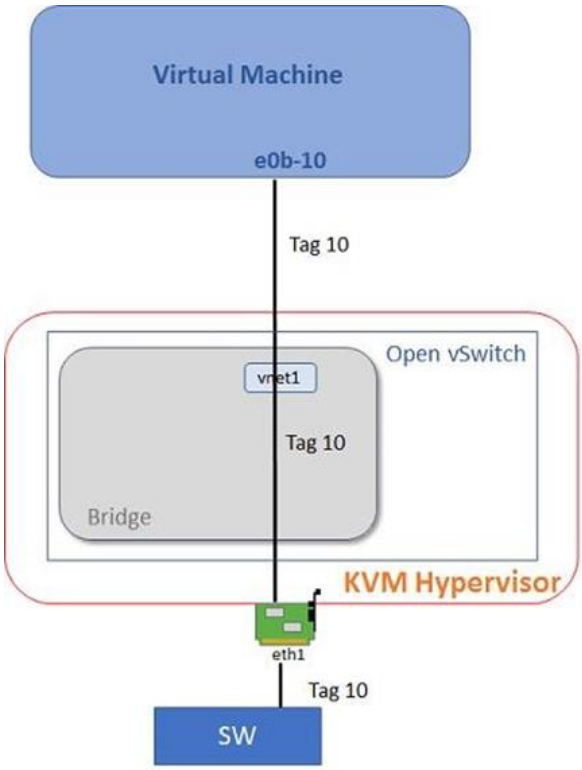
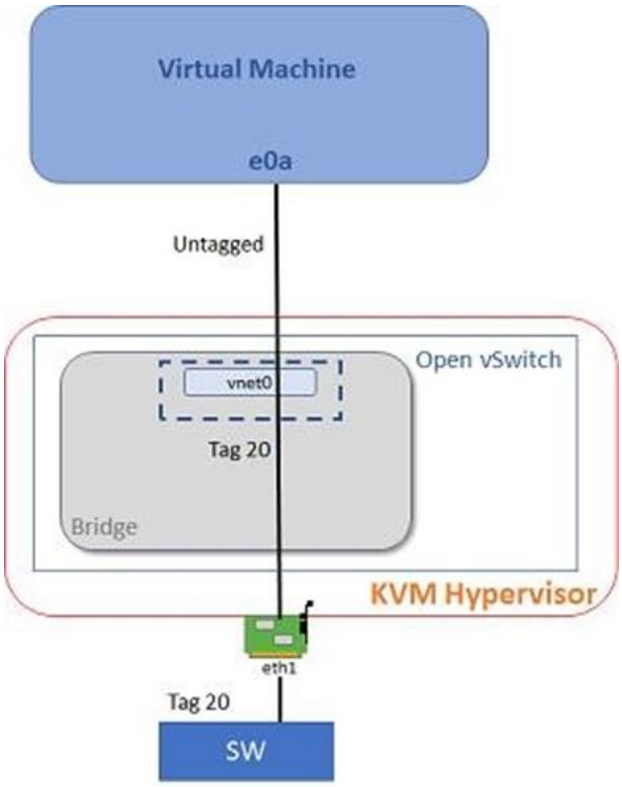


Figure 45) Single bridge (guest tagged or native tagged; unsupported ONTAP Select configuration).



In Figure 44, we see a vnet port vnet0 associated with virtual port e0a. In Figure 45, we see vnet1 associated with e0b-10. vnet0 can be configured as a native VLAN port and thus requires no changes in OVS. Vnet1 should be configured as a trunk port to enable OVS to pass tagged frames through to the VM. In Figure 45, the VM receives frames with VLAN tag 10 on virtual port e0b-10. This tagging feature in which the VLAN tagging is handled by the VM is also referred to as Virtual Guest Tagging (VGT) in VMware parlance. For our purposes, we can refer to this type of tagging as guest VLAN tagging. Note that Figure 44 and Figure 45 depict unsupported ONTAP Select configurations.

Figure 46) OVS bridge with switch tagging (unsupported single-port configuration for ONTAP Select).



In Figure 46, we see a single vnet port, vnet0, associated with virtual port e0a. In this case, vnet0 would be configured as an access port, which is the default mode. Even though it's tagged downstream from the physical switch and belongs to VLAN 20, the VM itself sees only untagged frames. This tagging feature, in which VLAN tagging is handled by OVS and tags are not seen by the VM, is also referred to as Virtual Switch Tagging (VST) in VMware parlance. For our purposes, this form of tagging can be referred to as switch tagging.

Open vSwitch Bonds

Bonds allow a logical grouping of physical interfaces to aggregate bandwidth, improve robustness, and balance load. Bonds are also equivalent to port groups in some literature.

VMs using a bond can help with the following tasks:

- Potentially help use both interfaces at the same time (for example, balance-tcp mode)
- Use one port at a time in an active/backup configuration for which one of the physical interfaces is active and the others are in standby
- Help balance load (example, balance-slb mode) when multiple VMs are using the same bond

The Link Aggregation Control Protocol (LACP) allows negotiation and use of multiple physical links between two endpoints. Bonds also present interesting ways to use VLAN tags.

Note: A mode of active-backup and balance-tcp is not supported for the current version of ONTAP Select. The only supported options are a mode of balance-slb along with LACP configured as active.

Figure 47) VMs using the same bond interface for load balancing or fault tolerance (unsupported configuration).

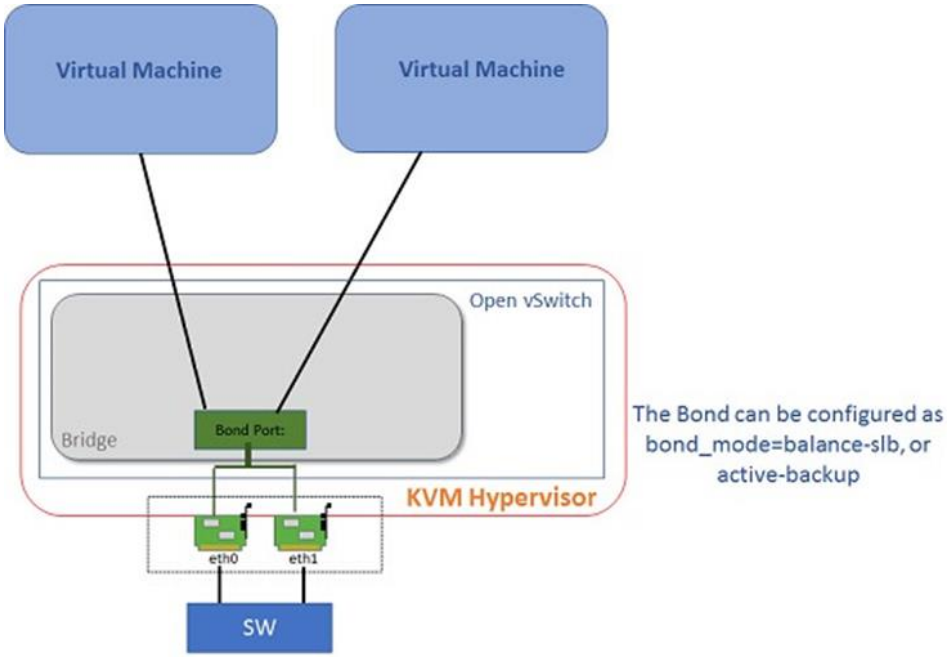
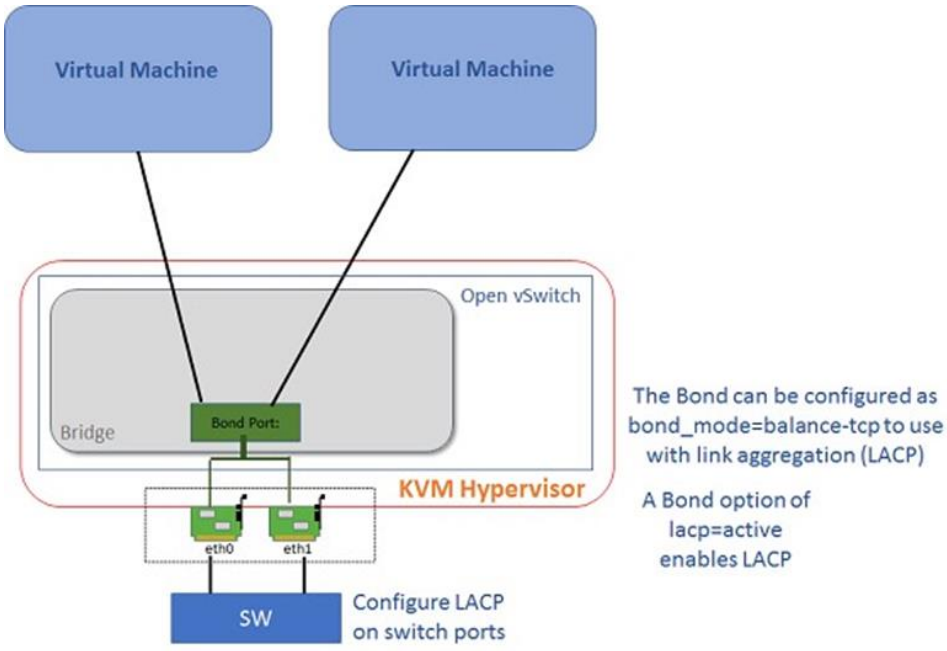


Figure 48) VMs using a bond for link aggregation (unsupported ONTAP Select configuration).



Note that Figure 47 and Figure 48 show multiple VMs using a single bond. In Figure 36, we can see that a single bond can be used to provide fault tolerance (active backup when only a single physical link is active). It can also provide load-balancing using a source MAC address hash to evenly spread traffic across the two physical interfaces (balance-slb). ONTAP Select only supports the balance-slb mode.

Figure 37 shows a configuration of balance-tcp (unsupported). This configuration can be used to provide higher aggregate throughput by enabling the physical interfaces to be used simultaneously.

The bond that ONTAP Select uses can be shared with other VMs such as the Deploy VM. ONTAP Select currently supports LACP bonds with balance-slb mode only. Hence these two configurations (Figure 36 and Figure 37) are not supported.

Open vSwitch Configuration

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

Standard Open vSwitch

All OVS configurations require a minimum of two physical network adapters bundled into a single aggregation group (referred to as a bond). On a KVM host, bonds created out of ports on the same bridge are the aggregation construct used to bundle multiple physical network adapters into a single logical channel. This configuration allows the network load to be shared across all member ports.

Bridges are logical groupings of ports, and several bonds can exist within the same bridge. It's important to remember that bonds can be created without support from the physical switch. Load balancing and failover policies can be applied directly to a bond, which is unaware of the upstream switch configuration. In this case, policies are only applied to outbound traffic. To balance inbound traffic, the physical switch must be properly configured. Port channels configured on physical switches are the primary way this is accomplished.

Bonds created out of physical ports should be configured with a mode of balance-slb. This mode makes sure that flows are load balanced among the ports based on source MAC addresses and output VLANs, with periodic rebalancing as traffic patterns change. A configuration of fast is appropriate for faster link detection failures and faster switching to other active ports. Only a bond mode of balance-slb is currently supported.

Example Commands to Configure Open vSwitch

The following commands add a bond to a bridge bridge0 using physical interfaces eth0 and eth1:

```
ovs-vsctl add-br bridge0
ovs-vsctl add-bond bridge0 bond0 eth0 eth1 bond_mode=balance-slb lacp=active other_config:lacp-time=fast
```

To set a port's MTU to 9000, use the following command:

```
ip link set eth0 mtu 9000 up
```

You might want to enable the ONTAP Select guest VM to receive tagged frames in a guest environment and use the VM interfaces as trunk VLAN ports. If so, configure the corresponding vnet (guest NIC as seen in the host) to trunk VLAN tagged frames:

```
ovs-vsctl set port vnet0 vlan_mode=trunk trunks=10,20
```

To set a vnet (VM interface as seen on the host) to a specific VLAN, use the following command:

```
ovs-vsctl set port vnet0 vlan_mode=access tag=30
```

Appendix B: Linux Logical Volume Manager

This section provides a deeper understanding of the LVM and can aid in preinstallation and preparation of host storage. However, this section by no means provides best practices or otherwise serves as a sole guide to use the system.

Physical volumes are either whole raw disks or partitions made from a single raw disk. The underlying physical storage unit of an LVM logical volume is a block device such as a partition or a whole disk. LVM allows you to create physical volumes out of disk partitions. It is recommended that you create a single partition that covers the whole physical disk (dev handle or, for example, a SCSI block device) to label as an LVM physical volume. Thus, physical volumes are either whole raw disks or partitions made from a single raw disk.

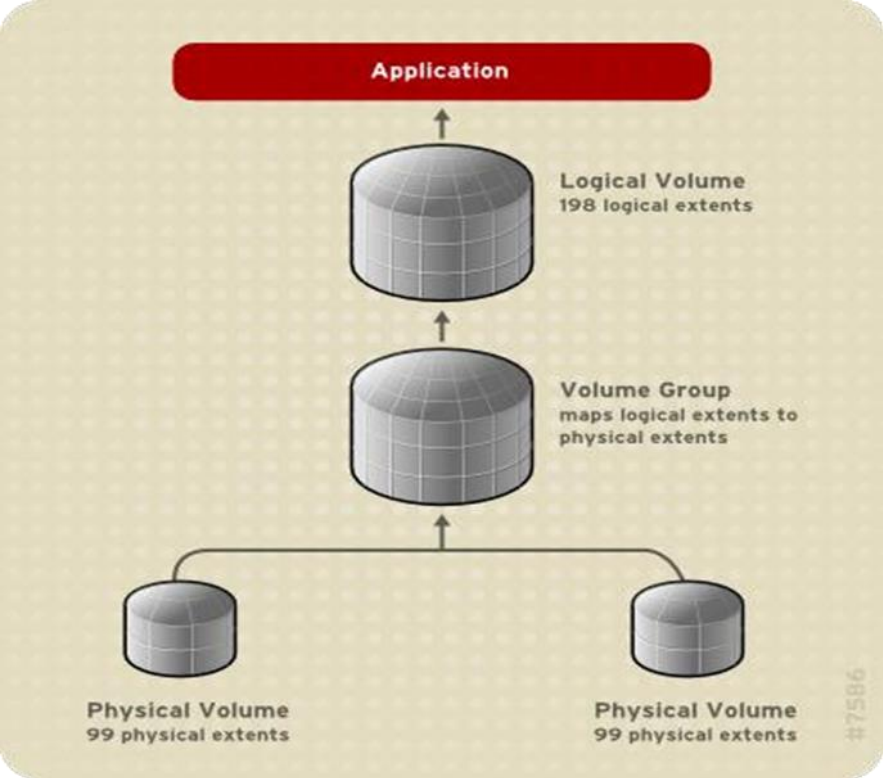
Physical volumes are combined into volume groups. This creates a pool of disk space out of which logical volumes can be allocated. Within a volume group, the disk space available for allocation is divided into units of a fixed size called extents. An extent is the smallest unit of space that can be allocated. Within a physical volume, extents are referred to as physical extents.

A logical volume is allocated into logical extents of the same size as the physical extents. The extent size is thus the same for all logical volumes in the volume group. The volume group maps the logical extents to physical extents.

A linear volume aggregates space from one or more physical volumes into one logical volume. For example, if you have two 60GB disks, you can create a 120GB logical volume. The physical storage is concatenated.

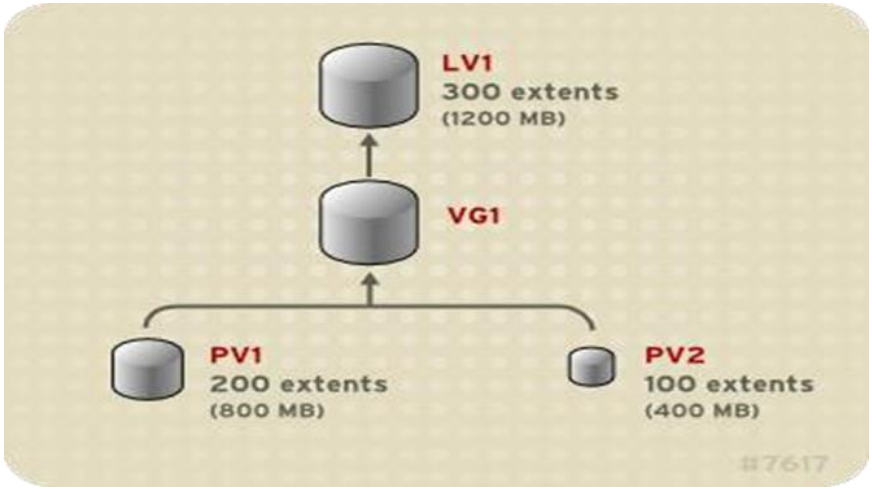
Creating a linear volume assigns a range of physical extents to an area of a logical volume in order. For example, as shown in Figure 49, logical extents 1 to 99 can map to one physical volume, and logical extents 100 to 198 can map to a second physical volume. From the point of view of the application, there is one device that is 198 extents in size.

Figure 49) Creation of a logical volume using physical extents.



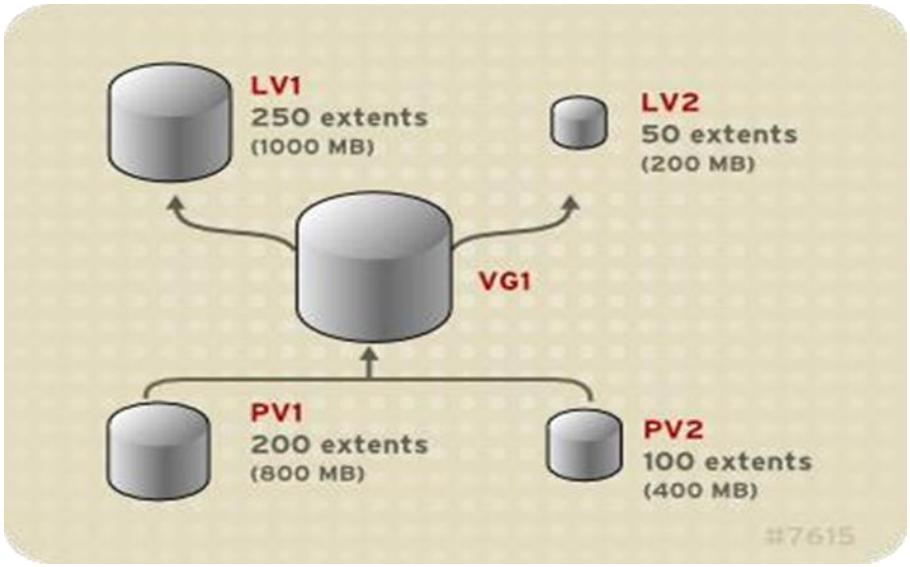
The physical volumes that make up a logical volume do not have to be the same size. Figure 50 shows volume group VG1 with a physical extent size of 4MB. This volume group includes two physical volumes named PV1 and PV2. The physical volumes are divided into 4MB units, because that is the extent size. In this example, PV1 is 200 extents in size (800MB), and PV2 is 100 extents in size (400MB). You can create a linear volume any size between 1 and 300 extents (4MB to 1200MB). Figure 50 shows that the linear volume named LV1 is 300 extents in size.

Figure 50) Creation of a logical volume from a volume group and multiple physical extents.



You can also configure more than one linear logical volume of whatever size you require from the pool of physical extents. Figure 51 shows the same volume group as in Figure 38, but in this case two logical volumes have been carved out of the volume group: LV1, which is 250 extents in size (1000MB), and LV2, which is 50 extents in size (200MB).

Figure 51) Creation of multiple logical volumes from a volume group and physical extents.



Creating Volume Groups

To create a volume group from one or more physical volumes, use the vgcreate command. The vgcreate command creates a new volume group by name and adds at least one physical volume to it.

The following command creates a volume group named vg1 that contains physical volumes `/dev/sdd1` and `/dev/sde1`:

```
# vgcreate vg1 /dev/sdd1 /dev/sde1
```

When physical volumes are used to create a volume group, its disk space is divided into 4MB extents by default. This extent is the minimum amount by which the logical volume can be increased or decreased in size. A large number extents have no effect on I/O performance of the logical volume.

You can specify the extent size with the -s option to the `vgcreate` command if the default extent size is not suitable. You can put limits on the number of physical or logical volumes the volume group can have by using the -p and -l arguments of the `vgcreate` command.

Extending Volume Groups

To add additional physical volumes to an existing volume group, use the `vgextend` command. The `vgextend` command increases a volume group's capacity by adding one or more free physical volumes. The following command adds the physical volume /dev/sdf1 to the volume group vg1:

```
# vgextend vg1 /dev/sdf1
```

Creating Logical Volumes from Volume Groups

To create a logical volume, use the `lvcreate` command. When you create a logical volume, the logical volume is carved from a volume group using the free extents on the physical volumes that make up the volume group.

The default unit for logical volume size is MB. The following command creates a 1500MB linear logical volume named testlv in the volume group testvg, creating the block device `/dev/testvg/testlv`.

```
# lvcreate -L 1500 -n testlv testvg
```

The following command creates a 50GB logical volume named gfslv from the free extents in volume group vg0:

```
# lvcreate -L 50G -n gfslv vg0
```

To create a logical volume to be allocated from a specific physical volume in the volume group, specify the physical volume or volumes at the end at the `lvcreate` command line. The following command creates a logical volume named testlv in volume group testvg allocated from the physical volume `/dev/sdg1`:

```
# lvcreate -L 1500 -n testlv testvg /dev/sdg1
```

Appendix C: Linux External Shared Storage and Host Clustering

There are different ways of using external shared storage on Linux. Primary among them is to configure external storage as storage pools. However, LUNs exposed through physical interfaces such as FC/FCoE can be also used. Some examples of storage pools and LUNs are as follows.

Storage Pools

- **iSCSI pool.** An iSCSI pool provides aggregate storage based on an iSCSI target. Volumes (LUNs) are pre-allocated on the iSCSI server and cannot be created with the libvirt APIs. For an iSCSI storage pool, there are three pieces of information that you must provide. The source is the unique identified (IQN) used when creating the iSCSI target. The target determines how libvirt exposes device paths for the pool. The host is the fully qualified domain name/IP address of the iSCSI server.

It is possible to create a volume group (storage pool) local to the host that is not visible to all the hosts accessing the shared storage.

- **NFS pools.** NFS pools allow you to manage storage with NFS, CIFS (SMB or CIFS), or FUSE (glusterFS) protocols. In NFS pools, a directory is considered a storage pool and each file within the directory is a volume. For a NFS storage pool, there are three pieces of information that you must provide. The host name is the storage server name. The source path is the path of the directory on the storage server. The target path directory is where the storage pool is created (mounted) on the host. The following volume formats are supported by libvirt: raw, bochs, cloop, cow, dmg, iso, qcow, qcow2, qed, vmdk, and vpc. The raw volume format is always available.

Storage LUNs

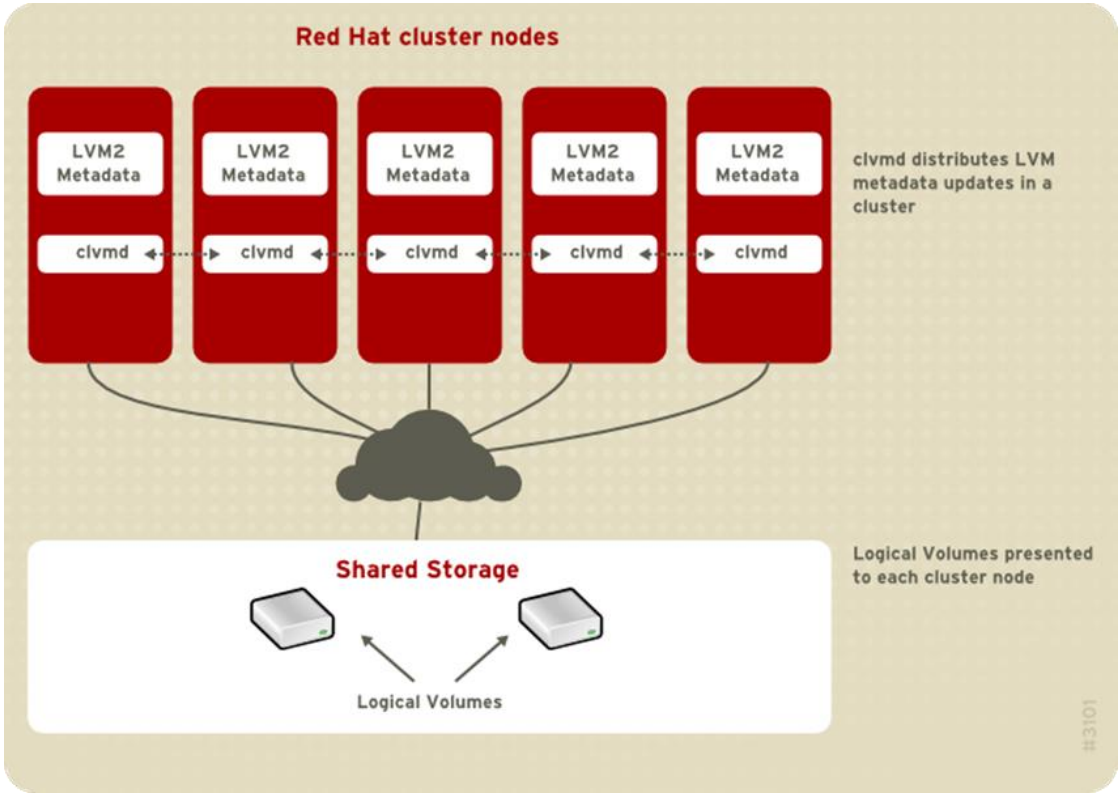
- **FC/FCoE LUNs.** When an external array is connected through an FC interface on the back-end, such as to a SAN, the Linux host sees a group of LUNs that the FC driver exposes. These LUNs can then be used to carve out storage for the guest VM.

Clustered Logical Volume Manager

Next, we cover components required for clustering. Clustering of multiple Linux hosts sharing external storage can be accomplished by using CLVM, Pacemaker, and the PCS utility.

CLVM must be used when there is a need to create LVM logical volumes from external storage that then must be shared by multiple hosts. Creating LVM logical volumes in a cluster environment is identical to creating LVM logical volumes on a single host. There is no difference in the LVM commands themselves. To enable the LVM volumes, the cluster infrastructure must be running and the cluster must be in quorum. By default, logical volumes created with CLVM on shared storage are visible to all systems that have access to the shared storage (Figure 52).

Figure 52) Sharing storage between multiple hosts enabled by CLVM.



Pacemaker

Pacemaker and the PCS utility should be used along with CLVM. Pacemaker is responsible for all cluster-related activities, such as monitoring cluster membership, managing the services and resources, and fencing cluster members. PCS provides a CLI to create, configure, and control every aspect of a Pacemaker cluster.

Pacemaker is a high availability resource manager with many useful features:

- Detection and recovery from machine and application-level failures
- Support for many redundancy configurations
- Support for quorate and resource-driven clusters
- Configurable strategies for dealing with quorum loss when multiple machines fail
- Support for specifying application startup and shutdown ordering, regardless of which machine the applications are on
- Support for specifying that applications must or must not run on the same machine
- Support for specifying that an application should be active on multiple machines
- Support for multiple modes for applications, such as master and slave
- Provably correct responses to any failure or cluster state
- Offline testing of responses to any situation, before the situation arises

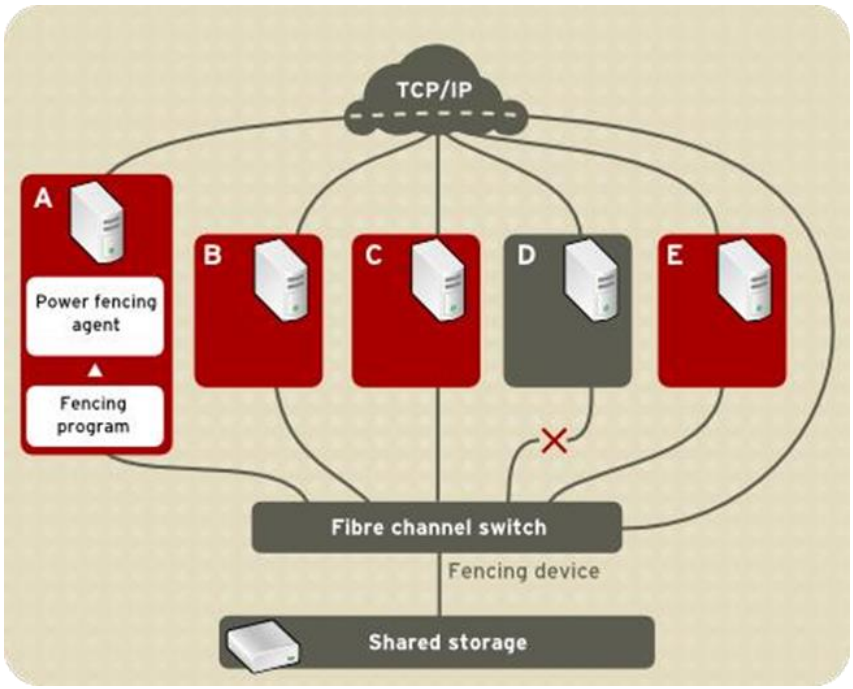
Fencing

If communication with a single host in the cluster fails, then other hosts in the cluster must restrict or release access to resources that the failed cluster host might have access to. This cannot be accomplished by contacting the cluster host itself because the cluster host might not be responsive. Instead, an external entity must be used. This external entity is called a fence agent or a fence device. It can be used by the cluster to restrict access to shared resources by an errant host or to issue a hard reboot on the cluster host.

Without a fence device configured, there is no way to know whether the resources previously used by the disconnected cluster host have been released. This is a situation that could prevent services from running on the other cluster hosts. Conversely, the system might assume erroneously that the cluster host has released its resources, which can lead to data corruption and data loss. Without a fence device configured, data integrity cannot be guaranteed, and the cluster configuration is unsupported. When fencing is in progress, no other cluster operation is allowed to run.

In Figure 53, the fencing program in node A causes the FC switch to disable the port for node D, disconnecting node D from storage.

Figure 53) Fencing used to enable access from other hosts during a host disruption.



Configuring External Shared Storage with Clustered Logical Volume Manager and Pacemaker

1. To configure external shared storage with CLVM and Pacemaker, complete the following steps on all the hosts of the cluster:

   a. Install the required packages.

```
fence-agents-all, lvm2-cluster, pacemaker, pcs
```

   b. Enable cluster locking.

```
$ lvmconf --enable-cluster
```

   c. Verify that the locking type is 3 for clustering-wide locking.

```
$ cat /etc/lvm/lvm.conf |grep -v '#' | grep locking_type
locking_type = 3
```

   d. Start the Pacemaker daemon.

```
$ systemctl start pcsd
```

   e. Enable Pacemaker to start automatically on host boot-up.

```
$ systemctl enable pcsd
```

   f. Set the PCS cluster admin password (the same password) in the cluster for the PCS admin, hacluster.

```
$ passwd hacluster
```

   g. Enable the ports required by the Linux high-availability add-on.

```
firewall-cmd --permanent --add-service=high-availability
```

   h. Verify that the PCS daemon is enabled and active.

```
$ systemctl is-enabled pcsd
$ systemctl is-active pcsd
```

2. Perform the following steps on one host in the cluster:
   a. Authenticate each host in the cluster on the host where you will run PCS.

```
$ pcs cluster auth sdot-node-1.company.com sdot-node-2.company.com
```

   b. Create the cluster.

```
$ pcs cluster setup --force --name test_clvm sdot-node-1.company.com sdot-node-2.company.com
```

   c. Start the cluster.

```
$ pcs cluster start –all o Check the cluster is online
$ pcs cluster status
```

3. Enable fencing.
4. Use an iSCSI LUN as a STONITH device.

Note: STONITH (shoot the other node in the head) is a mechanism to maintain the integrity of hosts in a high-availability cluster. A STONITH process automatically powers down a host that is not working correctly. For example, a STONITH process can be used if one of the hosts in a cluster cannot be reached by the other hosts in the cluster.

```
$ ll /dev/disk/by-id | grep sdq lrwxrwxrwx. 1 root root 9 Aug 13 19:23 scsi-
3600a098038303041765d4b32656d7461 -> ../../sdq
lrwxrwxrwx. 1 root root 9 Aug 13 19:23 wwn-
0x600a098038303041765d4b32656d7461 -> ../../sdq

$ iscsiadm -m node -T iqn.1992-
08.com.company:sn.f5e4b6654c2a11e79d9900a09854e741:vs.12 -p 10.20.19.28 -l

$ pcs stonith create scsi-shooter fence_scsi devices=/dev/disk/byid/wwn-
0x600a098038303041765d4b32656d7461 meta provides=unfencing

$ pcs property set no-quorum-policy=freeze
$ pcs stonith show scsi-shooter
```

5. Create DLM and CLVMD resources.

Note: DLM is a distributed lock manager that runs in each cluster host. Lock management is distributed across all hosts in the cluster. CLVM uses locks from the lock manager to synchronize updates to LVM volumes and volume groups (also on shared storage).

DLM works along with Pacemaker and CLVMD. DLM provides its own mechanisms to support its locking features. These mechanisms include inter-host communication to manage lock traffic and recovery protocols to remaster locks after a host failure or to migrate locks when a host joins the cluster.

Note: To make sure that resources remain healthy, you can add a monitoring operation to a resource's definition. In this case, resources are created to monitor DLM and CLVMD.

```
$ pcs resource create dlm ocf:pacemaker:controld op monitor interval=30s on-fail=fence clone
interleave=true ordered=true
$ pcs resource create clvmd ocf:heartbeat:clvm op monitor interval=30s on-fail=fence clone
interleave=true ordered=true
$ pcs constraint order start dlm-clone then clvmd-clone Adding dlm-clone clvmd-clone (kind:
Mandatory) (Options: firstaction=start then-action=start)
$ pcs constraint colocation add clvmd-clone with dlm-clone
```

6. Make sure that the resources are started.

```
$ pcs status resources
```

7. Discover and list the shared LUNs.
   a. To perform iSCSI LUN discovery and to log in from the KVM host, enter the following commands to discover and log in to the iSCSI LUN:

```
$ iscsiadm -m discovery -t st -p <iSCSI lif IP>
```

```
$ iscsiadm -m node -T <Target's (backend FAS) iqn> -p <iSCSI LIF IP> -l
```

For example, suppose that an iSCSI device is visible from the host with the following properties:

```
iqn.1992-08.com.netapp:sn.f5e4b6654c2a11e79d9900a09854e741:vs.12 -p
10.270.19.28 -l
```

b. List the discovered LUNs on the KVM host.

```
$ multipath -ll
```

8. Create shared FC LUNs on the host. Suppose that an FC LUN has been exported to the host, and the host sees the 19.51GB physical volume /dev/mapper/mpatha2, as shown:

```
$ pvs
 PV VG Fmt Attr PSize PFree
 /dev/mapper/mpatha2 rhel lvm2 a-- 19.51g 40.00m
$ vgs
 VG #PV #LV #SN Attr VSize VFree  rhel 1 2 0 wz--n- 19.51g 40.00m
```

Then create a storage pool out of the FC LUN. For example, create a pool called `ontap-select-storage-pool` on both hosts sharing the FC LUN. mpathb is the exposed FC LUN.

```
$ virsh pool-define-as ontap-select-storage-pool logical --source-dev /dev/mapper/mpathb --
target=/dev/ontap-select-storage-pool
```

Appendix D: Guest VM Migration on Linux

VM migration is possible because guest VMs run in a virtualized environment instead of directly on the hardware. Migration works by transferring the guest VM's memory state and virtualized devices to another host machine. NetApp recommends that guest VMs use shared, networked storage to store images for migration. NetApp also recommends basing this shared storage on libvirt-managed storage pools when migrating VMs.

Live Migration of Online VMs

In a live migration, the guest VM continues to run on the source host machine while its memory pages are transferred to the destination host physical machine. During migration, KVM monitors the source for any changes in pages it has already transferred and begins to transfer these changes when all of the initial pages have been transferred.

KVM also estimates transfer speed during migration. The remaining amount of data to transfer takes a certain configurable period of time (10ms by default). So KVM suspends the original guest VM, transfers the remaining data, and resumes the same guest VM on the destination host physical machine. If the original guest VM modifies pages faster than KVM can transfer them to the destination, then offline migration must be used, because live migration would never complete.

Migration of Offline VMs

For offline migration, the guest VM is suspended. An image of the guest VM's memory is then moved to the destination host physical machine. The guest VM is then resumed on the destination host physical machine, and the memory the guest VM used on the source host physical machine is freed. The time it takes to complete an offline migration depends on network bandwidth and latency. If the network is experiencing heavy use or low bandwidth, the migration takes much longer.

Migration Example

The following example shows how to perform live migration of the ONTAP Select VM test-vm-01 from the host sdot-node-1 to the target sdot-node-2. The node test-vm-01 belongs to a single-node cluster.

```
[root@sdot-node-1 ]# virsh migrate test-vm-01 --undefinesource -persistent --live --verbose --
desturi qemu+ssh://sdot-node-2/system
```

The following example shows how to perform offline migration of the ONTAP Select VM test-vm-01 from the host sdot-node-1 to sdot-node-2 when the node is offline.

```
[root@sdot-node-1 ]# virsh migrate test-vm-migration-01 --undefinesource -
-persistent --offline --verbose --desturi qemu+ssh://sdot-node-2/system
```

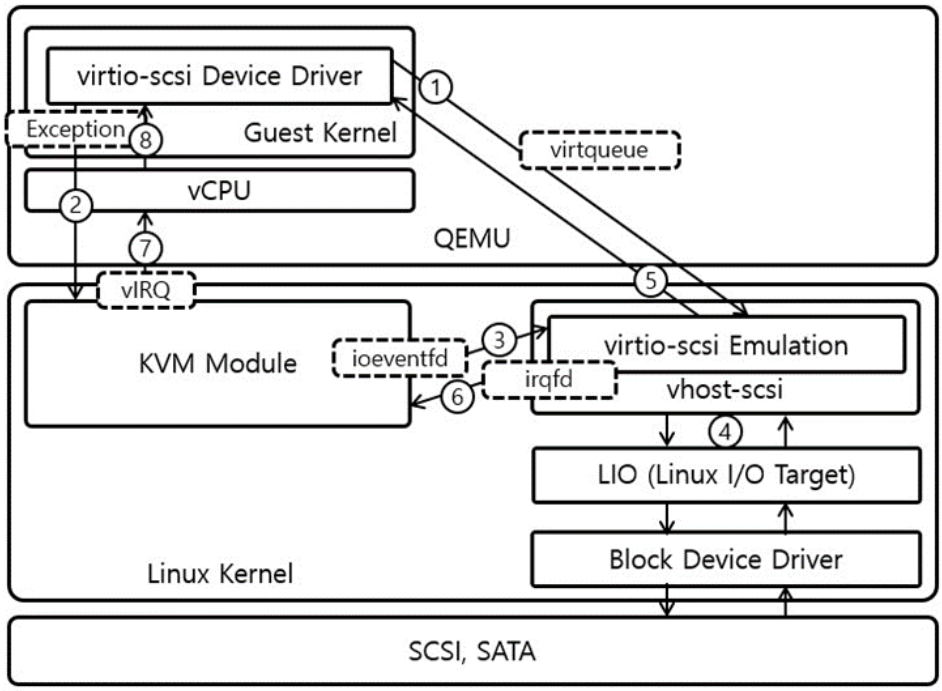
Appendix E: VirtIO Interface for Raw Device Mapping

The vhost drivers in Linux provide in-kernel virtio device emulation. Normally, the QEMU user-space process emulates I/O accesses from the guest. Vhost puts virtio emulation code into the kernel, taking QEMU user-space out of the picture. This allows device emulation code to directly call into kernel subsystems instead of performing system calls from user space.

Vhost has no dependency on the KVM kernel module. It is a user-space interface and does not emulate a complete virtio PCI adapter. Instead it restricts itself to virtqueue operations only. This means a vhost driver is not a self-contained virtio device implementation, but rather depends on the user-space to handle the control plane while the data plane is performed in-kernel. The vhost instance only knows about guest memory mapping, a kick eventfd, and a call eventfd.

Figure 54 shows the typical interactions during SCSI commands and response processing.

Figure 54) virtio-scsi interface and vhost-scsi emulation.



I/O Processing

Vhost processes I/O from Linux guests to LIO backstores as follows:

1. The KVM guest enqueues the SCSI I/O descriptor(s) to its virtio ring.
2. The KVM guest kicks LIO to wake up.
3. LIO wakes up, dequeues the I/O descriptor(s) off the virtio ring and processes them.
4. LIO dispatches the I/O to the back-end storage device (HDDs, SSDs, flash, RAM, and so on).
5. The LIO back-end storage device completes the I/O.
6. LIO enqueues the resulting I/O descriptor(s) to the KVM guest virtio ring.
7. LIO kicks the KVM guest to wake up.
8. The KVM guest wakes up and dequeues the I/O descriptor(s) off the virtio ring.

Where to Find Additional Information

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

- Red Hat Linux Logical Volume Manager Administration
  https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/htmlsingle/Logical_Volume_Manager_Administration/index.html
- Red Hat Linux Open vSwitch Bonding Options
  https://access.redhat.com/documentation/en-us/red_hat_openstack_platform/8/html/director_installation_and_usage/appe-bonding_options
- Bonding on Open vSwitch
  http://brezular.com/2011/12/04/openvswitch-playing-with-bonding-on-openvswitch/
- Scott's Weblog on OVS
  http://blog.scottlowe.org/2012/10/19/link-aggregation-and-lacp-with-open-vswitch/
- Open vSwitch VLAN Configuration FAQ
  http://docs.openvswitch.org/en/latest/faq/vlan/
- Creating a RedHat High-Availability Cluster with Pacemaker
  https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/html/High_Availability_AddOn_Administration/ch-startup-HAAA.html
- LVM Configuration Examples
  https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/logical_volume_manager_administration/lvm_administration#cluster_setup
- Basic Cluster configuration examples
  http://clusterlabs.org/quickstart-redhat.html
- Pacemaker Configuration
  https://www.server-world.info/en/note?os=CentOS_7&p=pacemaker&f=3
- Live Migration with Libvirt
  https://libvirt.org/migration.html
- KVM Migration
  https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/html/Virtualization_Deployment_and_Administration_Guide/chap-KVM_live_migration.html
- LVM Clustering Overview
  https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/logical_volume_manager_administration/lvm_cluster_overview
- KVM Live Migration with Virsh
  https://access.redhat.com/documentation/enus/red_hat_enterprise_linux/7/html/virtualization_deployment_and_administration_guide/sect-kvm_live_migration-live_kvm_migration_with_virsh

Version History

Version	Date	Document Version History
Version 1.0	June 23, 2017	Initial version for ONTAP 9.2 Deploy 2.5
Version 2.0	Oct 31, 2017	Updates for ONTAP Deploy 2.6
Version 2.1	March 12, 2018	Updates for ONTAP 9.3 Deploy 2.7
Version 3	June 15, 2018	Software RAID, V3, multiple instances, multinode with ONTAP 9.4 Deploy 2.8
Version 4	August 2018	ONTAP 9.4 Deploy 2.9 kernel and packages update
Version 5	April 2019	Updates for ONTAP 9.5 Deploy 2.11.

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.

Copyright Information

Trademark Information