Storage

ONTAP Select

NetApp
June 11, 2024

Table of Contents

Storage

Storage: General concepts and characteristics

Discover general storage concepts that apply to the ONTAP Select environment before exploring the specific storage components.

Phases of storage configuration

The major configuration phases of the ONTAP Select host storage include the following:

- Pre-deployment prerequisites
  - Make sure that each hypervisor host is configured and ready for an ONTAP Select deployment.
  - The configuration involves the physical drives, RAID controllers and groups, LUNs, as well as related network preparation.
  - This configuration is performed outside of ONTAP Select.
- Configuration using the hypervisor administrator utility
  - You can configure certain aspects of the storage using the hypervisor administration utility (for example, vSphere in a VMware environment).
  - This configuration is performed outside of ONTAP Select.
- Configuration using the ONTAP Select Deploy administration utility
  - You can use the Deploy administration utility to configure the core logical storage constructs.
  - This is performed either explicitly through CLI commands or automatically by the utility as part of a deployment.
- Post-deployment configuration
  - After an ONTAP Select deployment completes, you can configure the cluster using the ONTAP CLI or System Manager.
  - This configuration is performed outside of ONTAP Select Deploy.

Managed versus unmanaged storage

Storage that is accessed and directly controlled by ONTAP Select is managed storage. Any other storage on the same hypervisor host is unmanaged storage.

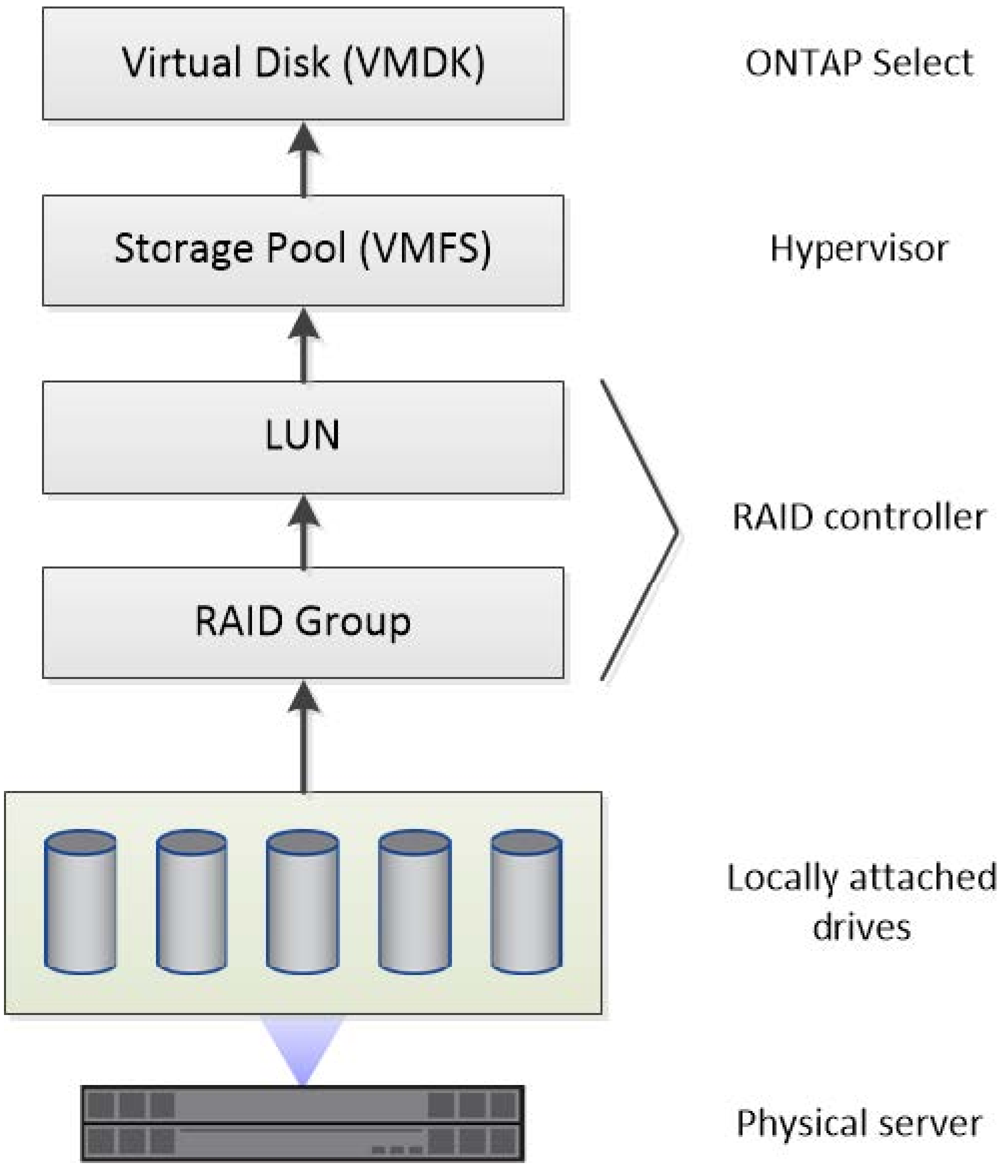
Homogeneous physical storage

All the physical drives comprising the ONTAP Select managed storage must be homogeneous. That is, all the hardware must be the same regarding the following characteristics:

- Type (SAS, NL-SAS, SATA, SSD)
- Speed (RPM)

Illustration of the local storage environment

Each hypervisor host contains local disks and other logical storage components that can be used by ONTAP Select. These storage components are arranged in a layered structure, from the physical disk.



Characteristics of the local storage components

There are several concepts that apply to the local storage components used in an ONTAP Select environment. You should be familiar with these concepts before preparing for an ONTAP Select deployment. These concepts are arranged according to category: RAID groups and LUNs, storage pools, and virtual disks.

Grouping physical drives into RAID groups and LUNs

One or more physical disks can be locally attached to the host server and available to ONTAP Select. The physical disks are assigned to RAID groups, which are then presented to the hypervisor host operating system as one or more LUNs. Each LUN is presented to the hypervisor host operating system as a physical hard drive.

When configuring an ONTAP Select host, you should be aware of the following:

- All managed storage must be accessible through a single RAID controller
- Depending on the vendor, each RAID controller supports a maximum number of drives per RAID group

One or more RAID groups

Each ONTAP Select host must have a single RAID controller. You should create a single RAID group for ONTAP Select. However, in certain situations you might consider creating more than one RAID group. Refer to Summary of best practices.

Storage pool considerations

There are several issues related to the storage pools that you should be aware of as part of preparing to deploy ONTAP Select.

> ⓘ In a VMware environment, a storage pool is synonymous with a VMware datastore.

Storage pools and LUNs

Each LUN is seen as a local disk on the hypervisor host and can be part of one storage pool. Each storage pool is formatted with a file system that the hypervisor host OS can use.

You must make sure that the storage pools are created properly as part of an ONTAP Select deployment. You can create a storage pool using the hypervisor administration tool. For example, with VMware you can use the vSphere client to create a storage pool. The storage pool is then passed in to the ONTAP Select Deploy administration utility.

Manage the virtual disks on ESXi

There are several issues related to the virtual disks that you should be aware of as part of preparing to deploy ONTAP Select.

Virtual disks and file systems

The ONTAP Select virtual machine is allocated multiple virtual disk drives. Each virtual disk is actually a file contained in a storage pool and is maintained by the hypervisor. There are several types of disks used by ONTAP Select, primarily system disks and data disks.

You should also be aware of the following regarding virtual disks:

- The storage pool must be available before the virtual disks can be created.
- The virtual disks cannot be created before the virtual machine is created.
- You must rely on the ONTAP Select Deploy administration utility to create all virtual disks (that is, an administrator must never create a virtual disk outside of the Deploy utility).

Configuring the virtual disks

The virtual disks are managed by ONTAP Select. They are created automatically when you create a cluster using the Deploy administration utility.

Illustration of the external storage environment on ESXi

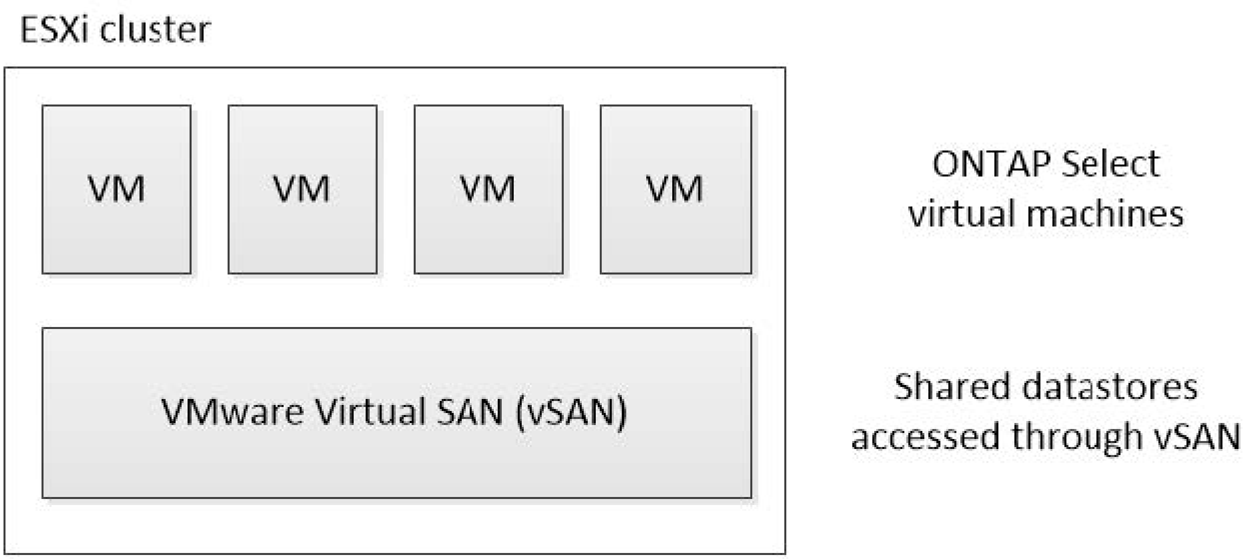
The ONTAP Select vNAS solution enables ONTAP Select to use datastores residing on storage that is external to the hypervisor host. The datastores can be accessed through the network using VMware vSAN or directly at an external storage array.

ONTAP Select can be configured to use the following types of VMware ESXi network datastores which are external to the hypervisor host:

- vSAN (Virtual SAN)
- VMFS
- NFS

vSAN datastores

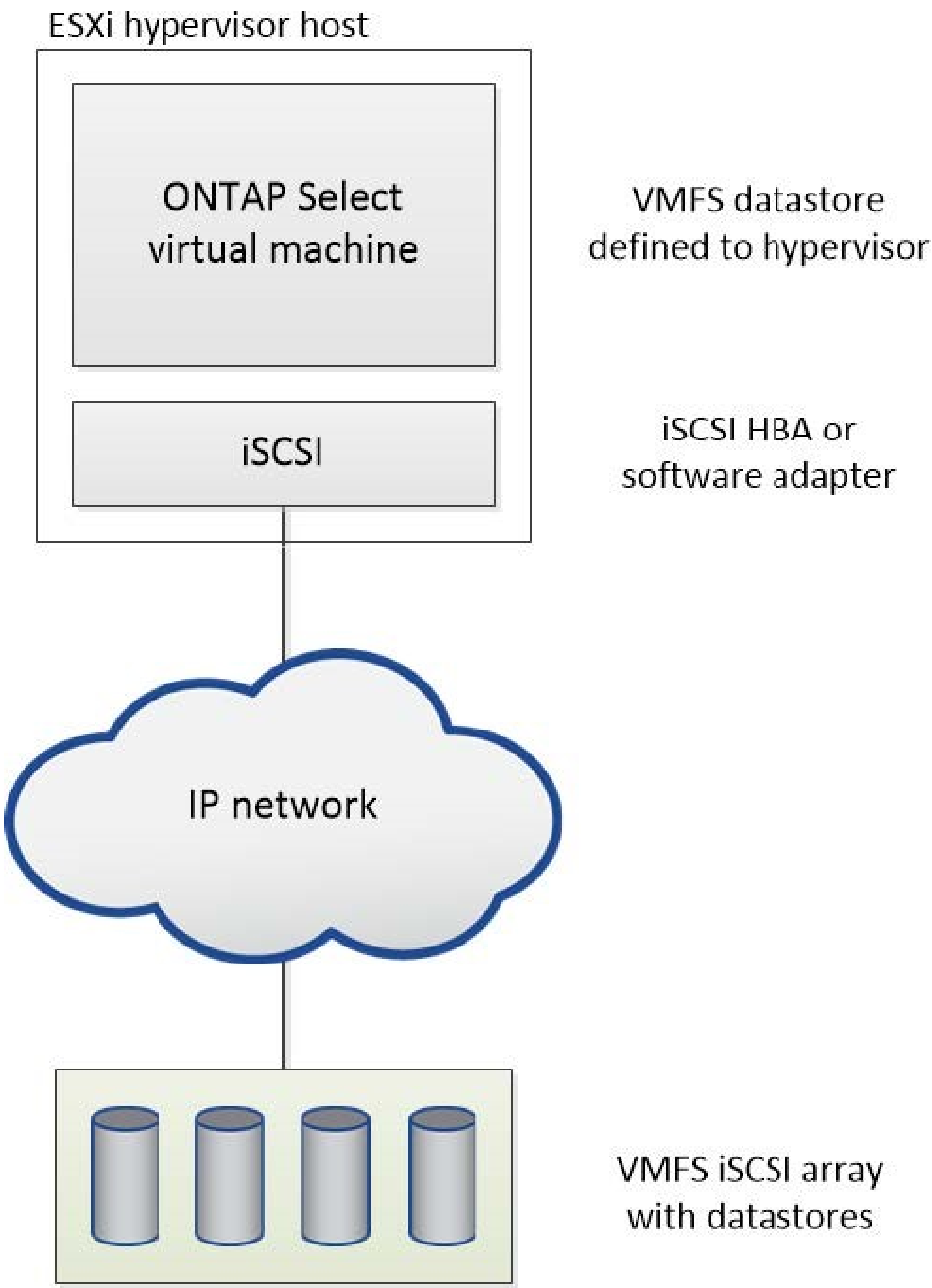
Every ESXi host can have one or more local VMFS datastores. Normally these datastores are only accessible to the local host. However, VMware vSAN allows each of the hosts in an ESXi cluster to share all of the datastores in the cluster as if they were local. The following figure illustrates how vSAN creates a pool of datastores that are shared among the hosts in the ESXi cluster.



VMFS datastore on external storage array

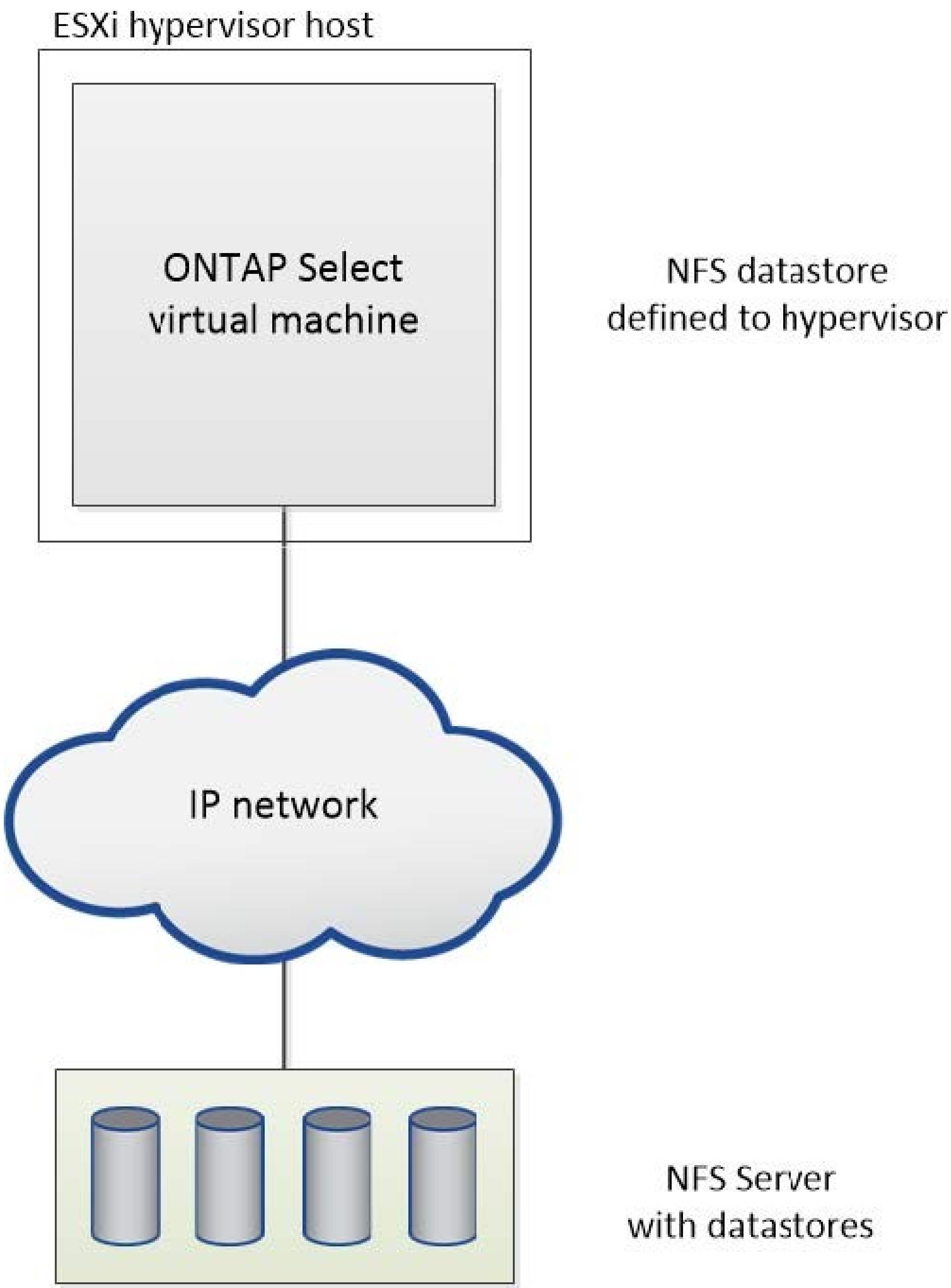
You can create a VMFS datastore residing on an external storage array. The storage is accessed using one of several different network protocols. The following figure illustrates a VMFS datastore on an external storage array accessed using the iSCSI protocol.

ⓘ ONTAP Select supports all external storage arrays described in the VMware Storage/SAN Compatibility documentation, including iSCSI, Fiber Channel, and Fiber Channel over Ethernet.



NFS datastore on external storage array

You can create an NFS datastore residing on an external storage array. The storage is accessed using the NFS network protocol. The following figure illustrates an NFS datastore on external storage that is accessed through the NFS server appliance.



Hardware RAID services for local attached storage

When a hardware RAID controller is available, ONTAP Select can move RAID services to the hardware controller for both a write performance boost and protection against physical drive failures. As a result, RAID protection for all nodes within the ONTAP Select cluster is provided by the locally attached RAID controller and not through ONTAP software RAID.

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

RAID controller configuration for local attached storage

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

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

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

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

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

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

RAID mode

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

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

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

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

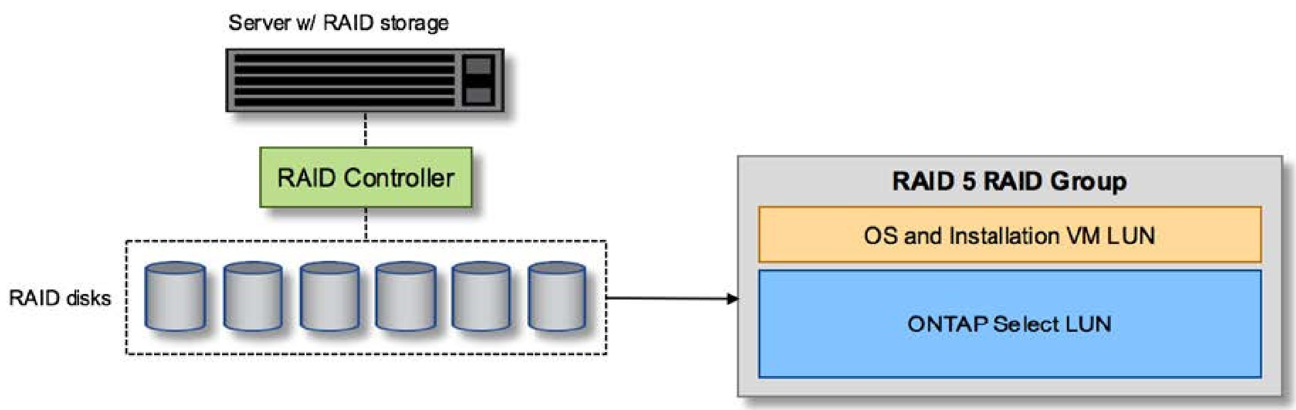
Local disks shared between ONTAP Select and OS

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

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

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

Server LUN configuration with only RAID-managed spindles



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

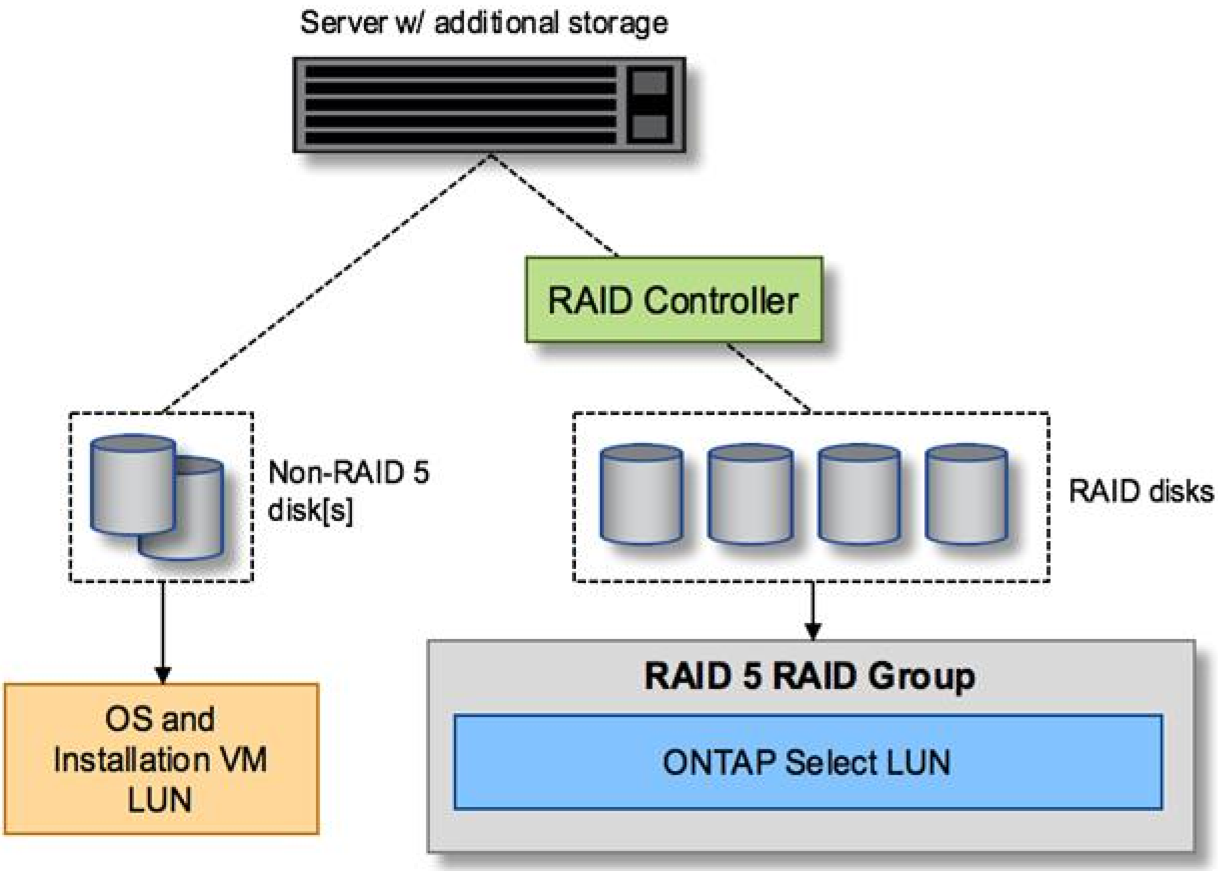
Local disks split between ONTAP Select and OS

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

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

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

Server LUN configuration on mixed RAID/non-RAID system



Multiple LUNs

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

VMware vSphere virtual machine file system limits

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

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

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

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

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

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

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

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

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

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

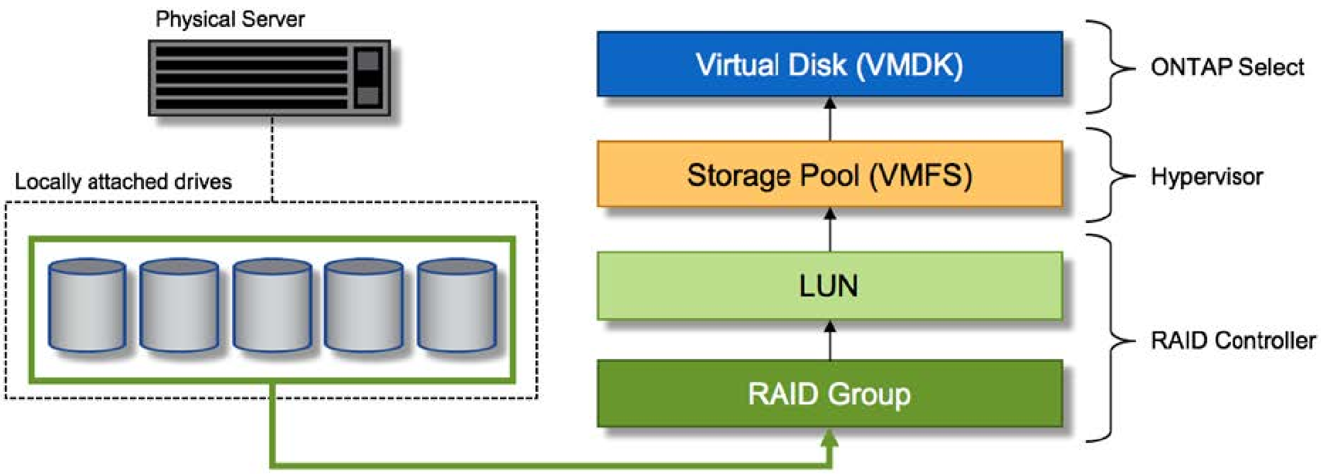
VMware does not support in-place upgrades from VMFS 5 to VMFS 6. Therefore, Storage vMotion is the only mechanism that allows any VM to transition from a VMFS 5 datastore to a VMFS 6 datastore. However, support for Storage vMotion with ONTAP Select and ONTAP Deploy was expanded to cover other scenarios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

ONTAP Select virtual disks

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

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

Virtual disk to physical disk mapping



Virtual disk provisioning

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

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

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

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

Virtualized NVRAM

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

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

NVRAM is placed on its own VMDK. Splitting the NVRAM in its own VMDK allows the ONTAP Select VM to use the vNVMe driver to communicate with its NVRAM VMDK. It also requires that the ONTAP Select VM uses hardware version 13, which is compatible with ESX 6.5 and newer.

Data path explained: NVRAM and RAID controller

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

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

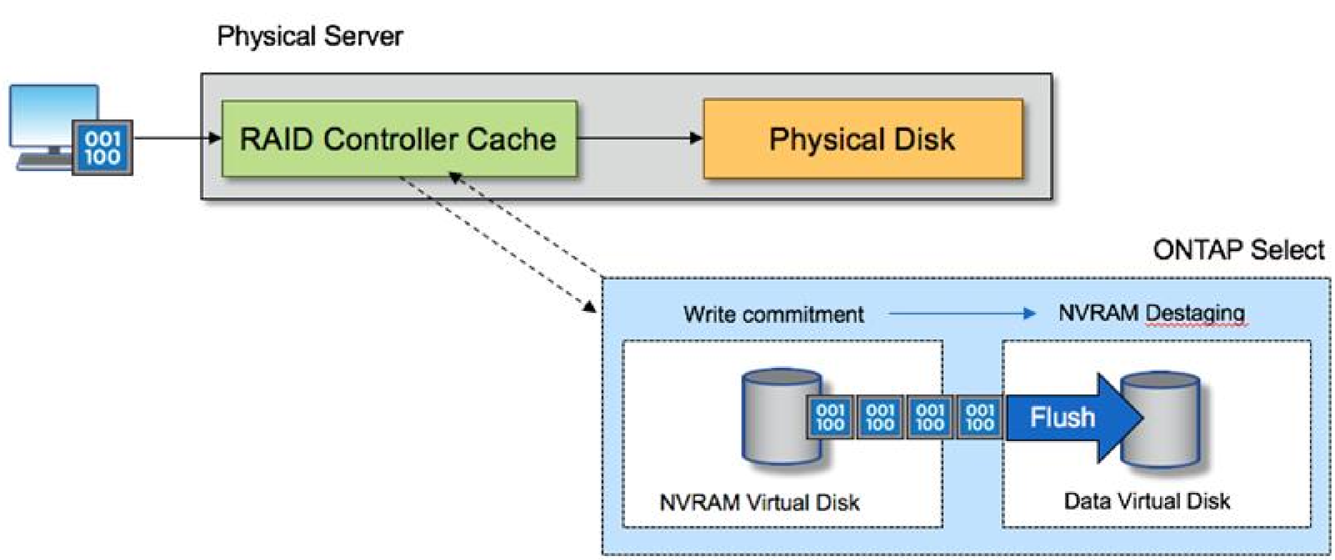
At this point, physically, the block resides in the RAID controller cache, waiting to be flushed to disk. Logically, the block resides in NVRAM waiting for destaging to the appropriate user data disks.

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

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

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

Incoming writes to ONTAP Select VM



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

Software RAID services for local attached storage

Software RAID is a RAID abstraction layer implemented within the ONTAP software stack. It provides the same functionality as the RAID layer within a traditional ONTAP platform such as FAS. The RAID layer performs drive parity calculations and provides protection against individual drive failures within an ONTAP Select node.

Independent of the hardware RAID configurations, ONTAP Select also provides a software RAID option. A hardware RAID controller might not be available or might be undesirable in certain environments, such as when ONTAP Select is deployed on a small form-factor commodity hardware. Software RAID expands the available deployment options to include such environments. To enable software RAID in your environment, here are some points to remember:

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

Notes

- The terms service disk and system disk are used interchangeably.
    - Service disks are the VMDKs that are used within the ONTAP Select VM to service various items such as clustering, booting, and so on.
    - Service disks are physically located on a single physical disk (collectively called the service/system physical disk) as seen from the host. That physical disk must contain a DAS datastore. ONTAP Deploy creates these service disks for the ONTAP Select VM during cluster deployment.
- It is not possible to further separate the ONTAP Select system disks across multiple datastores or across multiple physical drives.
- Hardware RAID is not deprecated.

Software RAID configuration for local attached storage

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

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

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

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

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

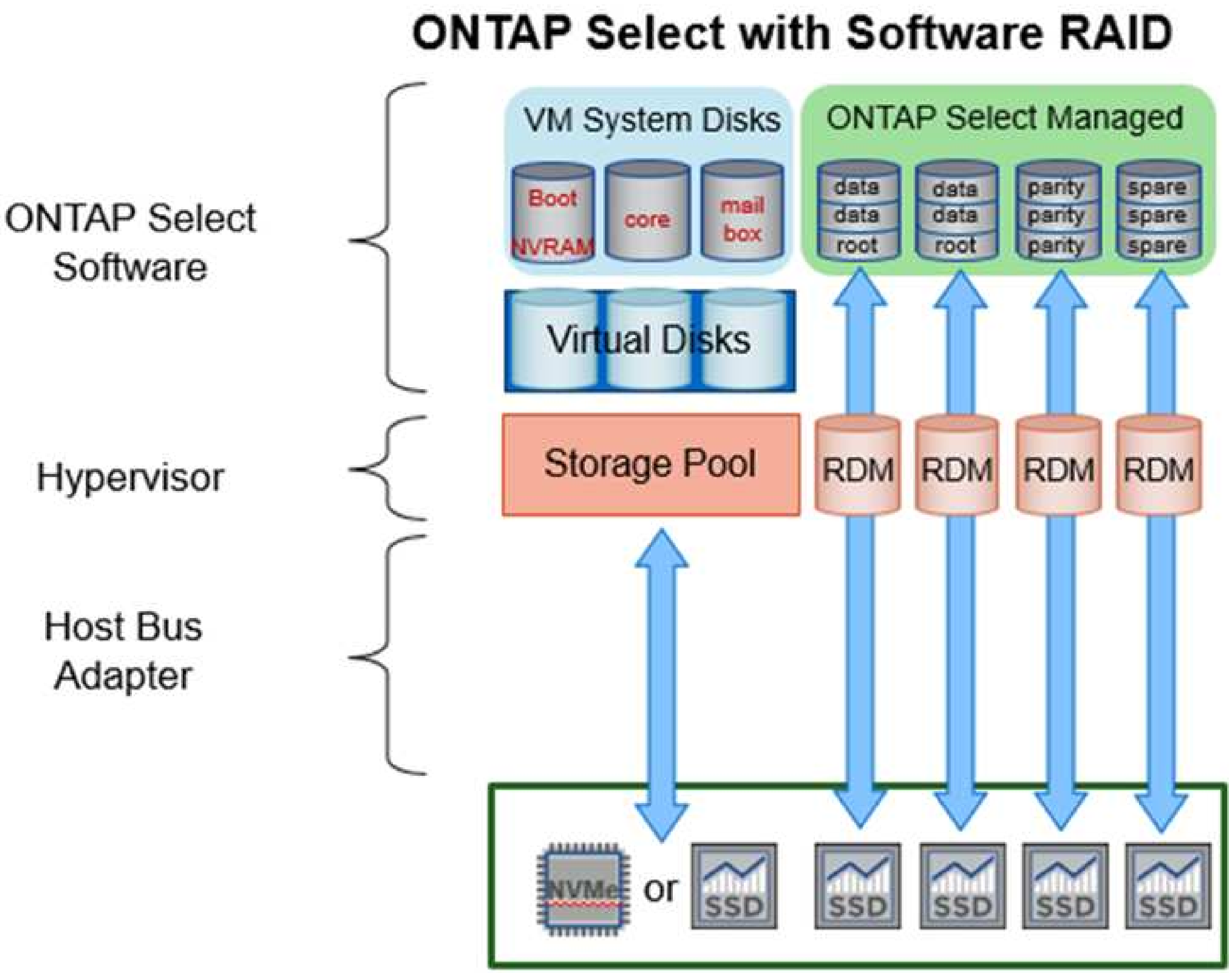
ONTAP Select virtual and physical disks

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

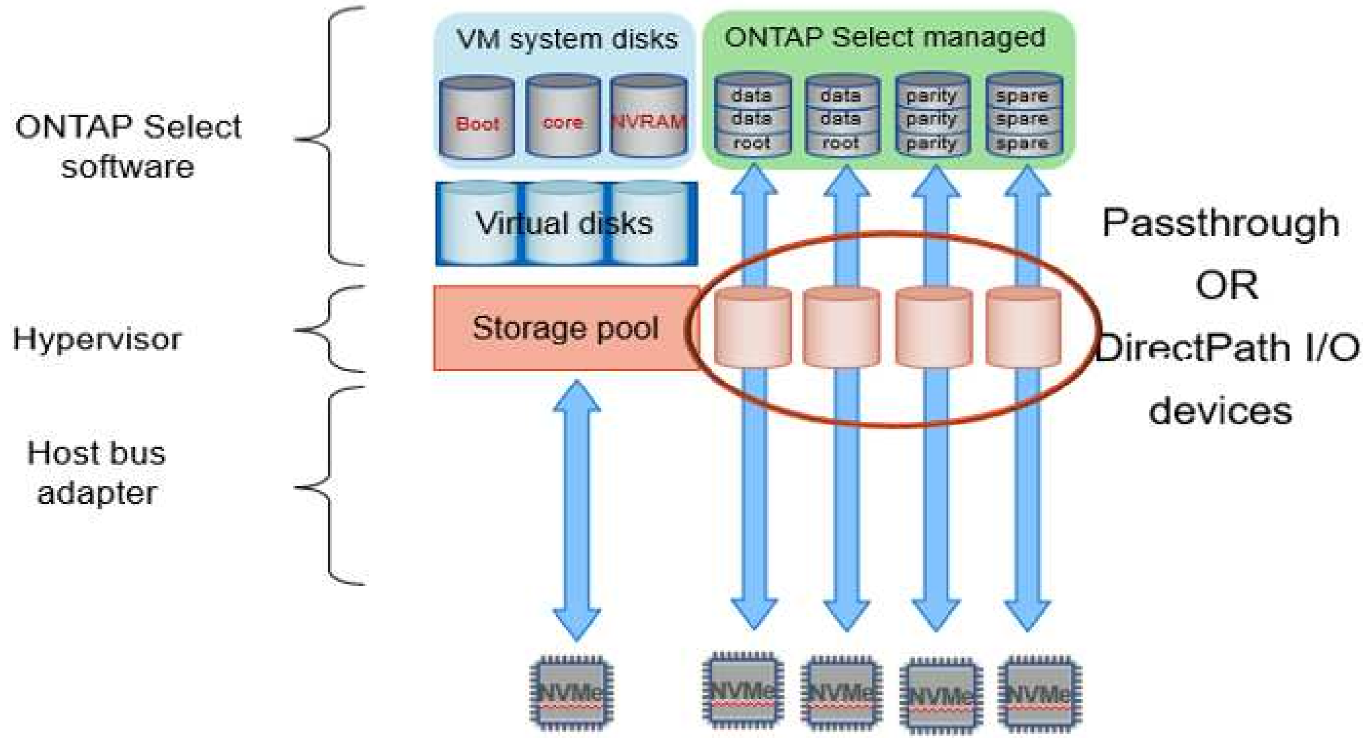
When using software RAID, ONTAP Deploy presents ONTAP Select with a set of virtual disks (VMDKs) and physical disks Raw Device Mappings [RDMs] for SSDs and passthrough or DirectPath IO devices for NVMes.

The following figures show this relationship in more detail, highlighting the difference between the virtualized disks used for the ONTAP Select VM internals and the physical disks used to store user data.

ONTAP Select software RAID: use of virtualized disks and RDMs



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



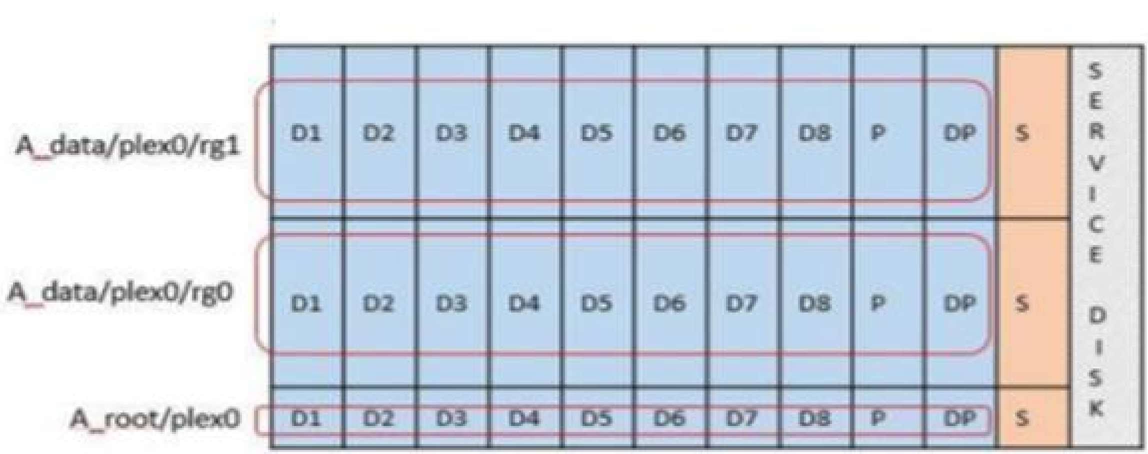
The system disks (VMDKs) reside in the same datastore and on the same physical disk. The virtual NVRAM disk requires a fast and durable media. Therefore, only NVMe and SSD-type datastores are supported. When using NVMe drives for data, the system disk should also be an NVMe device for performance reasons. A good candidate for the system disk in an all NVMe configuration is an INTEL Optane card.

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

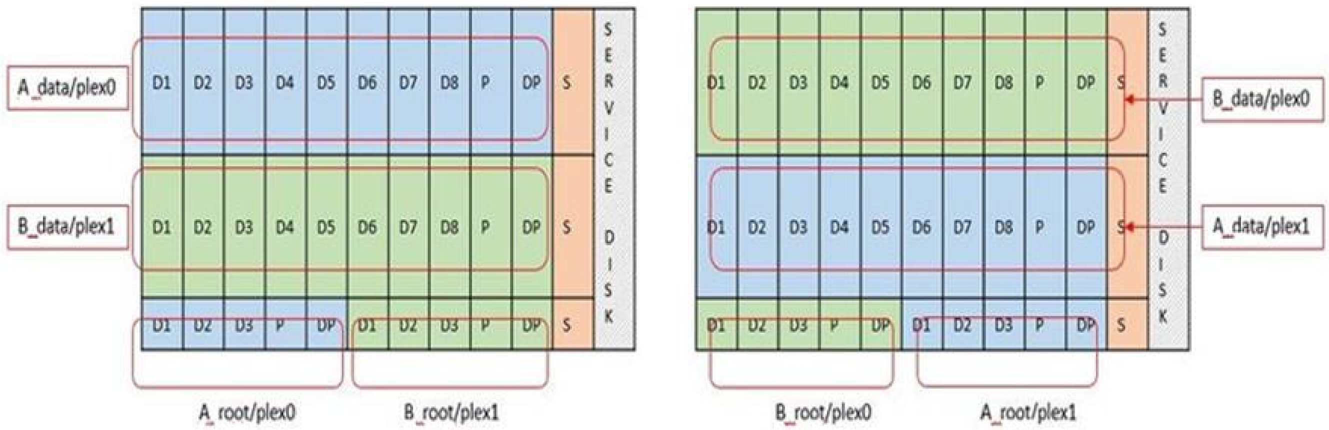
Each data disk is divided into three parts: a small root partition (stripe) and two equal-sized partitions to create two data disks seen within the ONTAP Select VM. Partitions use the Root Data Data (RD2) schema as shown in the following figures for a single node cluster and for a node in an HA pair.

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

RDD disk partitioning for single-node clusters



RDD disk partitioning for multinode clusters (HA pairs)



ONTAP software RAID supports the following RAID types: RAID 4, RAID-DP, and RAID-TEC. These are the same RAID constructs used by FAS and AFF platforms. For root provisioning ONTAP Select supports only RAID 4 and RAID-DP. When using RAID-TEC for the data aggregate, the overall protection is RAID-DP. ONTAP Select HA uses a shared-nothing architecture that replicates each node's configuration to the other node. That means each node must store its root partition and a copy of the its peer's root partition. Since a data disk has a single root partition, that the minimum number of data disks will vary depending on whether the ONTAP Select node is part of an HA pair or not.

For single node clusters, all data partitions are used to store local (active) data. For nodes that are part of an HA pair, one data partition is used to store local (active) data for that node and the second data partition is used to mirror active data from the HA peer.

Passthrough (DirectPath IO) devices vs. Raw Device Maps (RDMs)

VMware ESX does not currently support NVMe disks as Raw Device Maps. For ONTAP Select to take direct control of NVMe disks, the NVMe drives must be configured in ESX as passthrough devices. Please note that configuring an NVMe device as a passthrough devices requires support from the server BIOS and it is a disruptive process, requiring an ESX host reboot. Furthermore, the maximum number of passthrough devices per ESX host is 16. However, ONTAP Deploy limits this to 14. This limit of 14 NVMe devices per ONTAP Select node means that an all NVMe configuration will provide a very high IOPs density (IOPs/TB) at the expense of total capacity. Alternatively, if a high performance configuration with larger storage capacity is desired, the recommended configuration is a large ONTAP Select VM size, an INTEL Optane card for the system disk, and a nominal number of SSD drives for data storage.

> ⓘ To take full advantage of NVMe performance, consider the large ONTAP Select VM size.

There is an additional difference between passthrough devices and RDMs. RDMs can be mapped to a running VM. Passthrough devices require a VM reboot. This means that any NVMe drive replacement or capacity expansion (drive addition) procedure will require an ONTAP Select VM reboot. The drive replacement and capacity expansion (drive addition) operation is driven by a workflow in ONTAP Deploy. ONTAP Deploy manages the ONTAP Select reboot for single node clusters and failover / failback for HA pairs. However it is important to note the difference between working with SSD data drives (no ONTAP Select reboot / failovers are required) and working with NVMe data drives (ONTAP Select reboot / failover is required).

Physical and virtual disk provisioning

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

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

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

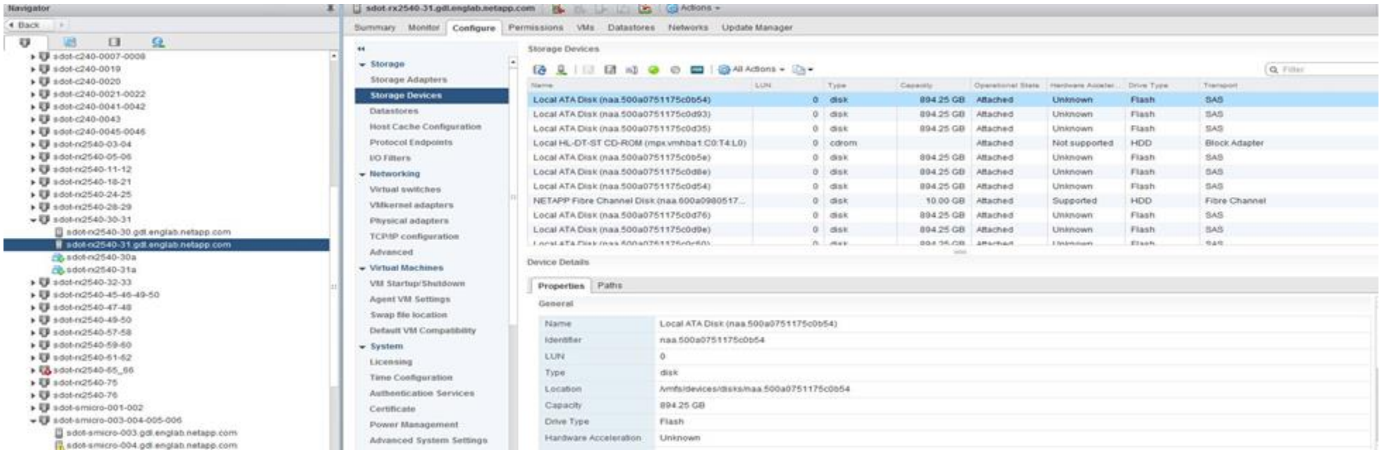
When adding capacity to an ONTAP Select VM using software RAID, the administrator must consider the physical drive size and the number of drives required. For details, see the section Increase storage capacity.

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

Match an ONTAP Select disk to the corresponding ESX disk

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

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



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

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

Multiple drive failures when using software RAID

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

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

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

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

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

```
C3111E67::> storage aggregate plex delete -aggregate aggr1 -plex plex1
Warning: Deleting plex "plex1" of mirrored aggregate "aggr1" in a non-
shared HA configuration will disable its synchronous mirror protection and
disable
         negotiated takeover of node "sti-rx2540-335a" when aggregate
"aggr1" is online.
Do you want to continue? {y|n}: y
```

```
[Job 78] Job succeeded: DONE

C3111E67::> storage aggregate mirror -aggregate aggr1
Info: Disks would be added to aggregate "aggr1" on node "sti-rx2540-335a"
in the following manner:
      Second Plex
        RAID Group rg0, 5 disks (advanced_zoned checksum, raid_dp)
                                                              Usable
Physical
          Position   Disk                       Type           Size
Size
          ---------- ------------------------- ---------- --------
--------
          shared     NET-3.2                    SSD               -
-
          shared     NET-3.3                    SSD               -
-
          shared     NET-3.4                    SSD         208.4GB
208.4GB
          shared     NET-3.5                    SSD         208.4GB
208.4GB
          shared     NET-3.12                   SSD         208.4GB
208.4GB

      Aggregate capacity available for volume use would be 526.1GB.
      625.2GB would be used from capacity license.
Do you want to continue? {y|n}: y

C3111E67::> storage aggregate show-status -aggregate aggr1
Owner Node: sti-rx2540-335a
 Aggregate: aggr1 (online, raid_dp, mirrored) (advanced_zoned checksums)
  Plex: /aggr1/plex0 (online, normal, active, pool0)
   RAID Group /aggr1/plex0/rg0 (normal, advanced_zoned checksums)
                                                                Usable
Physical
     Position Disk                        Pool Type     RPM     Size
Size Status
     -------- --------------------------- ---- ----- ------ --------
-------- ----------
     shared   NET-1.1                      0   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-1.2                      0   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-1.3                      0   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-1.10                     0   SSD        -  205.1GB
```

```
447.1GB (normal)
     shared   NET-1.11                    0   SSD        -  205.1GB
447.1GB (normal)
  Plex: /aggr1/plex3 (online, normal, active, pool1)
   RAID Group /aggr1/plex3/rg0 (normal, advanced_zoned checksums)
                                                              Usable
Physical
     Position Disk                        Pool Type     RPM     Size
Size Status
     -------- --------------------------- ---- ----- ------ --------
-------- ----------
     shared   NET-3.2                     1   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-3.3                     1   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-3.4                     1   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-3.5                     1   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-3.12                    1   SSD        -  205.1GB
447.1GB (normal)
10 entries were displayed..
```

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

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

The output for the last command should be empty.

Virtualized NVRAM

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

Commodity systems are not typically fitted with this type of equipment. Therefore, the functionality of the NVRAM card has been virtualized and placed into a partition on the ONTAP Select system boot disk. It is for this reason that placement of the system virtual disk of the instance is extremely important.

VSAN and external array configurations

Virtual NAS (vNAS) deployments support ONTAP Select clusters on virtual SAN (VSAN), some HCI products, and external array types of datastores. The underlying infrastructure of these configurations provide datastore resiliency.

The minimum requirement is that the underlying configuration is supported by VMware and should be listed on the respective VMware HCLs.

vNAS architecture

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

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

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

vNAS architectures versus local DAS with hardware RAID controllers

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

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

vNAS NVRAM

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

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

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

The SIDL feature is not compatible with all ONTAP Select storage efficiency features. The SIDL feature can be disabled at the aggregate level using the following command:

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

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

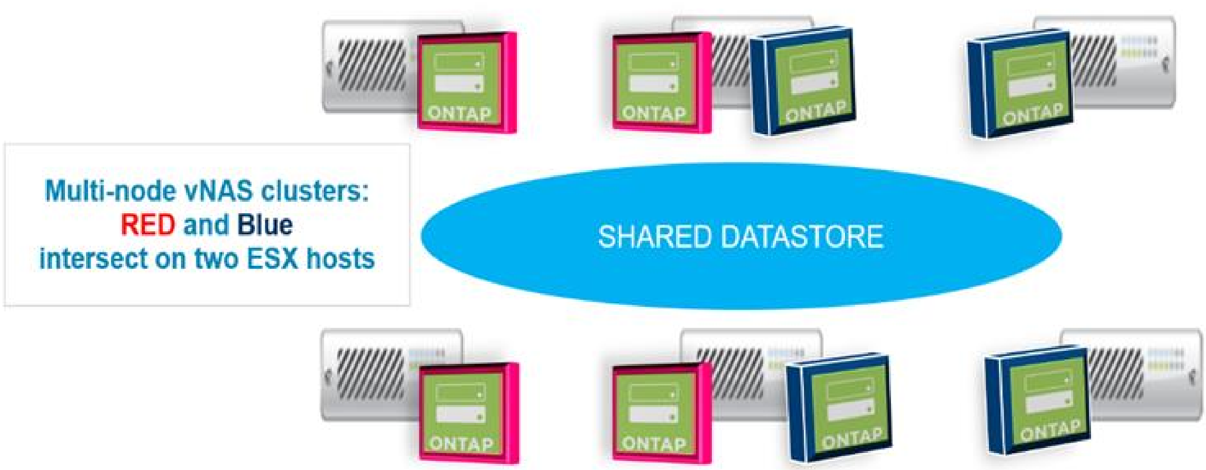
```
volume efficiency stop -all true -vserver * -volume * (all volumes in the
affected aggregate)
```

Collocate ONTAP Select Nodes When Using vNAS on ESXi

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

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

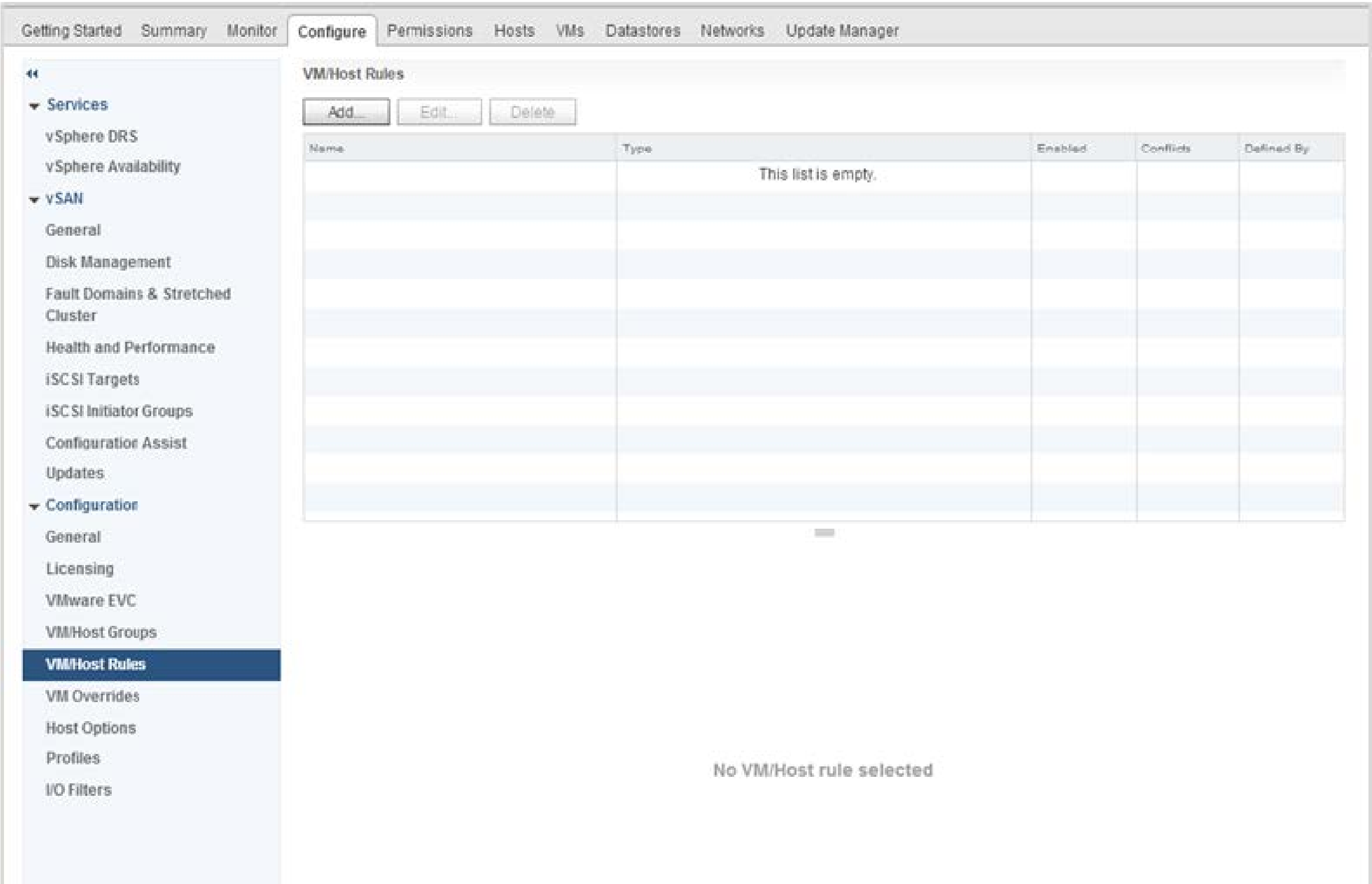
Initial deployment of multinode VNAS clusters

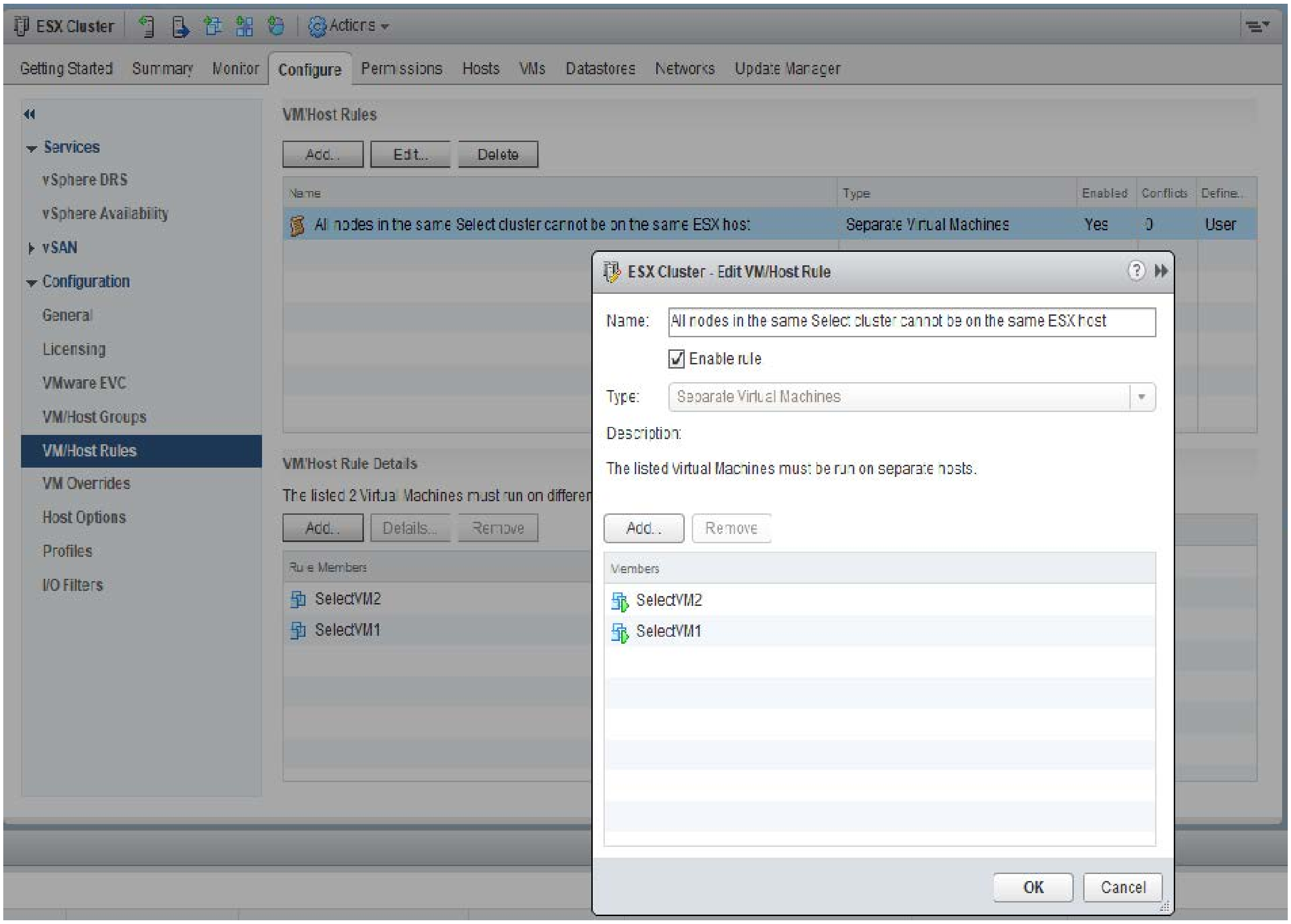


After deployment, the ONTAP Select nodes can be migrated between hosts. This could result in nonoptimal and unsupported configurations for which two or more ONTAP Select nodes from the same cluster share the same underlying host. NetApp recommends the manual creation of VM anti-affinity rules so that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

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

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





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

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

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

Increase storage capacity

ONTAP Deploy can be used to add and license additional storage for each node in an ONTAP Select cluster.

The storage-add functionality in ONTAP Deploy is the only way to increase the storage under management, and directly modifying the ONTAP Select VM is not supported. The following figure shows the “+” icon that initiates the storage-add wizard.

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

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

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

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

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

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

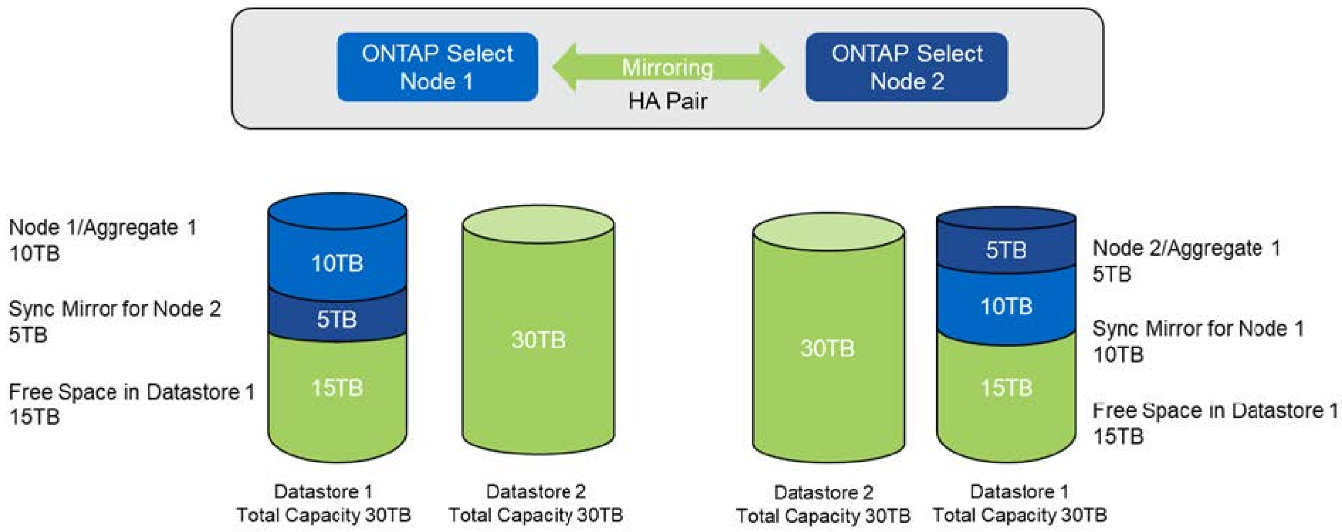
There is an additional consideration with regard to performance. The data on node 1 is synchronously replicated to node 2. Therefore, the performance of the new space (datastore) on node 1 must match the performance of the new space (datastore) on node 2. In other words, adding space on both nodes, but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

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

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

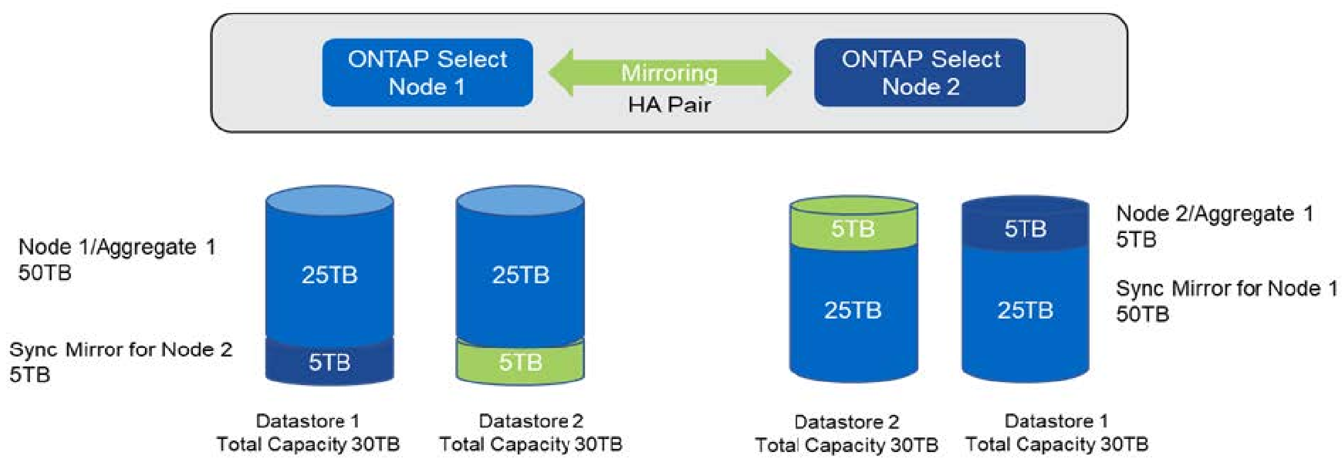
The following figure shows the results of a single storage-add operation for node 1. ONTAP Select still uses an equal amount of storage (15TB) on each node. However, node 1 has more active storage (10TB) than node 2 (5TB). Both nodes are fully protected as each node hosts a copy of the other node's data.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

Capacity distribution: allocation and free space after a single storage-add operation



Two additional storage-add operations on node 1 consume the rest of datastore 1 and a part of datastore 2 (using the capacity cap). The first storage-add operation consumes the 15TB of free space left in datastore 1. The following figure 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.

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



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

Increase capacity for ONTAP Select with Software RAID

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

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

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

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

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

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

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

Storage efficiency support

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

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

An AFF-like personality is automatically enabled on new installations as long as you have DAS storage with SSD drives and a premium license.

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

- Inline zero pattern detection

- Volume inline deduplication
- Volume background deduplication
- Adaptive inline compression
- Inline data compaction
- Aggregate inline deduplication
- Aggregate background deduplication

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

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

For ONTAP Select upgrades from 9.6 and later, you must install ONTAP Select on DAS SSD storage with a premium license. In addition, you must check the **Enable Storage Efficiencies** check box during the initial cluster installation with ONTAP Deploy. Enabling an AFF-like personality post-ONTAP upgrade when prior conditions have not been met requires the manual creation of a boot argument and a node reboot. Contact technical support for further details.

ONTAP Select storage efficiency configurations

The following table summarizes the various storage efficiency options available, enabled by default, or not enabled by default but recommended, depending on the media type and software license.

ONTAP Select features	DAS SSD (premium or premium XL[1])	DAS HDD (all licenses)	vNAS (all licenses)
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)	Not available	Not supported

ONTAP Select features	DAS SSD (premium or premium XL[1])	DAS HDD (all licenses)	vNAS (all licenses)
32K inline compression (secondary compression)	Yes Enabled by user on a per volume basis.	Yes Enabled by user on a per-volume basis	Not supported
8K inline compression (adaptive compression)	Yes (default)	Yes Enabled by user on a per volume basis	Not supported
Background compression	Not supported	Yes Enabled by user on a per volume basis	Yes Enabled by user on a per-volume basis
Compression scanner	Yes	Yes	Yes Enabled by user on a per-volume basis
Inline data compaction	Yes (default)	Yes Enabled by user on a per volume basis	Not supported
Compaction scanner	Yes	Yes	Not supported
Aggregate inline deduplication	Yes (default)	N/A	Not supported
Volume background deduplication	Yes (default)	Yes Enabled by user on a per volume basis	Yes Enabled by user on a per-volume basis
Aggregate background deduplication	Yes (default)	N/A	Not supported

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

Notes on upgrade behavior for DAS SSD configurations

After upgrading to ONTAP Select 9.6 or later, wait for the `system node upgrade-revert show` command to indicate that the upgrade has completed before verifying the storage efficiency values for existing volumes.

On a system upgraded to ONTAP Select 9.6 or later, 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. Existing volumes that undergo the ONTAP Select code upgrade have most of the same storage efficiency policies as a newly created volume with some variations:

Scenario 1

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

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

Scenario 2

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

- Volumes with `space guarantee = volume` do not see any difference after upgrade.

- Volumes with `space guarantee = none` have aggregate background deduplication turned on.
- Volumes with `storage policy inline-only` have their policy set to auto.
- Volumes with user defined storage efficiency policies have no change in policy, with the exception of volumes with `space guarantee = none`. These volumes have aggregate background deduplication enabled.