Hardware RAID services for local attached storage

ONTAP Select

Barb Einarsen
November 21, 2019

Table of Contents

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.

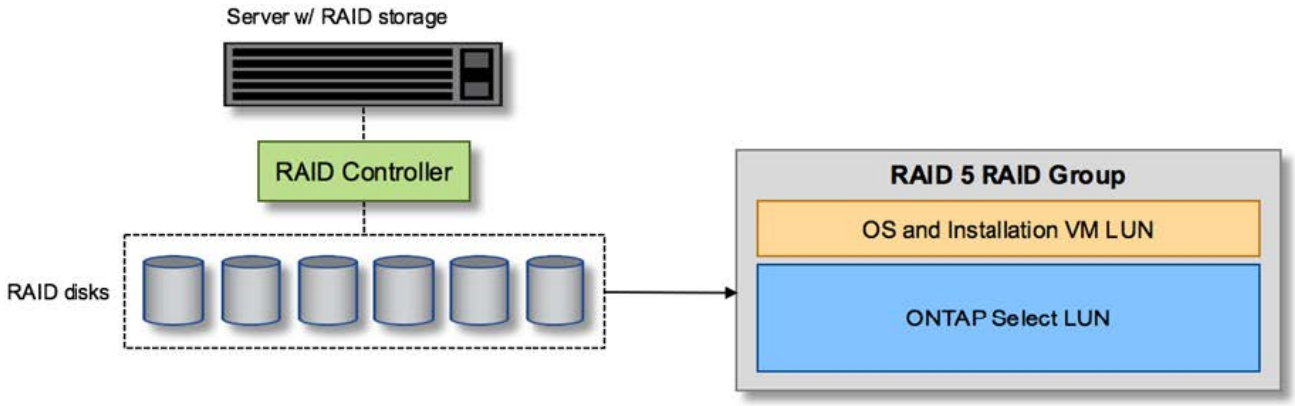
Local disks shared between ONTAP Select and OS

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.

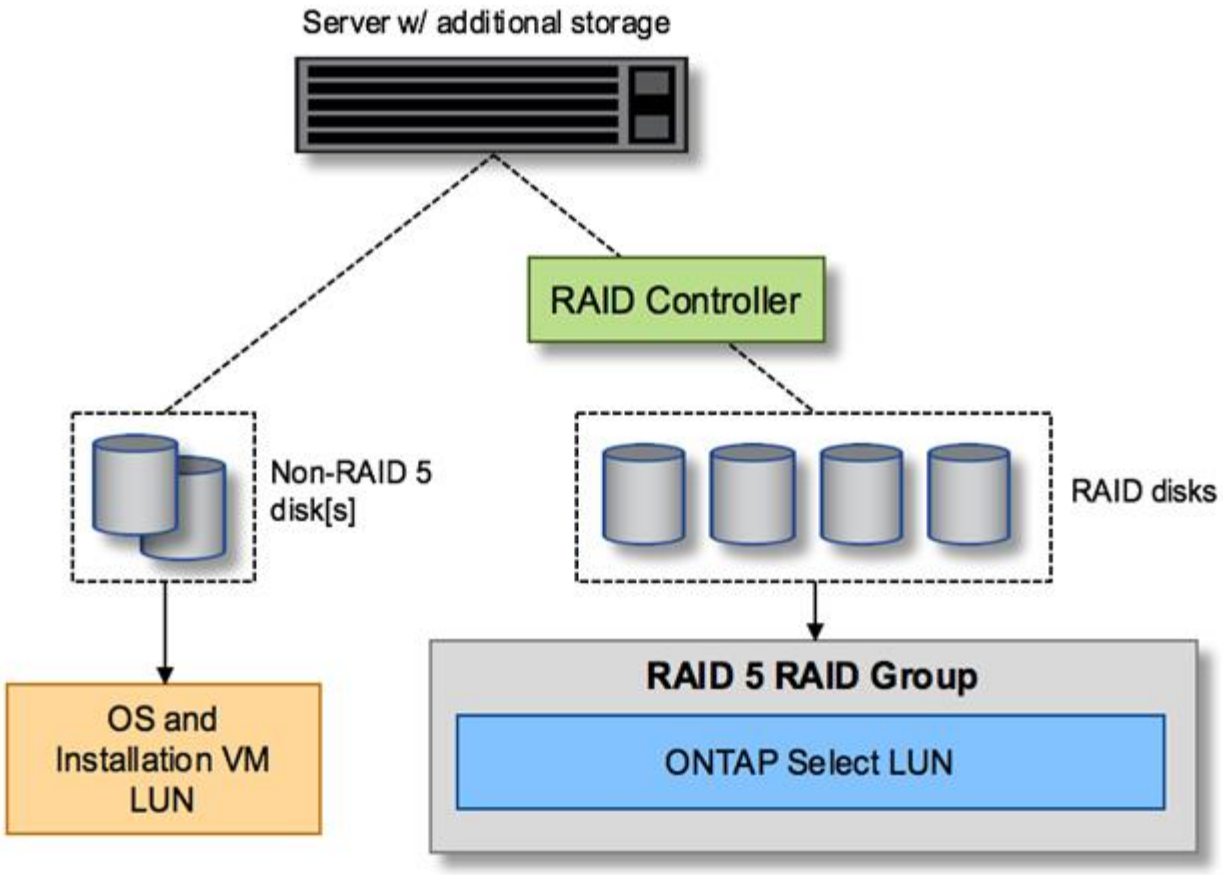
Local disks split between ONTAP Select and OS

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 Increasing 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.

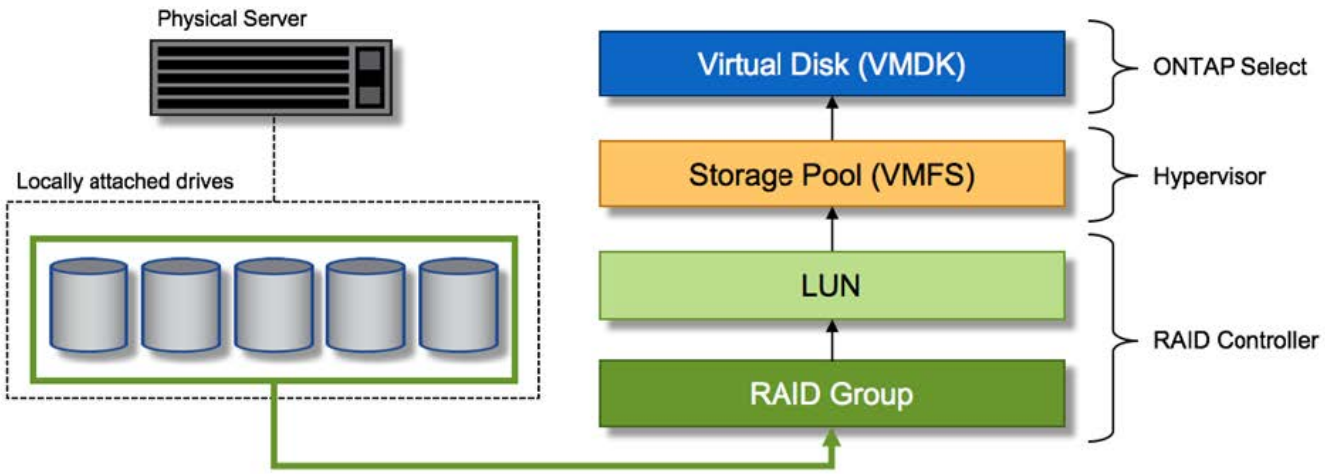
ONTAP Select virtual disks

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 Increasing 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.

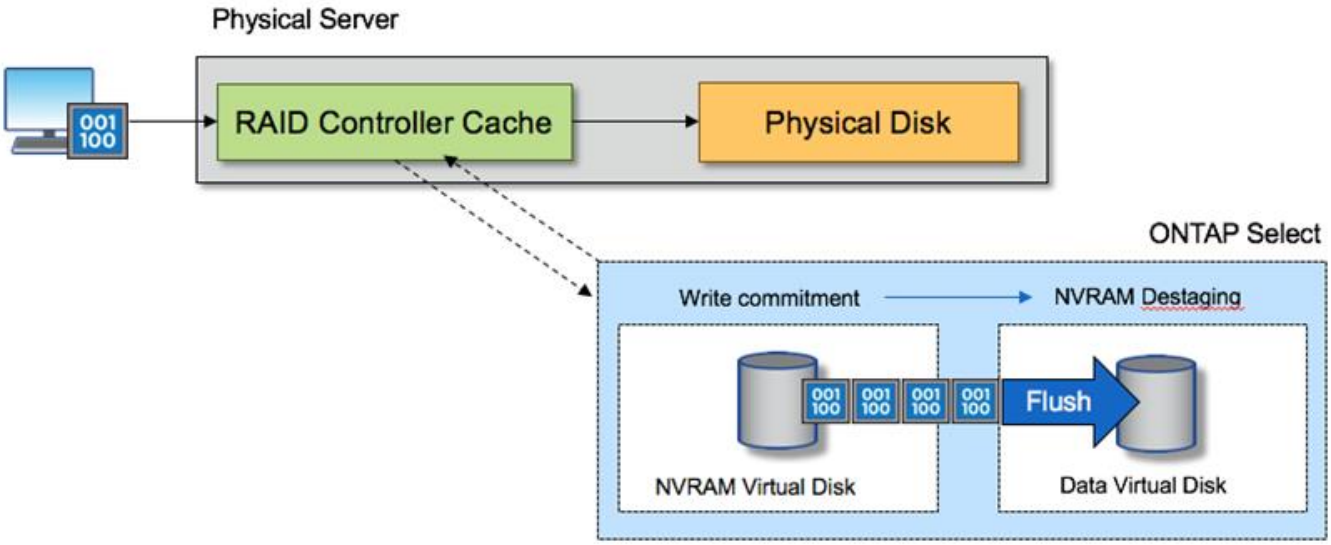
Because changed blocks are automatically stored within the RAID controller's local cache, incoming writes to the NVRAM partition are automatically cached and periodically flushed to physical storage media. This should not be confused with the periodic flushing of NVRAM contents back to ONTAP data disks.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



Note that 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.