vSAN Space Efficiency Technologies

Space efficiency capabilities with vSAN using VMware Cloud Foundation 9.0

September 16, 2025

vmware by Broadcom

Table of Contents

Introduction 3

Opportunistic Space Efficiency Features 3

- Compression (ESA) 3
- Global Deduplication (ESA) 4
- Deduplication & Compression (OSA) 6
- Compression-Only (OSA) 9
- Deduplication & Compression or Compression-Only (OSA) – Which one is Right for you? 11
- Thin Provisioning 13
- TRIM/UNMAP Space Reclamation 13

Deterministic Space Efficiency Features 19

- Data Placement Schemes and Erasure Code Concepts 20
- RAID-5 Erasure Coding (ESA) 21
- RAID-6 Erasure Coding (ESA) 22
- RAID-5 Erasure Coding (OSA) 22
- RAID-6 Erasure Coding (OSA) 23
- Erasure Coding Recommendations (OSA) 24
- RAID-1, RAID-5, or RAID-6 – What to Choose? 25

Summary 26

- Additional Resources 26
- About the Author 26

Introduction

Space efficiency technologies in enterprise storage play an important role improving value and decreasing costs. VMware vSAN has several technologies in place to help improve storage efficiency.

Space efficiency techniques can be categorized into the following:

- **Opportunistic.** These space efficiency techniques are dependent on conditions of the data, and not guaranteed to return a predetermined level of savings. vSAN offers several types of opportunistic space efficiency features such as Deduplication & Compression (in vSAN OSA), Compression-only, TRIM/UNMAP space reclamation, and thin provisioning.
- **Deterministic.** These space efficiency techniques can be relied upon to deliver a guaranteed level of capacity savings. vSAN offers deterministic space efficiency capabilities through data placement schemes that are optimized for storing data in a resilient but efficient manner. This includes RAID-5/6 erasure codes.

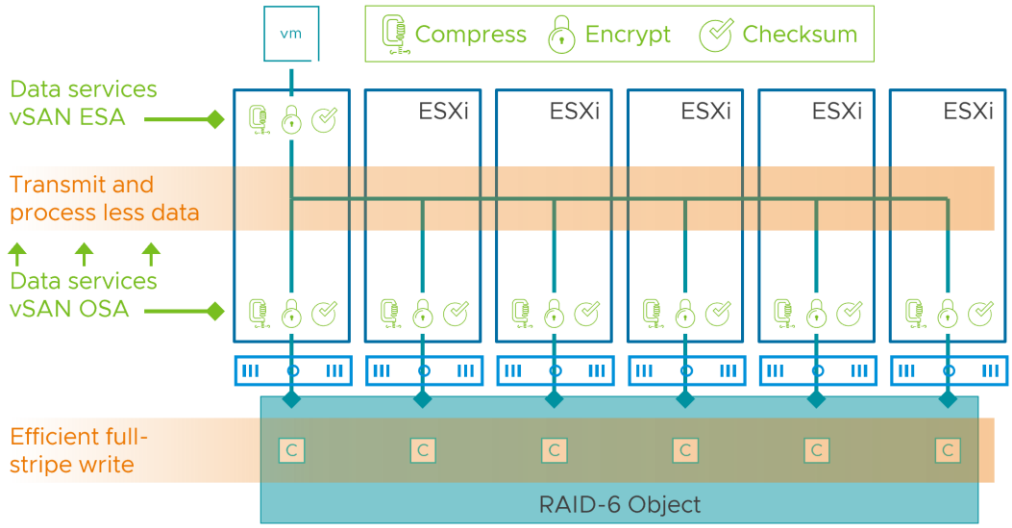
In vSAN, opportunistic and deterministic space efficiency features can be used independently or together. The specific trade-offs when doing so will be discussed in this document. **The tradeoffs will vary depending on if the cluster is using the much newer vSAN Express Storage Architecture (ESA), or the Original Storage Architecture (OSA).** The differences will be noted throughout this document.

Opportunistic Space Efficiency Features

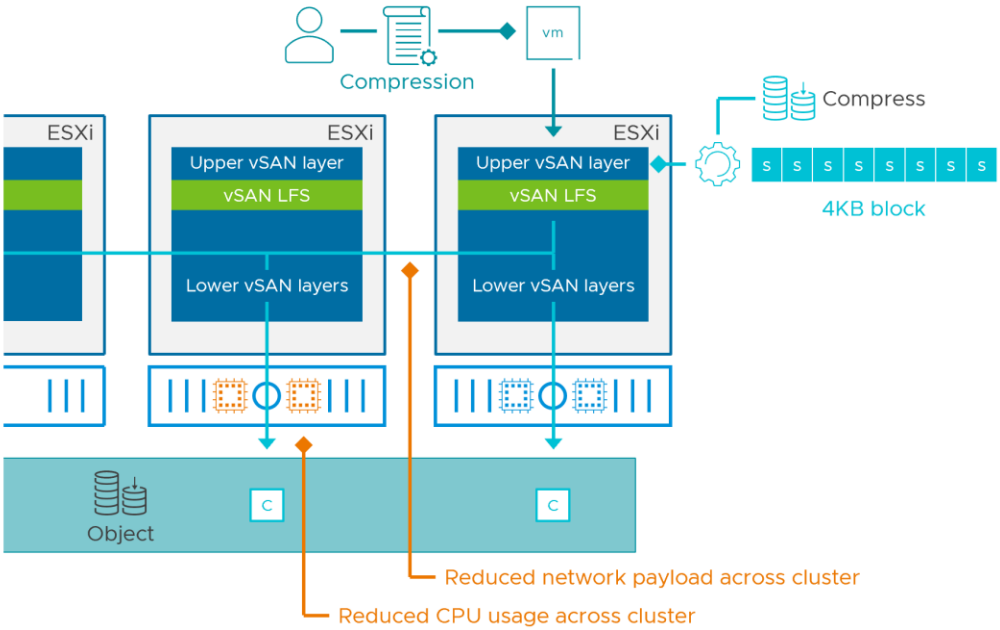
Opportunistic space efficiency techniques do just as the name implies. If a given set of conditions is ideal for saving capacity, it will do so based on that given set of conditions. The degree to which the savings will occur will be highly dependent on the technology in place, the workloads, and even the host hardware configuration.

Compression (ESA)

vSAN 8 introduced a new optional architecture, known as the vSAN Express Storage Architecture, or ESA. Compression is implemented quite differently in the ESA versus the OSA described elsewhere in this document. In the ESA, data compression (and other services such as encryption, and checksum processing) have been moved to the top of the storage stack. When a guest VM issues a write operation, it will compress the data the moment it enters the top of the vSAN stack. Unlike the OSA, this is performed once, and not only eliminates the need to compress the data on the other hosts holding the object, but will reduce the amount of data transmitted across the network. This reduces CPU and network resources across the cluster. This can even improve performance and reduce resources when using the vSAN ESA in a stretched cluster topology.



The compression mechanisms in the ESA evaluates and compresses data differently than in the OSA. In the ESA, each incoming 4KB block is evaluated on a 512 Byte sector size. Relative to the OSA, this smaller size of compression means that data that can be compressed at finer levels of granularity if the data written is actually compressible. With 8 sectors to a 4KB block, this means that the 4KB block can be reduced down in increments of 512 bytes, depending on how compressible the 4KB block is. For example, a 4KB block could be compressed down to 7/8ths its original size is if it is not very compressible, or all the way down to 1/8th its original size, if it is highly compressible.



While this can offer much better compression, it is entirely dependent on the customer's data, and how compressible it may or may not be. For example, an already compressed image or video file format will not be compressed beyond the native file format. As a result, we recommend taking a pretty conservative approach in estimating what your compression rates may be for real world data.

Data compression is controllable using storage policies in the vSAN ESA. It is on by default, and we recommend that it remains on, unless there is a specific application that performs its own compression. Compression in the ESA can deliver unique efficiencies in a stretched cluster environment, where data sent across the ISL is already compressed, thereby increasing the effective potential bandwidth of the ISL. See the post, "Using the vSAN ESA in a Stretched Cluster Environment" for more information.

For more information compression using the ESA, see the post: "vSAN 8 Compression - Express Storage Architecture" and "An Introduction to the vSAN Express Storage Architecture."

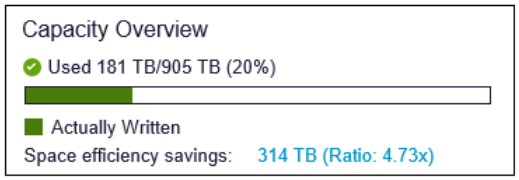
Global Deduplication (ESA)

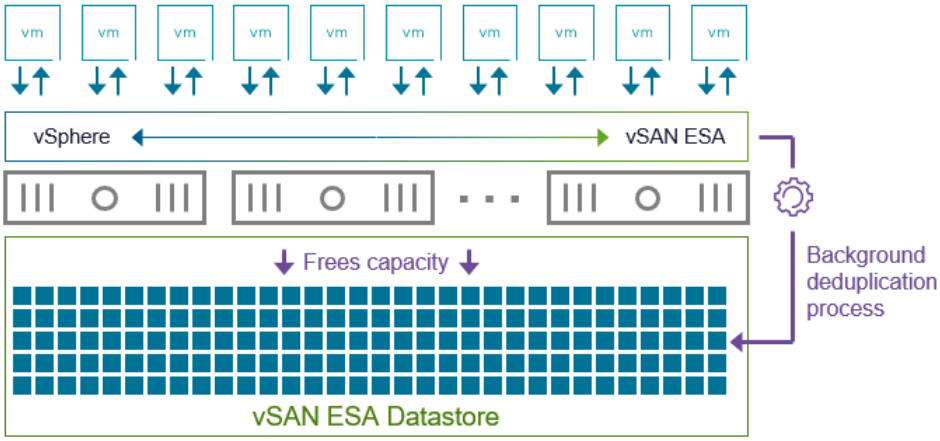
vSAN ESA in VCF 9.0 introduces global deduplication. Built specifically for the Express Storage Architecture, it can deliver space efficiency and performance at levels that are not possible with the original storage architecture. It reduces or eliminates all the technical considerations and limitations found with deduplication in vSAN OSA, including:

- **Larger deduplication domain.** The deduplication domain in vSAN ESA is the entire ESA cluster. This dramatically improves the potential of duplicate blocks of data to be found and deduplicated, improving the effective data reduction rates. This deduplication domain will automatically grow as the host count of the cluster grows. vSAN OSA used a much smaller deduplication domain of a discrete disk group, which limited its effectiveness.
- **Adaptively throttled post-processing.** Performing deduplication outside of the write path helps ensure that there is no interference with ingesting write operations from guest VMs. This helps ensure that guest VM write latency

remains low. Since vSAN is fully aware of the amount of resources used at any point in time, advanced algorithms will adaptively throttle the amount of resources used for the purposes of deduplication. During quiet times, the cluster will consume more resources for deduplication than it will during busier times in the day.

- **Minimal performance impact.** The architecture used in ESA's deduplication virtually eliminates all of the performance challenges associated with deduplication in vSAN OSA. Deduplication in vSAN ESA is appropriate for all workload types.





When compared to deduplication in vSAN OSA, the result is much better data reduction, virtually no impact on performance, and an automated, intelligent way that data storage is reduced. When paired with compression in vSAN ESA, the data reduction savings may meet or exceed the deduplication rates observed with the same data on a traditional storage array. For more information, see the post: "Save Costs and Scale Efficiently with vSAN Deduplication in VMware Cloud Foundation 9.0."

Note that as of vSAN 9.0 P01, vSAN encryption services is not currently compatible with the new global deduplication feature.

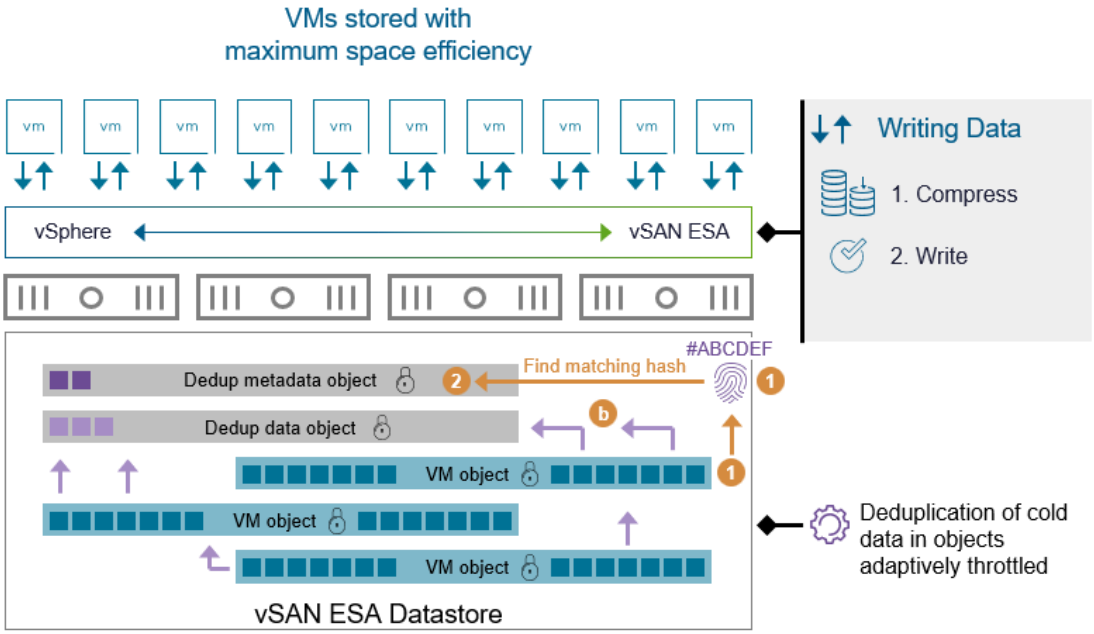
Deduplication Post-Processing

All deduplication activities occur after data has been persisted to storage. This type of post-processing helps minimize the impact on guest VM performance. Since ESA's metadata mapping understands recently written data versus cold data, it will always prioritize the deduplication of cold data first to minimize the deduplication of frequently written data that may change rapidly in a short period of time. vSAN's sophisticated set algorithms will dynamically throttle the amount of resources used for the purpose of deduplication. vSAN will categorize any data movement related to deduplication as resynchronization traffic, where its adaptive resync and adaptive network traffic shaping will manage resources properly to ensure VM workloads maintain priority. There are no administrative tuning or operations to worry about.

The deduplication process will occur in the following manner.

1. vSAN will read a discrete 4KB block and generate a secure cryptographic hash to be stored in the dedup metadata object.
2. vSAN will look for a matching hash entry in the dedup metadata.
    a. If a match has been found with data in the dedup data object, it will update the block with a metadata pointer and reclaim the space

b. If a match has been found with no data in the dedup data object, it will move both the current data and the original data (using the back-pointer discussed below) to the dedup object, update the blocks with a metadata pointer and reclaim the space.
c. If no match has been found, it will leave the data as-is. Hash entries will be created in the metadata object with a back-pointer where the data resides so that if and when a duplicate entry is identified, it can be deduplicated as described above.



Both the data and the metadata used in vSAN ESA's deduplication process reside in unique objects that are instantiated at the time that deduplication is enabled. Depending on the size and other conditions of the cluster, a cluster will have one or more of the following:

- **Deduplication metadata object.** Maintains a hash entry for every chunk in the system. It is the hash entry that helps identify other instances that contain the same data.
- **Deduplication data object.** This will be comprised of all of the chunks of data that have been deduplicated. Deduplicated data are moved to these objects to prevent referential hot spots on VMs.

Availability of the Feature

Initially, global deduplication for vSAN ESA will have additional limitations. For more information, see the post: "Global Deduplication in vSAN ESA for VCF 9.0"

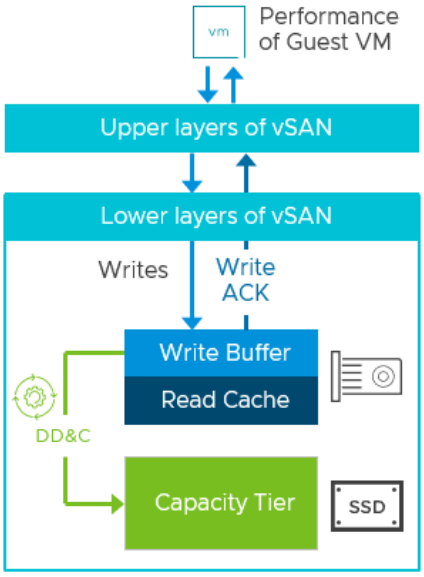
Deduplication & Compression (OSA)

When using the vSAN Original Storage Architecture (OSA), deduplication and compression (DD&C) in vSAN is enabled at the cluster level, as a single space efficiency feature. The process occurs as the data is destaged to the capacity tier - well after the write acknowledgments have been sent back to the VM. Minimizing any form of data manipulation until after the acknowledgment has been sent help keeps write latency seen by the guest VM low.

Recommendation: Use a cluster running vSAN ESA for optimal space efficiency and performance. It is simpler, faster, and more effective that deduplication in vSAN OSA. The information shared below about deduplication in OSA is for existing customers running legacy hardware that have not moved to vSAN ESA yet.

In vSAN OSA, as data is destaged, the deduplication process will look for opportunities to deduplicate the 4KB blocks of data it finds within a disk group: vSAN's deduplication domain. This task is followed by the compression process. If the

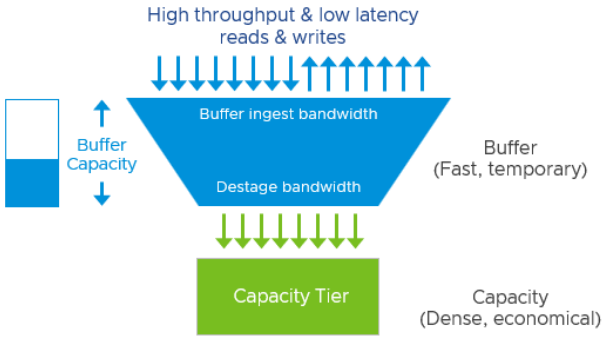
4KB block can be compressed by 50% or more, it will do so. Otherwise, it will leave as-is, and continue destaging the data to the capacity tier.



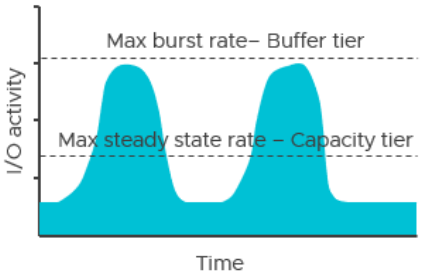
Implementing DD&C in this manner prevents the performance penalties found with inline systems that perform the deduplication prior to sending the write acknowledgment back to the guest. It also avoids the challenges of deduplicating data already at rest. While the DD&C process occurs after the write acknowledgment is sent to the guest VM, enabling it in vSAN can impact performance under certain circumstances, which will be discussed below.

Two-Tier Storage System Basics

The vSAN OSA is a two-tier distributed storage system. Incoming data is written to a write buffer with the write acknowledgment sent immediately back to the guest for optimal performance, and funneled down to the capacity tier at a time and frequency determined by vSAN. This architecture provides a higher level of storage performance while keeping the cost per gigabyte/terabyte of capacity reasonable.



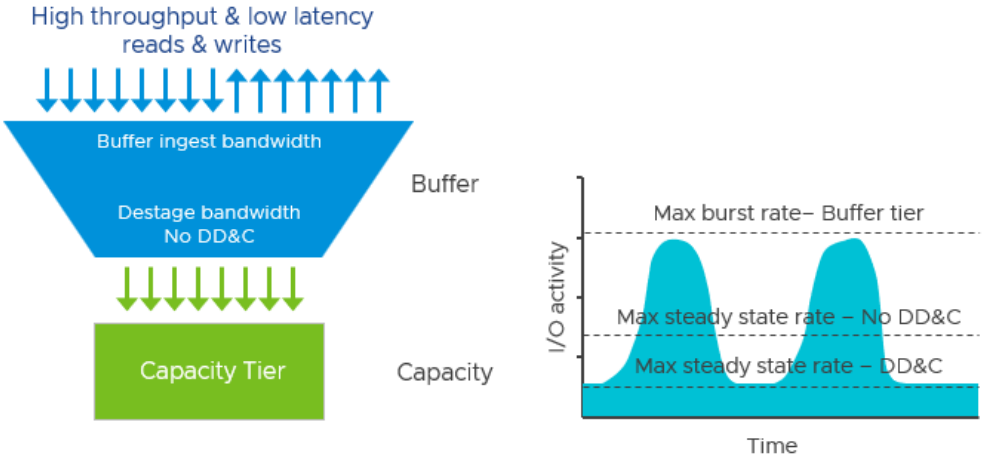
A two-tier system like vSAN has two theoretical performance maximums. A burst rate, representing the capabilities of buffer tier, and the steady-state rate, representing the capabilities of the capacity tier. The underlying hardware at each tier has a tremendous influence on the performance capabilities of each tier, but software settings, applications, and workloads can impact performance as well.



The performance maximums of your vSAN hosts will be somewhere in between the maximum burst rate, and the maximum steady-state rate. Synthetic testing for long periods using HCIBench will stress the environment enough to show these approximate values when looking at the time-based graphs. Production workloads may hit these maximums in an undersized design.

Deduplication in OSA and it's Potential Impact on Performance

Deduplication and compression require effort: Computational effort and the use of RAM and additional I/O that stems from it. This is true regardless of how it is implemented. It just depends on when, where, and how it occurs. In vSAN, since this effort occurs once the data in the write buffer begins to destage, the task reduces the effective destaging throughput to the capacity tier. This would lower the maximum steady-state rate that the cluster could provide. **In other words, an OSA cluster with DD&C enabled may have similar performance to a cluster with DD&C deactivated that use much lower performing capacity tier devices.**



Assuming all other variables remained the same, lowering the maximum steady-state rate of the capacity tier would demonstrate the following behaviors

- The write buffer may fill up more quickly because delta in performance between the two tiers has been increased through slowing down the performance of the capacity tier.
- The write buffer will destage more slowly because of the reduced destaging performance.
- The write acknowledgment time (write latency) of the guest VMs may be affected **if destaging has begun**. The degree of impact depends on several factors, including the destage rate capable by the capacity tier. This scenario is most common if the aggregate working set of data far surpasses the capacity of the buffer tier, or a fast duty cycle, which places more significant demands on the capacity tier.
- The write acknowledgment time (write latency) of the guest VM will be unaffected **IF the buffer has not reached any destaging thresholds.** This would be common with a small aggregate working set of data that fits well within the buffer tier and not a lot of pressure on the buffer to destage.

The elevator algorithms in the vSAN OSA detect a variety of conditions that will help determine if, when, and how much destaging should occur. It gradually introduces the destaging of data, and does not go as fast as it can, but only as fast as it needs to. This helps keep hot data in the buffer for subsequent overwrites, which reduces unnecessary destaging activity and potential impacts on performance.

Customization Options

vSAN's architecture gives customers numerous options to tailor their clusters to meet their requirements. Design and sizing correctly means clearly understanding the requirements and priorities. For example, in a given cluster, is capacity the higher priority, or is it performance? Are those priorities reflected in the existing hardware and software settings? Wanting the highest level of performance but choosing the lowest grade componentry while also using space efficiency techniques is a conflict in the objective. Space efficiency can be a way to achieve capacity requirements but comes with a cost.

Once the priorities are established, adjustments can be made to accommodate the need of the environment. This could include:

- **Faster devices in the capacity tier.** If you enabled DD&C and notice higher than expected VM latency, consider faster capacity devices at the capacity tier. This can help counteract the reduced performance of the capacity tier when DD&C is enabled. Insufficient performing devices at the capacity tier is one of the most common reasons for performance issues.
- **Use more disk groups.** This will add more buffer capacity, increasing the capacity for hot working set data, and reduce the urgency at which data is destaged. Two disk groups would be the minimum, with three disk groups the preferred configuration.
- **Upgrade to vSAN 8 to enable large write buffer support.** The OSA in vSAN 8 has increased the maximum logical buffer limit from 600GB to 1.6TB per disk group. See the post: "Increased Write Buffer Capacity for the vSAN 8 Original Storage Architecture" for more information:
- **Look at newer high-density storage devices for the capacity tier to meet your capacity requirements.** Assuming they meet your performance requirements, these new densities may offset your need to use DD&C.
- **Run the very latest version of vSAN.** Recent editions of vSAN have focused on improving performance for clusters running DD&C: Improving the consistency of latency to the VM, and increasing the destage rate through software optimizations.
- **Enable only in select clusters.** Only enable DD&C in clusters that hardware that can provide sufficient performance to support the needs of the workloads. Optionally, space-efficient storage policies like RAID-5/6 could be applied to discrete workloads where it makes the most sense. Note there are performance tradeoffs with erasure coding as well.

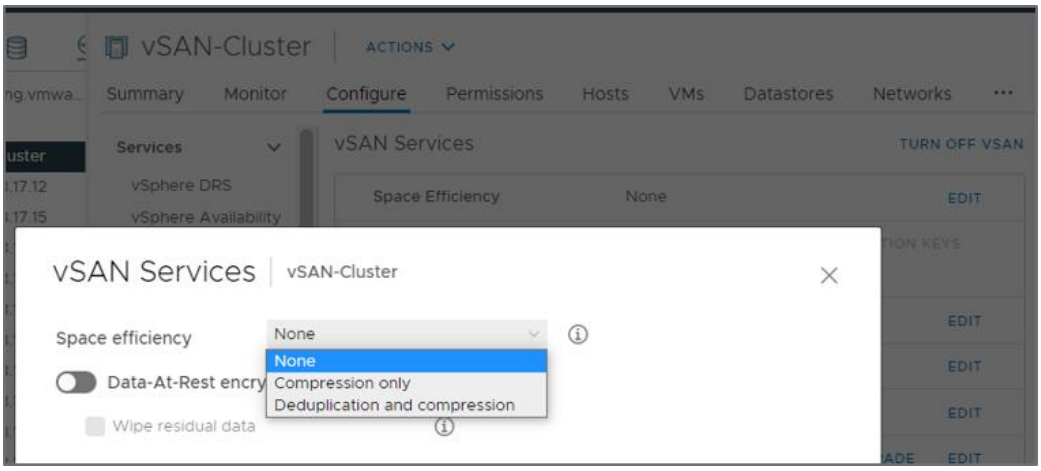
Compression-Only (OSA)

VMware vSAN introduced deduplication and compression (DD&C) at a time many years ago in which NAND flash was much more expensive than it is today. The economics of flash at the time inspired VMware to provide a cluster-based service that combined two space-saving techniques as a single feature for maximum space efficiency and simplicity for all-flash storage environments. Space efficiency techniques are a marvelous innovation, but each type introduces tradeoffs. Some workloads and data may not be ideally suited for certain types of space efficiency. When using the vSAN OSA, deduplication engines will run without regard to the data it is processing, which means that if the data does not deduplicate well, the additional computational effort and I/O amplification provide no benefit in those conditions.

"Compression only" option in vSAN 7 U1 and newer (OSA)

A "Compression only" option in the OSA alleviates the challenge described above. vSAN administrators can use this setting for clusters with demanding workloads that typically cannot take advantage of deduplication techniques. It accommodates today's economics of flash storage while maintaining an emphasis on delivering performance for high demand, latency-sensitive workloads.

Selecting the desired space efficiency option is easy. At the cluster level, the vCenter Server UI now presents three options: 1.) None 2.) Compression only 3.) Deduplication and compression.

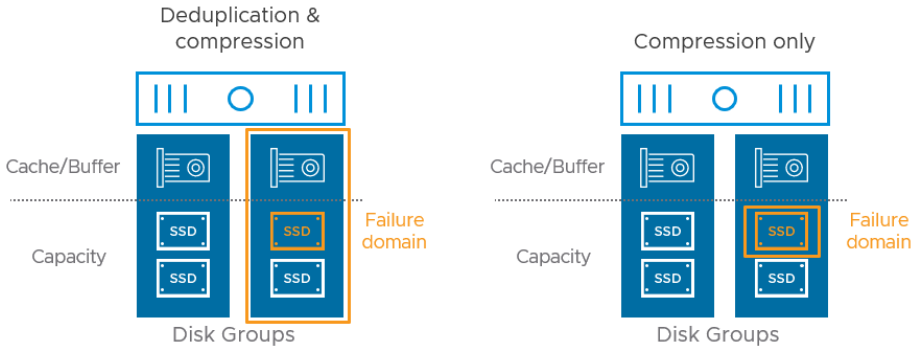


Note that changing this cluster-level setting does require a rolling evacuation of the data in each disk group. This is an automated process but does require resources while the activity is performed.

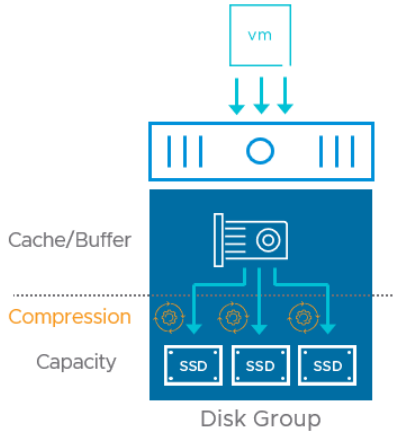
Advantages

When compared to the DD&C option in the OSA, the "Compression only" option in the OSA offers interesting advantages

- **Reduce the failure domain of a capacity device failure.** A failure of a capacity device in a disk group for a cluster using "Compression only" will only impact that discrete storage device, whereas the same failure using DD&C would impact the entire disk group. This reduced impact area of a device failure also reduces the amount of potential data that vSAN needs to rebuild upon a device failure.



- **Increased destaging rates of data from the buffer tier to the capacity tier**. As described in "vSAN Design Considerations - Deduplication and Compression" vSAN's two-tier system ingests writes into a high-performance buffer tier, while destaging the data to the more value-based capacity tier at a later time. The space efficiency processes occur at the time of destaging, and as described in that post, may have a potential impact on performance. When compared to DD&C, the "Compression only" feature improves destage rates in two ways:
  - Avoids the inherent write amplification required with deduplication techniques.
  - Uses multiple elevator processes to destage the data.



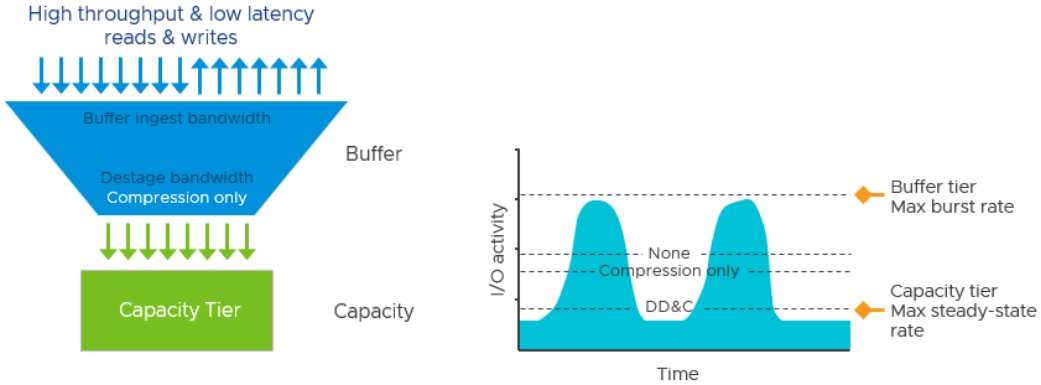
Capacity Savings

How much space savings can one expect using the "Compression only" feature in the OSA? The answer to this depends on the workload, and the type of data being stored. Both of the DD&C and "Compression only" features are opportunistic, which means that space savings are not guaranteed. This capacity savings through compression can be easily viewed in the vCenter Server UI. Note that it may take some time before the savings ratio stabilizes.

Performance

What will the levels of performance be like when using the "Compression only" feature? This will land somewhere in between the performance of your hosts not running any space efficiency, and the performance of your hosts running DD&C.

Performance using "Compression only" could be superior when compared to the same environment using DD&C. This improvement would show up most where there are workloads with large working sets issuing large sequential writes and medium-sized random writes. In these cases, the absence of the deduplication engine and the improved parallelization of destaging will allow the data to be destaged faster, and less likely to hit buffer fullness thresholds that begin to impact the guest VM latency.



The performance capabilities of vSAN are still ultimately determined by the hardware used, the configuration of vSAN, the version of vSAN, the associated storage policies, and the characteristics of the application & workload. To better understand how hardware selection (including the type of flash devices) impact performance, see the post "Write buffer sizing in vSAN when using the very latest hardware."

Deduplication & Compression or Compression-Only (OSA) – Which one is Right for you?

Workloads and data sets do not provide an easy way to know if they are ideally suited for some space efficiency techniques versus others. When using the Original Storage Architecture (OSA) in vSAN, an administrator should decide based on the

requirements of the workloads and the constraints of the hardware powering the workloads. A comparison of design and operational considerations between the three options in **vSAN OSA** is provided below.

	None	Compression-only	DD&C
Capacity savings	None	Moderate*	High*
Resource overhead	None	Minimal	High
Failure domain of capacity device failure	Disk	Disk	Disk group
Impact on destaging performance	None	Minimal	High
Ideal for high-demand workloads	Yes	Yes	Conditional**

* Capacity savings not guaranteed

** Depends on workloads, working sets, and hardware configuration

For some environments, the minimal failure domain of a capacity device failure may be the only reason needed to justify the use of the "Compression only" feature versus the other options. Whatever the case, the configuration desired can be tailored on a per cluster basis. Note that this consideration only exists with the OSA. The vSAN ESA does not use a construct of disk groups, thus eliminating failure domains larger than a single storage device.

When using the vSAN OSA, VMware recommends the following settings for the best balance of capacity savings and performance impact. Workloads and environmental conditions vary, therefore these are generalized recommendations.

Workload	Recommendation	Capacity Savings	Performance Impact
OLTP databases	Compression only	Moderate	Minimal*
Mixed workloads	Compression only	Moderate	Minimal*
VDI using instant clones	Compression only	Moderate	Minimal*
VDI using linked clones	Compression only	Moderate	Minimal*
VDI using full clones	DD&C	High	Moderate*

* In OSA, If performance is of the highest priority, but are limited to using the original storage architecture in vSAN (opposed to the ESA), using no space efficiency would yield the highest sustained performance for the hardware configuration used.

Transition to vSAN ESA if possible, as there are no performance implications with space efficiency features in ESA. If you can't and must remain on OSA for a period of time, and are uncertain as to what is best for your clusters, and you prefer some degree of cluster-level space efficiency with minimal performance impact, choose the "Compression only" feature.

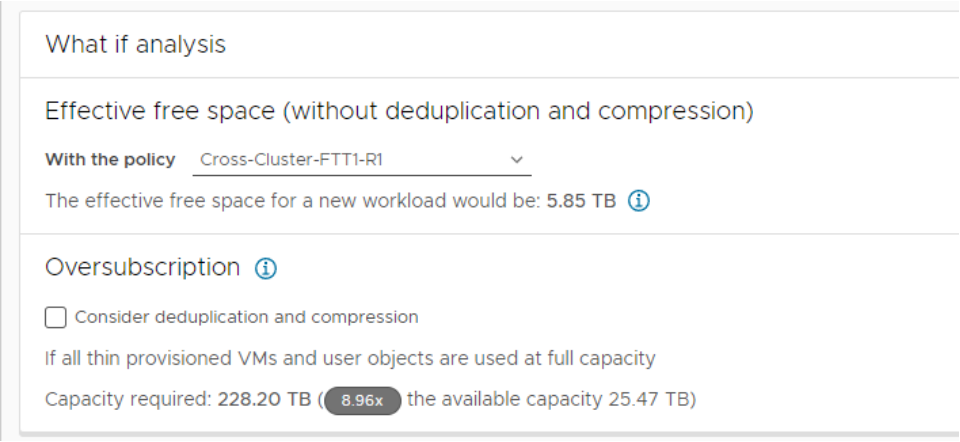
Viewing Space Efficiency Savings

The savings gained from Deduplication & Compression, or Compression-only can be viewed in the vSAN Capacity view in vCenter Server. This will present the savings as a ratio. A ratio of 1x would equal no savings, while a ratio of 2x would mean that the data is only consuming half the capacity that it would have without the space efficiency techniques.

Thin Provisioning

A thin-provisioned storage system provisions the storage on an as-needed basis. It does not provision the entire amount of capacity needed for say, a VMDK, but rather, only the space that has been initially needed by the VMDK. For more information on "provisioned" versus "used" space in vSAN, see the post: Demystifying Capacity Reporting in vSAN.

vSAN supports thin provisioning, which provides the minimum amount of storage capacity needed by the vSAN objects created on the datastore. It will then transparently increase the amount of used space as it is needed. As a result, it is entirely possible to initially provision for more capacity than the physical datastore is able to actually provide. This is known as "oversubscription" and is typical in any storage system that uses thin provisioning. vSAN provides an easy method to determine the level of oversubscription that a cluster is at any given time. This can be found in the cluster capacity view, under the "What if Analysis."



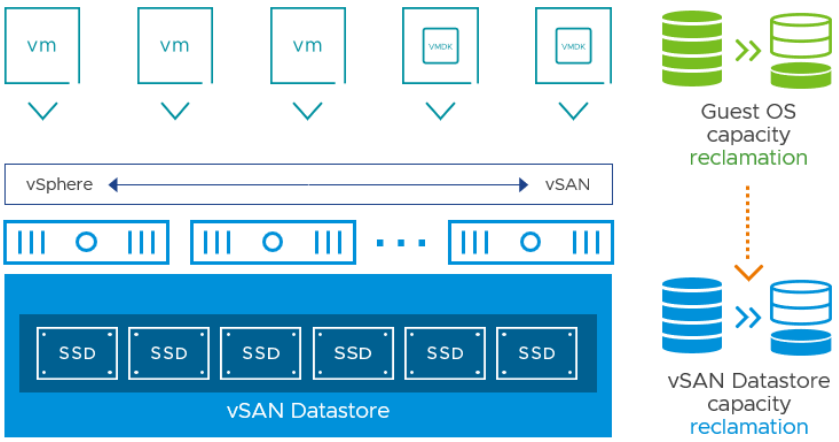
The example above indicates that this particular vSAN cluster is about 9x oversubscribed. Meaning that if all of the deployed VMDKs on the datastore were filled to capacity, the storage system would need to be 9x the size of the existing capacity. Having this ratio is good to know, as organizations can then ensure they stay within a certain oversubscription ratio for their clusters. Oversubscription is a common approach to using storage capacity most effectively.

Recommendation: Choose an oversubscription ratio for your organization that you would like your clusters to fall within. There is no correct answer as to the ideal oversubscription ratio, and the ideal ratio is dependent on the characteristics of an environment and the respective workloads that run in it.

One of the challenges to thin provisioned systems is that once a given entity has grown (such as a VM's VMDK), it will not shrink when data within the guest OS are deleted - common with databases that use transaction log files. This problem is amplified by the fact that many file systems will always direct new writes into free space. A steady set of writes to the same block of a single small file will eventually use significantly more space at the VMDK level. To solve this problem, automated TRIM/UNMAP space reclamation is available for vSAN.

TRIM/UNMAP Space Reclamation

In an attempt to be more efficient with storage space, modern guest OS file systems have had the ability to reclaim no longer used space using what are known as TRIM/UNMAP commands for the respective ATA and SCSI protocols. vSAN has full awareness of TRIM/UNMAP commands sent from the guest OS and can reclaim the previously allocated storage as free space. This is an opportunistic space efficiency feature that can deliver much better storage capacity utilization in vSAN environments.



This process carries benefits of freeing up storage space but also has other secondary benefits:

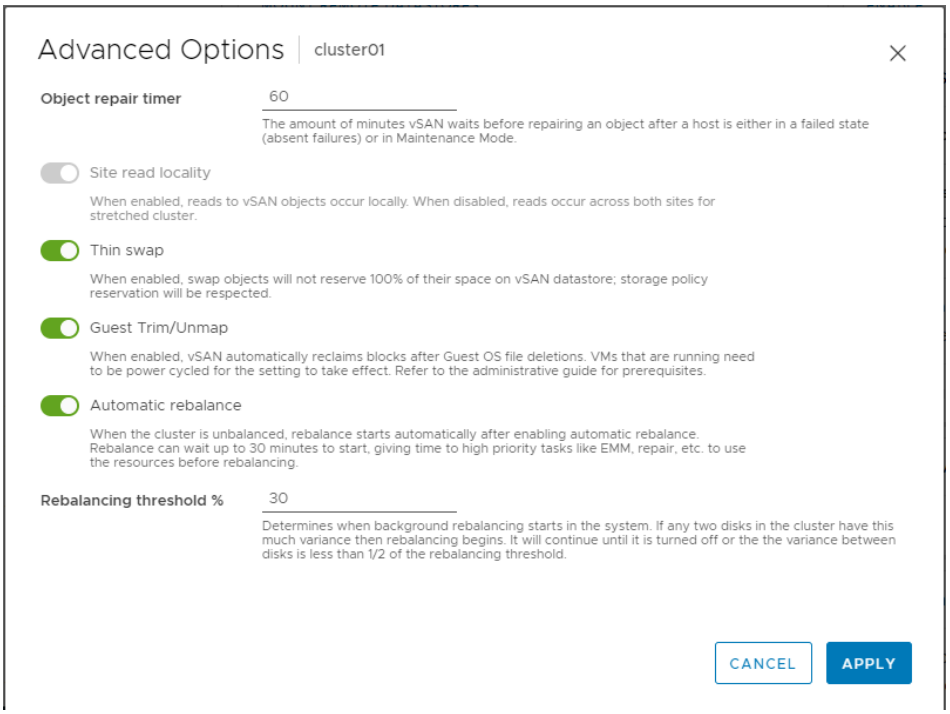
- **Faster repair** - Blocks that have been reclaimed do not need to be rebalanced, or re-mirrored in the event of a device failure.
- **Removal of dirty cache pages** - Read Cache can be freed up in the DRAM client Cache, as well as the Hybrid vSAN SSD Cache for use by other blocks. If removed from the write buffer this reduces the number of blocks that will be copied to the capacity tier.

Performance Impact

This process does carry performance impact as I/O must be processed to track pages that are no longer needed. The largest impact in vSAN OSA will be UNMAP's issued against the capacity tier directly. This will be less of an issue with ESA. For environments with high deletions performance should be monitored.

VMware Specific Guidance

In vSAN 8, TRIM/UNMAP can be enabled in the User Interface, under the services enablement as shown below.



TRIM/UNMAP can be enabled using PowerCLI or RVC. Enabling via PowerCLI is shown below.

Status query:

```
Get-Cluster -name R63*|get-VsanClusterConfiguration |ft GuestTrimUnmap

GuestTrimUnmap

--------------

         False
```

Enable:

```
Get-Cluster -name R63*|set-VsanClusterConfiguration -GuestTrimUnmap:$true

Cluster              VsanEnabled  IsStretchedCluster   Last HCL Updated

-------              -----------  ------------------   ----------------

R630-Cluster-70GA    True         True                 25/04/2020 16:03:00
```

Deactivate

```
Get-Cluster -name R63*|set-VsanClusterConfiguration -GuestTrimUnmap:$false

Cluster              VsanEnabled  IsStretchedCluster   Last HCL Updated

-------              -----------  ------------------   ----------------

R630-Cluster-70GA    True         True                 25/04/2020 16:03:00
```

TRIM/UNMAP can also be enabled per vSAN cluster using the RVC Console. The example below demonstrates how to enable it using RVC.

RVC Command: vsan.unmap_support

-e —enable unmap support on vSAN cluster

-d —disable unmap support on vSAN cluster.

Before running this command make sure the vSAN cluster is healthy, and all hosts are connected to the vCenter.

Example: Using RVC to enable unmap_support

To enable TRIM/UNMAP for vSAN, one first must SSH into the vCenter server and connect to the RVC console.

```
login as: administrator@vsphere.local

Command> rvc administrator@vsphere.local@localhost
```

Next, browse to compute, and identify the cluster name.

```
Welcome to RVC. Try the 'help' command.

0 /

1 localhost/

> ls

0 /

1 localhost/
```

```
> cd 1

/localhost> ls

0 vSAN-DC (datacenter)

/localhost> cd 0

/localhost/vSAN-DC> ls

0 storage/

1 computers [host]/

2 networks [network]/

3 datastores [datastore]/

4 vms [vm]/

/localhost/vSAN-DC> cd 1

/localhost/vSAN-DC/computers> ls

0 vSAN-Cluster (cluster): cpu 37 GHz, memory 108 GB

/localhost/vSAN-DC/computers>
```

Next, enable unmap support.

```
/localhost/vSAN-DC/computers> vsan.unmap_support vSAN-Cluster -e

2022-03-10 16:37:37 +0000: Enabling unmap support on cluster

 vSAN-Cluster: success

VMs need to be power cycled to apply the unmap setting

/localhost/vSAN-DC/computers>
```

Prerequisites - VM Level

Once enabled, there are several prerequisites that must be met for TRIM/UNMAP to successfully reclaim no longer used capacity.

- A minimum of virtual machine hardware version 11 for Windows
- A minimum of virtual machine hardware version 13 for Linux.
- disk.scsiUnmapAllowed flag is not set to false. The default is implied true. This setting can be used as a "stop switch" at the virtual machine level should you wish to disable this behavior on a per VM basis and do not want to use in guest configuration to disable this behavior. VMX changes require a reboot to take effect.
- The guest operating system must be able to identify the virtual disk as thin.
- After enabling at a cluster level, the **VM must be powered off and back on**. (A reboot is insufficient). This requirement is a one-time operation on each VM after it is enabled.

Linux Specific Guidance

There are two primary means of reclaiming thin provisioning.

- fstrim is used on a mounted filesystem to discard (or "trim") blocks which are not in use by the filesystem. This is useful for thinly-provisioned storage. Depending on the distribution, this may or may not be included in a cron job, such as /etc/cron.weekly. To manually perform the in-guest reclamation, perform the following:

```
/sbin/fstrim --all || true
```

- blkdiscard is used to discard device sectors. Unlike strip(8), this command is used directly on the block device. blkdisacrd is known to have more performance overhead than fstrim. As a result, fstrim is recommended over blkdiscard.

Other Concerns:

- If using encrypted file systems you may need to add discard to /etc/crypttab.
- If shrinking or deleting LVM volumes, the issue_discards configuration may be needed in /etc/lvm/lvm.conf
- Different options for automating the running of fstrim exist. These range from weekly cron tasks, to fstrim.timer
- The following file systems are reported to work with TRIM: btrfs, ecryptfs, ext3, ext4, f2fs, gfs2, jfs, nilfs2, ocfs2, xfs

Microsoft Specific Guidance

Automated Space Reclamation

Windows Server 2012 and newer support automated space reclamation. This behavior is enabled by default. To check this behavior, the following PowerShell can be used.

```
Get-ItemProperty

-Path

"HKLM:\System\CurrentControlSet\Control\FileSystem"

-Name

DisableDeleteNotification
```

To enable automatic space reclamation this value the following:

```
Set-ItemProperty

-Path

"HKLM:\System\CurrentControlSet\Control\FileSystem"

-Name

DisableDeleteNotification

-Value

0
```

Asynchronous Space Reclamation

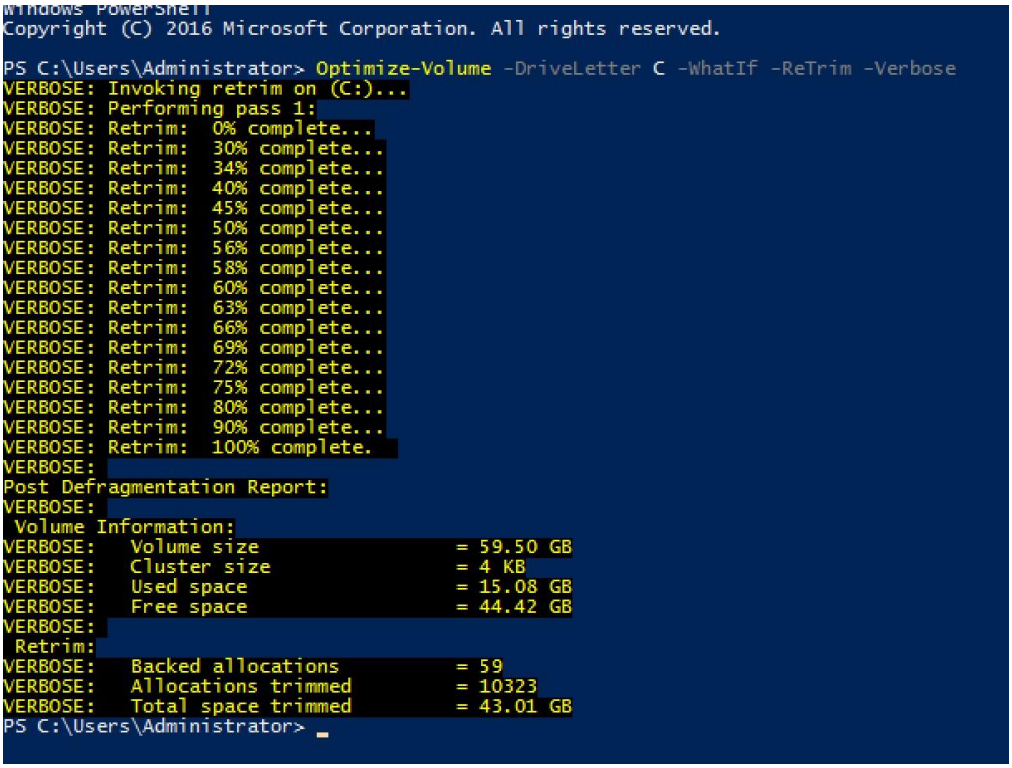
Two methods exist for asynchronously reclaiming space.

Example 1: Perform TRIM optimization

PowerShell

```
PS C:\>Optimize-Volume -DriveLetter H -ReTrim -Verbose
```

This example optimizes drive H by re-sending Trim requests. The -WhatIf flag can be added to test if TRIM commands are being passed cleanly to the backend.



Example 2: Perform TRIM optimization

defrag /L can also be used to perform the same operation as the Optimize-Storage -ReTrim command.

```
Defrag C: D: /L
```

This Example would reclaim space on both volume C: and D:.

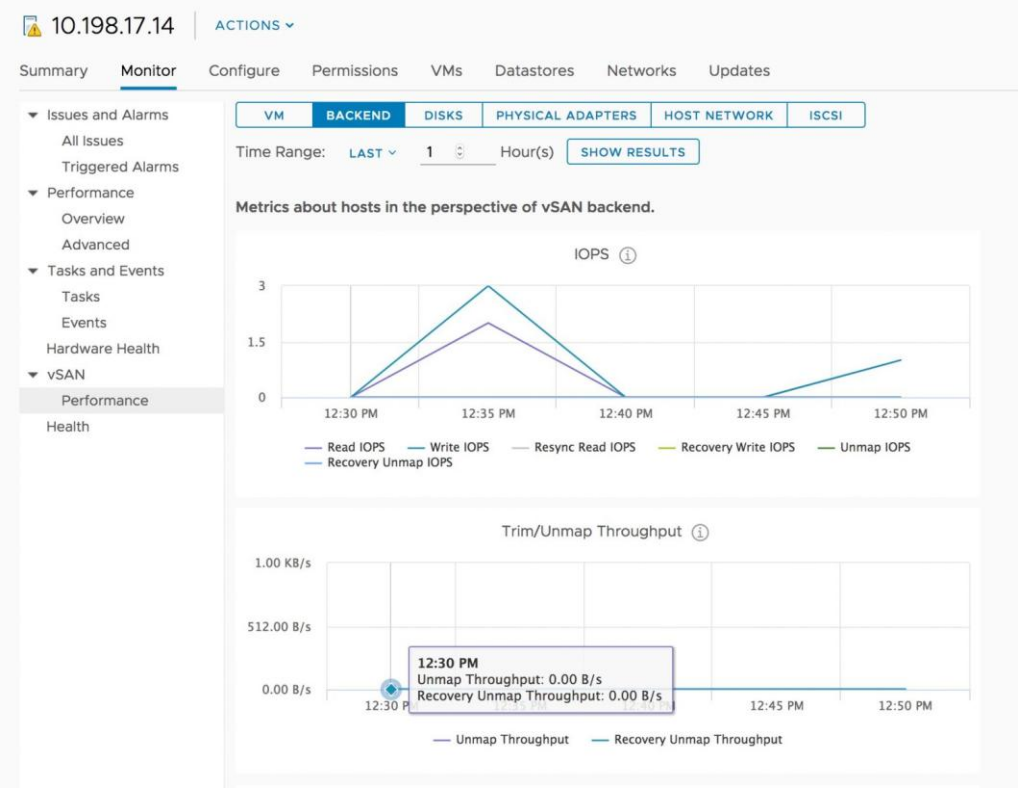
Other Concerns

- Windows when using Optimize-Storage or Defrag /L will report sending TRIM commands to all unused blocks. This reporting should not be relied upon up to determine how much space will be reclaimed.
- It is recommended to primarily use the automatic Delete Notification.

Monitoring TRIM/UNMAP

TRIM/UNMAP has the following counters in the vSAN performance service:

- **UNMAP Throughput.** The measure of UNMAP commands being processed by the disk groups of a host
- **Recovery UNMAP Throughput.** The measure of throughput of UNMAP commands be synchronized as part of an object repair following a failure or absent object.



VADP Backup Considerations

Snapshots are commonly used by backup vendors to provide a known state for backing up VM data at a given point in time. When a snapshot is no longer needed by the backup process, those changes are merged with the base disk. Snapshots and other tasks such as Storage vMotion use VMware's mirror driver to ensure that ongoing changed blocks are mirrored to more than one location. In these circumstances where the mirror driver is used, TRIM/UNMAP commands will not be passed down to the base disk, preventing full reclamation from occurring. One method of accommodating for these conditions is to use a pre-freeze-script.

For Windows:

```
C:\Windows\pre-freeze-script.bat
```

For all other operating systems:

```
/usr/sbin/pre-freeze-script
```

Running the fstrim or DiskOptimize before a snapshot will clean out any deleted files that happened during a previous backup window.

UNMAP with vSAN File Services

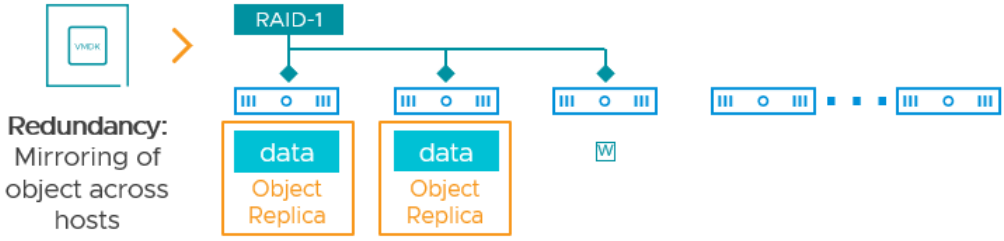
vSAN File Services (in vSAN 7 U2 and later) supports UNMAP and will automatically reclaim storage when files are deleted from the NFS or SMB share. TRIM/UNMAP must be enabled on the cluster in order for this process to take place. Depending on the circumstances, it make take a few minutes for the reclaimed storage to be reported correctly within vCenter Server.

Deterministic Space Efficiency Features

Deterministic space efficiency techniques will result space efficiency that can be specifically determined. The degree to which the savings will occur will depend on the form of space efficiency used.

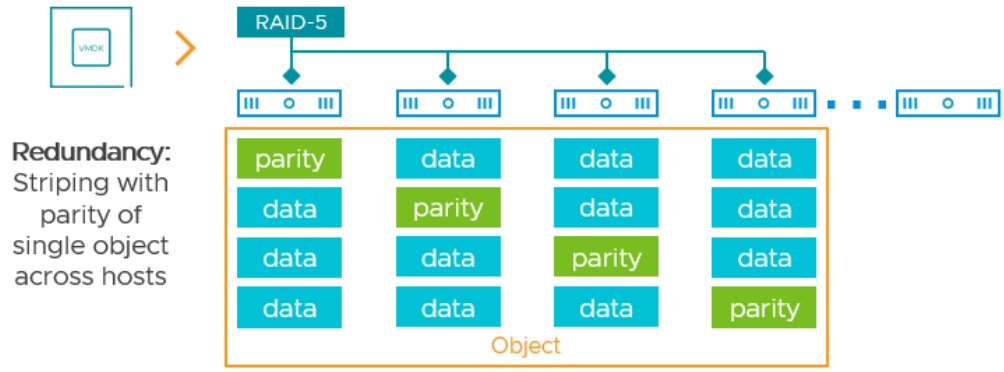
Data Placement Schemes and Erasure Code Concepts

vSAN achieves resilience of data in different ways. One way is by having a copy, or a mirror of a chunk of data (an object in vSAN) to one or more locations, or hosts. This replica of an object resides somewhere else in the cluster to provide resilience. The level of resilience is defined in the assigned storage policy, and vSAN takes care of the rest, placing it in the cluster to achieve the desired result. A level of failure to tolerate of 1 (FTT=1) when using RAID-1 mirroring creates two copies of that object. FTT=2 creates three copies of that object, and an FTT=3 creates four copies of the object.



Data mirroring is a simple data placement scheme that uses minimal computational overhead but comes with the tradeoff of using an equal amount of capacity somewhere else in the cluster to protect the object at the level of resilience you desire.

The other way **vSAN achieves data resilience is through the use of erasure codes**. Erasure coding is a method of fragmenting data across some physical boundary in a manner that maintains access to the data in the event of a fragment or fragments missing. In vSAN's case, erasure codes do this striping data with parity across hosts. Unlike a RAID-1 mirror where there are two or more copies of the data, there will only be a single instance of an object using RAID-5/6 erasure coding. The data with parity is spread across the hosts to provide this resilience. In the vSAN OSA, an object assigned FTT=1 using erasure coding (RAID-5) will maintain availability in the event of a single failure (e.g. host) and will spread that data across 4 hosts . An object assigned an FTT=2 using erasure coding (RAID-6) will maintain availability in the event of a double failure, spreading that data across 6 hosts. **As a reminder, an object using erasure coding is not spread across all hosts.** vSAN's approach offers superior resilience under failure conditions and simplified scalability.



The benefit of erasure codes is predictable space efficiency when compared to mirroring data. In the OSA, providing resilience for a single failure (FTT=1) for an object using erasure coding consumes just 1.33x the capacity of the single object. Providing resilience for a double failure (FTT=2) for an object using erasure coding consumes just 1.5x the capacity of the single object. With mirroring, FTT=1 and FTT=2 would consume 2x and 3x the capacity, respectively. The Original Storage Architecture (OSA) in vSAN has seen several improvements over recent releases, but cannot match the performance capabilities of erasure codes in vSAN ESA.

Recommendation: If you are running vSAN OSA, upgrade to the very latest version. The performance improvements introduced in recent editions of vSAN can make a profound difference in the effective performance of VMs using RAID-5/6 erasure coding.

Implications on Storage Capacity

Traditional storage array capacity is sometimes described in "raw" capacity, and sometimes in "usable" capacity. Raw capacity is finite based on the hardware in the array, but usable capacity can fluctuate. This fluctuation can occur depending on what data protection techniques are used. Any capacity overhead they incur is directly reflected in the amount of usable capacity. Data protection techniques are typically masked by the array, and vSphere uses only what is presented to it. An array with 100TB of capacity with only data mirroring in use would provide 50TB of usable capacity. vSAN differs from traditional storage, as the raw capacity is exposed directly in the vSAN datastore. Data protection in vSAN is configured through storage policies and is independently configurable per object. **Effectively, the usable capacity changes depending on the policies applied to vSAN objects.**

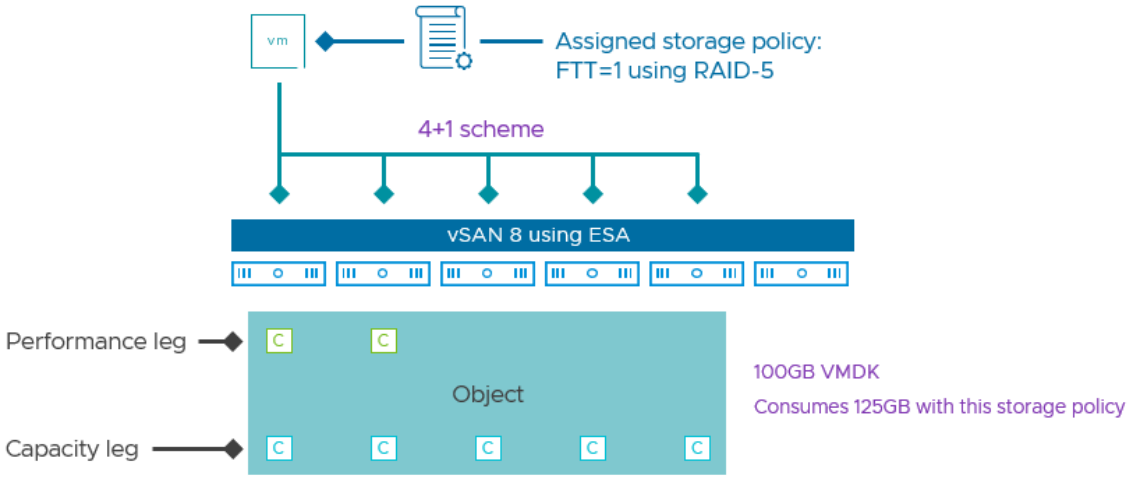
For more information on how storage policies impact capacity utilization, see the post: Demystifying Capacity Reporting in vSAN. To learn more about how to reserve the proper amount of capacity in a vSAN cluster, see the post: Understanding "Reserved Capacity" Concepts in vSAN.

RAID-5 Erasure Coding (ESA)

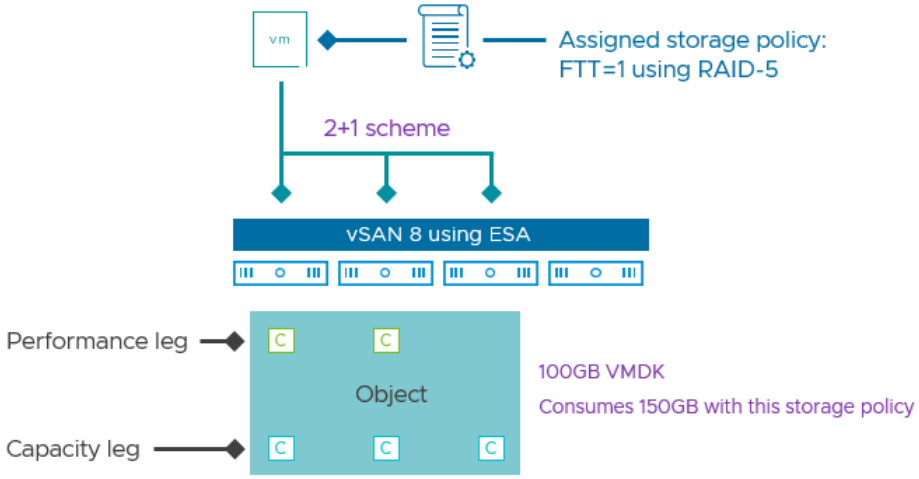
Erasure coding in the vSAN Express Storage Architecture (ESA) is significantly different than the implementation found in the vSAN OSA. For the ESA, **this yields deterministic space efficiency without any compromise in storage performance.**

Where the OSA uses a 3+1 scheme, spreading data and parity fragments across at least 4 hosts, the ESA uses two distinct RAID-5 erasure codes exclusive to the ESA. The choice of RAID-5 scheme used is entirely based on conditions that vSAN detects in the cluster. A user is presented with a single RAID-5 policy, and vSAN will determine which RAID-5 scheme to use.

For clusters with 6 or more hosts, it uses a scheme of 4+1, spreading the data and parity fragments across at least 5 hosts. This consumes just 1.25x the capacity of the primary data (compared to 1.33x when using RAID-5 in the OSA), making it extremely space efficient without any performance tradeoff.



For clusters with fewer than 6 hosts, it will use a scheme of 2+1, spreading the data and parity fragments across at least 3 hosts. While this consumes a bit more capacity to make the data resilient (1.5x the capacity of the primary data), it is MUCH more space efficient than mirroring (2x capacity of the primary data), and can run on clusters with as few as three hosts, while also not impacting performance in any way. This makes RAID-5 an ideal choice for smaller clusters powered by the ESA.

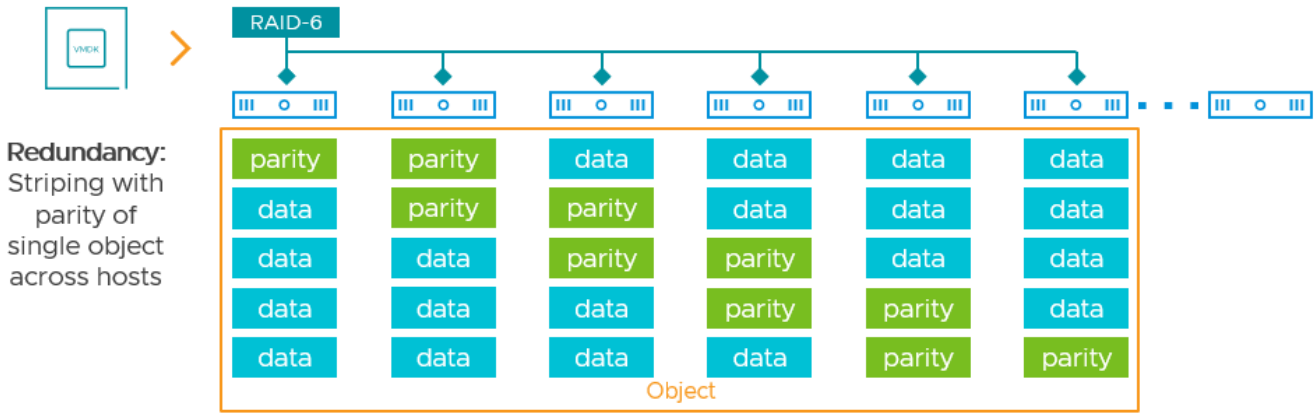


As a result of effective behavior of erasure codes in the ESA, RAID-5/6 erasure coding should be the data placement schemes of choice when the topology supports it. For stretched clusters, and 2-node topologies will still require the use of RAID-1 mirroring to protect the data across a site, or host respectively, but the secondary levels of resilience for these topologies can use erasure coding.

For more information, see the post: "RAID-5/6 with the Performance of RAID-1 using the vSAN Express Storage Architecture." And it also provides the ability to use RAID-5 erasure coding on as few as three hosts. For more information, see the post: "Adaptive RAID-5 Erasure Coding with the Express Storage Architecture in vSAN 8."

RAID-6 Erasure Coding (ESA)

A RAID-6 erasure code in the ESA uses a 4+2 scheme, where data and parity fragments are spread across at least 6 hosts, similar to what is shown below.



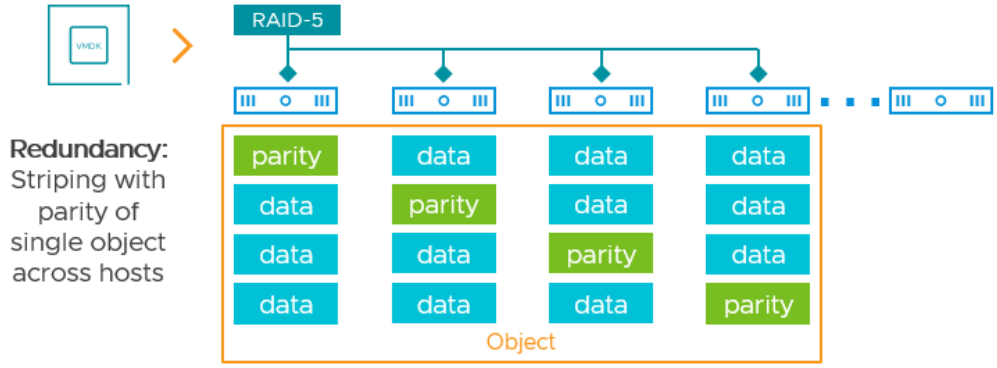
A 4+2 scheme allows vSAN to suffer the failure of two hosts for the hosts that are used to store this given object while maintaining availability of the object. The most significant difference with the ESA and RAID-6 is the complete elimination of CPU and I/O amplification that occurred with RAID-6 in the OSA. *In the ESA, RAID-6 is your best option for offering supreme levels of storage resilience while maintaining very good space efficiency, and no compromises on performance.*

RAID-5 Erasure Coding (OSA)

Data Placement

When using a RAID-5 data placement scheme in a storage policy, the objects using that policy will spread the data with parity, in the form of components, across hosts to provide the assigned level of resilience. Using the RAID-5 erasure code will have an implied level of failure to tolerate (FTT) of 1. It will maintain the availability of an object if one of the 4 hosts it

resides on is offline. For RAID-5, the vSAN OSA uses a 3+1 data placement scheme, where a minimum of one component resides on each of the 4 hosts holding the object. Each component will consist of both data and parity fragments. There is not a dedicated parity component.



Space Savings

When comparing to RAID-1 mirroring using an FTT=1, the effective space savings for storing data in a resilient manner can be summarized as the following:

Number of Failures to Tolerate (FTT)	Capacity Consumption using RAID-1 mirroring	Capacity Consumption using RAID-5 erasure coding (OSA)	Capacity Savings in OSA
1	2x	1.33x	33% reduction

As an example, a VM comprised of 100GB of data would consume 200GB worth of data to store in a resilient manner using RAID-1 mirroring, but would consume just 133GB worth of data to store in a resilient manner using RAID-5 erasure coding.

Host Requirements

When comparing to RAID-1 mirroring, the absolute minimum, and recommended minimum number of hosts required to use RAID-5 in the OSA can be summarized as the following:

Number of Failures to Tolerate (FTT)	Absolute Minimum number of Hosts	Recommended Minimum number of Hosts
FTT=1 using RAID-1 (OSA)	3	4
FTT=1 using RAID-5 (OSA)	4	5

The "Recommended minimum" reflects the desire to have an available host for the object to automatically rebuild to in order to regain its prescribed level of resilience.

These numbers are different when discussing the vSAN ESA. For a listing of host minimums, see Figure 5 in the post: "Adaptive RAID-5 Erasure Coding with the Express Storage Architecture in vSAN 8."

RAID-6 Erasure Coding (OSA)

Data Placement

When using a RAID-6 data placement scheme in a storage policy, the objects using that policy will spread the data with parity, in the form of components, across hosts to provide the assigned level of resilience. Using the RAID-6 erasure code will have an implied level of failure to tolerate (FTT) of 2. It will maintain the availability of an object if two of the 6 hosts it resides on is offline. For RAID-6, vSAN uses a 4+2 data placement scheme, where a minimum of one component resides on

each of the 6 hosts holding the object. Each component will consist of both data and parity fragments. There is not a dedicated parity component.

Space Savings

When comparing to RAID-1 mirroring using an FTT=2, the effective space savings for storing data in a resilient manner can be summarized as the following:

Number of Failures to Tolerate (FTT)	Capacity Consumption using RAID-1 mirroring	Capacity Consumption using RAID-5 erasure coding (OSA)	Capacity Savings in OSA
2	3x	1.5x	50% reduction

This applies to both the vSAN OSA and vSAN ESA. As an example, a VM comprised of 100GB of data would consume 300GB worth of data to store in a resilient manner using RAID-1 mirroring (with FTT=2) but would consume just 150GB worth of data to store in a resilient manner using RAID-6 erasure coding.

Host Requirements

When comparing to RAID-1 mirroring, the absolute minimum, and recommended minimum number of hosts required to use RAID-6 can be summarized as the following:

Number of Failures to Tolerate (FTT)	Absolute Minimum number of Hosts	Recommended Minimum number of Hosts
FTT=2 using RAID-1 (OSA)	5	6
FTT=2 using RAID-6 (OSA)	6	7

The "Recommended minimum" reflects the desire to have an available host for the object to automatically rebuild to in order to regain its prescribed level of resilience under a single host failure. To support the ability for an automatic rebuild to full prescribed compliance under a double failure, the minimum number of hosts need for RAID-6 would be 8 hosts.

Erasure Coding Recommendations (OSA)

While RAID-5/6 erasure coding in vSAN OSA provides significant capacity savings over RAID-1 mirroring, it requires more effort to place data. This effort can come in the form of CPU cycles needed to write data, update data, repair data under failure conditions or policy changes. In contrast, mirroring does not perform any parity calculations, and may involve writing to fewer hosts. Since erasure coding can be used by simply assigning a new storage policy to an object, it is easy to determine the level of impact in your own environment. The following recommendations may help you incorporate RAID-5/6 erasure coding in vSAN OSA in your environment. **These recommendations below do NOT apply to ESA, as erasure coding in ESA has no negative impact on performance.**

- **Consider the application.** Some applications such as ERP systems and OLTP applications are more sensitive to higher latency than others. Be sure to thoroughly test the resulting behavior of the application after the change has been made.
- **Be prescriptive.** Using RAID-5/6 is not an all or nothing decision. It can be prescribed on a per VM or per VMDK basis using storage policies. Use it where it makes sense and can save you capacity, while staying with mirroring if your hardware is not capable of delivering performance using RAID-5.
- **Exercise caution if using RAID-5/6 as the secondary level of resilience in a vSAN stretched cluster.** See the post: Performance with vSAN Stretched Clusters for more details.
- **Understand the impacts of storage policies when troubleshooting performance.** Changing storage policies can be a way of determining the root cause of performance issues. See the section of mitigations options in the Troubleshooting vSAN Performance Guide.

- **Watch the network.** RAID-5/6 erasure coding can be more demanding on a network than RAID-1 mirroring. 25Gb networking is now a cost-effective option that can relieve many of the pressures that may exist on networking in an HCI environment.
- **Opt for "Compression-only" as the data service to use in combination with erasure codes.** You may use Erasure coding policies, and cluster based space efficiency features together, but this can cause a performance burden if the hardware is insufficient. The "Compression-only" feature will be a better option to use with erasure codes than the Deduplication & Compression option. See "Using Space-Efficient Storage Policies (Erasure Coding) with Clusters Running DD&C" (5-12) in the vSAN Operations guide for more information.
- **Use fast storage devices.** Using fast storage devices at the buffer tier and the capacity tier can help with sustained write workloads.
- **Use the latest version of vSAN.** An extraordinary amount of effort has been made to improve the effective performance of vSAN, especially when using erasure coding. vSAN 7 U2, U2, and U3 all contained significant performance enhancements for data using a RAID-5/6 data placement scheme.
- **Be mindful of storage reclamation processes**. Storage reclamation through TRIM/UNMAP can play an important role in not only reducing used capacity in an environment, but improve the accuracy of capacity reporting. For more information, see the post: The Importance of Space Reclamation for Data Usage Reporting in vSAN

Note that more recent editions of vSAN OSA can drive better performance using erasure coding than in much older versions. Even with the OSA, this can place strain on the network if the workloads are demanding it. While the vSAN OSA has Adaptive Resync capabilities to manage fairness of different types of vSAN traffic, this only applies to the physical storage stack in each host (e.g. the disk groups). It cannot manage bandwidth across the network. The vSAN ESA has an adaptive network traffic shaping feature that will provide this ability but is limited to the ESA.

RAID-1, RAID-5, or RAID-6 – What to Choose?

The use of multiple data placement schemes helps a distributed storage system like vSAN store data in a resilient manner in several ways, depending on the topology. For example, in a 2-node cluster, the data can only be mirrored, while a larger cluster can use erasure coding to store the resilient data in a more space efficient way. The vSAN OSA also introduced considerations when using erasure codes, as it could impact performance.

With the vSAN ESA, **these considerations are simplified greatly.** Since there are no performance implications when using erasure codes in the ESA, erasure coding should be the choice when the cluster size supports it. The level of resilience (FTT=1 using RAID-5, or FTT=2 using RAID-6) is entirely up to you. The table below summarizes the "Failure to Tolerate" (FTT) resilience setting of a given storage policy applied to one or more VMs, and the associated multiplier used for actual space consumed on a vSAN datastore. By multiplier, it means that for every GB of data stored in a VM, it will consume that GB multiplied by the value listed below to store it in a resilient way.

Comparing Erasure Code Options in the vSAN ESA and OSA

Storage Policy Rule	vSAN Architecture	Placement scheme	Capacity consumption to make data resilient (lower is better)	Absolute minimum number of hosts in cluster	Recommended minimum number of hosts in cluster	Performance impact compared to RAID-1
FTT=1 using RAID-1	Both	2x mirror	2x	3	4	N/A
FTT=1 using RAID-5	OSA	3+1 erasure code	1.33x	4	5	Moderate
FTT=1 using RAID-5 (adaptable)	ESA	4+1 erasure code	1.25x	5	6	None
FTT=1 using RAID-5 (adaptable)	ESA	2+1 erasure code	1.5x	3	4	None
FTT=2 using RAID-1	Both	3x mirror	3x	5	6	N/A
FTT=2 using RAID-6	OSA	4+2 erasure code	1.5x	6	7	Moderate
FTT=2 using RAID-6	ESA	4+2 erasure code	1.5x	6	7	None

When using the vSAN OSA, while erasure coding is extremely efficient with capacity, it does consume additional CPU and network resources. For all new data written or updated, calculations must occur to complete the stripe with parity, which leads to an inherent amplification of I/Os to read the data, calculate the parity, and distribute the data and parity across hosts. The result is that this can impact the performance as seen by the guest VM. The amount will depend on the host hardware, the networking hardware, and the characteristics of the workload.

When using the vSAN ESA, the architecture allows for full use of RAID-5/6 erasure coding without any compromise in performance. And it can run on as few as three hosts. **When using the ESA, RAID-5/6 should be the standard in data placement schemes**, with the only exceptions being stretched clusters and 2-Node topologies.

The use of the ESA can also simplify your storage policy configurations. For performance focused applications, it wasn't uncommon to find customers apply perhaps a RAID-1 storage policy to a VMDK that required higher levels of performance while using a RAID-5/6 policy for VMDKs that required better space efficiency. With the ESA, you can use a single policy for these VMs, simplifying configuration and ongoing operations.

Recommendation: If you aren't on 25/100Gb Ethernet already, you should be planning for it. The capabilities of modern storage devices and the hosts they live on can saturate 10Gb ethernet and can no longer keep up with the power of other modern hardware. As the performance capabilities of the hosts' increase, so should your network.

Summary

Space efficiency techniques are a common way to improve efficiency with data storage, and as a result, drives down the effective cost of the storage solution. VMware vSAN (OSA and ESA) provides several forms of opportunistic and deterministic space efficiency features that are easy to implement. When using the OSA, while many of the techniques are applicable to most environments, the considerations in this document should be reviewed to ensure that your capacity management objectives align with your workload performance objectives. When using the ESA, vSAN will be able to deliver numerous space saving technologies with supreme levels of performance, and a simplified management experience.

Additional Resources

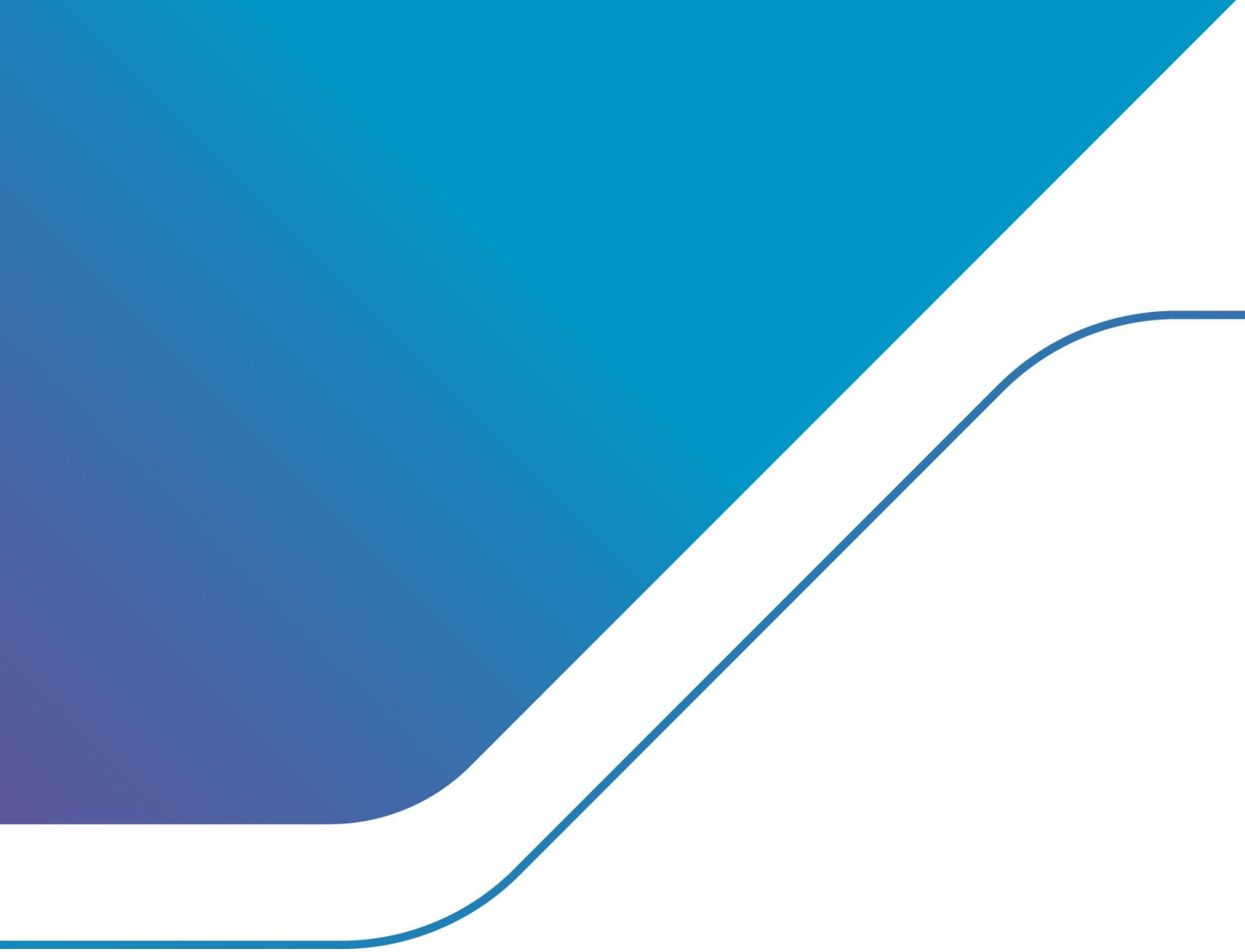
vSAN technical blogs. Stay up to date on the most recently published technical information about vSAN. These posts are created by the vSAN Technical Marketing team.

VMware Resource Center. The location for design guides, operations guides and other technical white papers on vSAN. These assets are created by the vSAN Technical Marketing and Product Enablement teams.

Official vSAN documentation. The location for all “how to” documentation on vSAN.

About the Author

Pete Koehler is a Product Marketing Engineer in the VCF division at Broadcom. With a primary focus on vSAN, Pete covers topics such as design and sizing, operations, performance, troubleshooting, and integration with other products and platforms.



vmware® by Broadcom

Copyright ©2025 Broadcom. All rights reserved.
The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries. For more information, go to www.broadcom.com. All trademarks, trade names, service marks, and logos referenced herein belong to their respective companies. Broadcom reserves the right to make changes without further notice to any products or data herein to improve reliability, function, or design. Information furnished by Broadcom is believed to be accurate and reliable. However, Broadcom does not assume any liability arising out of the application or use of this information, nor the application or use of any product or circuit described herein, neither does it convey any license under its patent rights nor the rights of others.