vSAN Space Efficiency Technologies

First Published On: 07-07-2016
Last Updated On: 04-28-2020

Table Of Contents

1. Introduction

About Storage Space Efficiency

1. 1 Overview

VMware Hyper-Converged Software transforms industry-standard x86 servers and directly attached storage into radically simple Hyper-Converged Infrastructure (HCI) to help eliminate high IT costs, management complexity, and performance limitations. VMware Hyper-Converged Software enables the lowest cost and highest performance HCI solutions available. The tightly integrated software stack includes VMware vSphere®, the market-leading hypervisor; VMware vSAN™, radically simple, enterprise-class native storage; and VMware vCenter Server™, a unified and extensible management solution.

vSAN is natively integrated with vSphere and it is configured with just a few mouse clicks. Since disks internal to the vSphere hosts are used to create a vSAN datastore, there is no dependency on external shared storage. Virtual machines can be assigned specific storage policies based on the needs of the applications. These workloads benefit from dependable shared storage with predictable performance and availability characteristics.

The new release of vSAN, version 6.2, further reduces TCO by reducing data capacity consumption by as much as 10x. vSAN 6.2 is optimized for modern all-flash storage, delivering efficient near-line deduplication, compression, and erasure coding capabilities that enable high performance all-flash systems for as low as $1 per GB of usable capacity—up to 50% less than the cost of lower-performing hybrid solutions from the leading competition. Before getting into the details of the space efficiency features of vSAN, it is important to better understand the economics of flash in today's datacenters.

1. 2 Why the Future is All-Flash

At the initial launch of vSAN in the March of 2014, flash was still relatively expensive. Hybrid vSAN with 10K RPM and 7.2K RPM drives offered a lot of value by enabling cost effective capacity to be blended with amazing performance. Over the past 18 months, flash has significantly closed in on the price advantage of traditional magnetic disks while still offering significantly better IOPS and latency than many workloads need.

A primary goal of space efficiency features in vSAN 6.2, was to close the cost per GB gap between flash and 10K RPM drives. We expect many customers to find that deduplication, compression and erasure coding, will enable a shift to all-flash vSAN designs, providing a balance of performance consistency and cost.

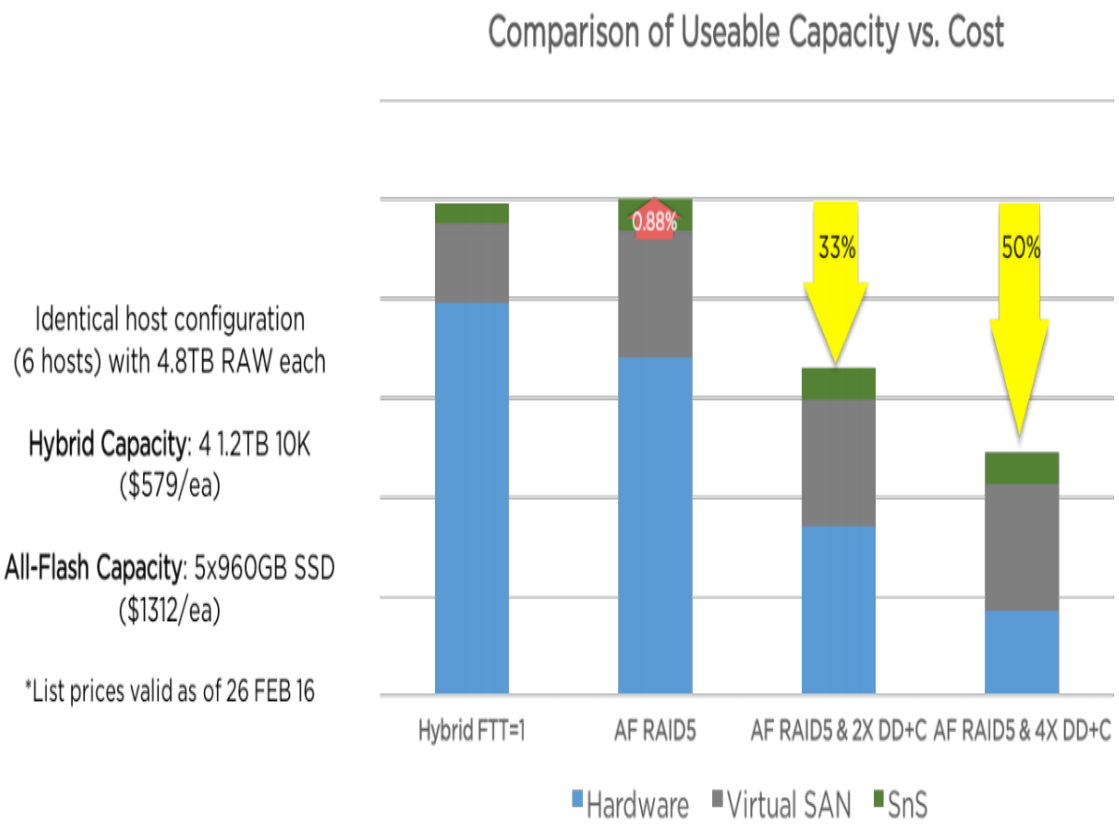


Figure 1 Usable Capacity versus Cost for FTT=1 Configurations

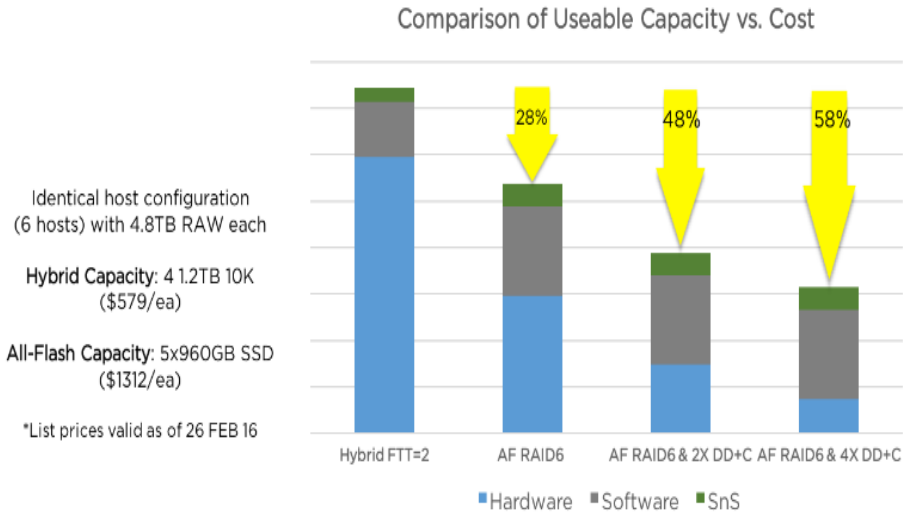


Figure 2 Usable Capacity versus Cost for FTT=2 Configurations

It is important to highlight the external factors making all-flash vSAN more affordable than ever. Technology advances with NAND such as 3D layering, multi-level cell (MLC) flash, and triple-level cell (TLC) flash will continue to drive down the cost of flash devices. New classes of persistent memory and new interfaces such as Non-Volatile Memory Express (NVMe) deliver even better performance.

Looking at the trends for traditional magnetic drives there are some interesting observations. We will see 7.2K RPM drives replace 10K as the dominant chosen drive for hybrid deployments focused on optimizing capacity cost. For working sets that can tolerate slight increases in latency on read cache miss or that have data growth increasing with working sets staying small, we expect this to continue to deliver great value. As new generations of NL-SAS drives increase in capacity the quantity of IOPS do not increase by the same ratio. This presents some challenges in that with the same read cache miss ratio, larger drives will increasingly risk contention and higher variations in latency.

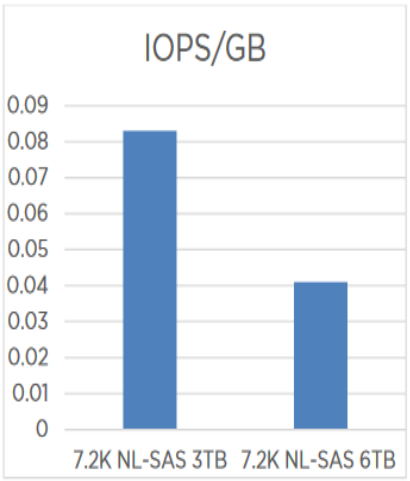
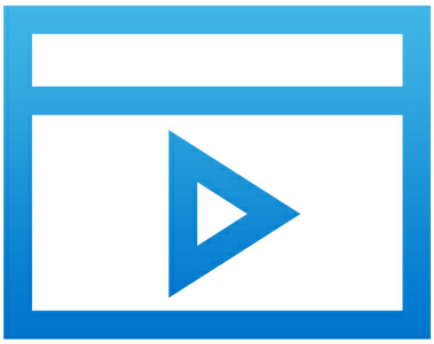


Figure 3 IOPS per Gigabyte as Drive Capacity increases

Storage trends are becoming more predictable. Hybrid will continue to provide excellent value for specific workloads, but falling flash prices and data reduction technologies will close the gap enough for many customers to make all-flash vSAN configurations the primary deployment method.



Click to see topic media

2. Space reclamation on VSAN

Guidance on VMware's implementation of vSAN Space Efficiency

2. 1 UNMAP/TRIM Space Reclamation on vSAN

Virtual Machine Space Reclamation

Thin Provisioning

vSAN supports thin provisioning, which lets you, in the beginning, use just as much storage capacity as currently needed and then add the required amount of storage space at a later time. Using the vSAN thin provisioning feature, you can create virtual disks in a thin format. For a thin virtual disk, ESXi commits only as much storage space as the disk needs for its initial operations. To use vSAN thin provisioning set the SPBM Policy for Object Space Reservation (OSR) to its default of 0.

One challenge to thin provisioning is that VMDK's once grown will not shrink when files within the guest OS are deleted. 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. Previous solutions to this required manual intervention and storage vMotion to external storage, or powering off a virtual machine. To solve this problem automated TRIM/UNMAP Space reclamation was created for vSAN 6.7U1

What is TRIM/UNMAP

In an attempt to be more efficient with storage space, modern guest OS filesystems 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 6.7 U1 now has full awareness of TRIM/UNMAP command 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.

Benefits

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 IO must be processed to track pages that are no longer needed. The largest impact will be UNMAP's issued against the capacity tier directly. For environments with high deletions performance should be monitored.

VMware Specific Guidance

TRIM/UNMAP can be enabled using PowerCLI.

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
```

Disable:

```
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 is enabled per vSAN cluster using the RVC Console.

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

First Connect to the RVC console

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

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

```
rvc administrator@vsphere.local@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 130 GHz, memory 560 GB
```

Now, enable unmap support.

```
/localhost/VSAN-DC/computers> vsan.unmap_support VSAN-Cluster
-e
Unmap support is already disabled
 VSAN-Cluster: success

VMs need to be power cycled to apply the unmap setting
/localhost/VSAN-DC/computers>
```

Virtual Machine Level

- 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 an implied true. This setting can be used as a "kill 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, virtual machines must be power cycled.

Linux Specific Guidance

There are two primary means of reclaiming thin provisioning.

1. 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.
2. 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

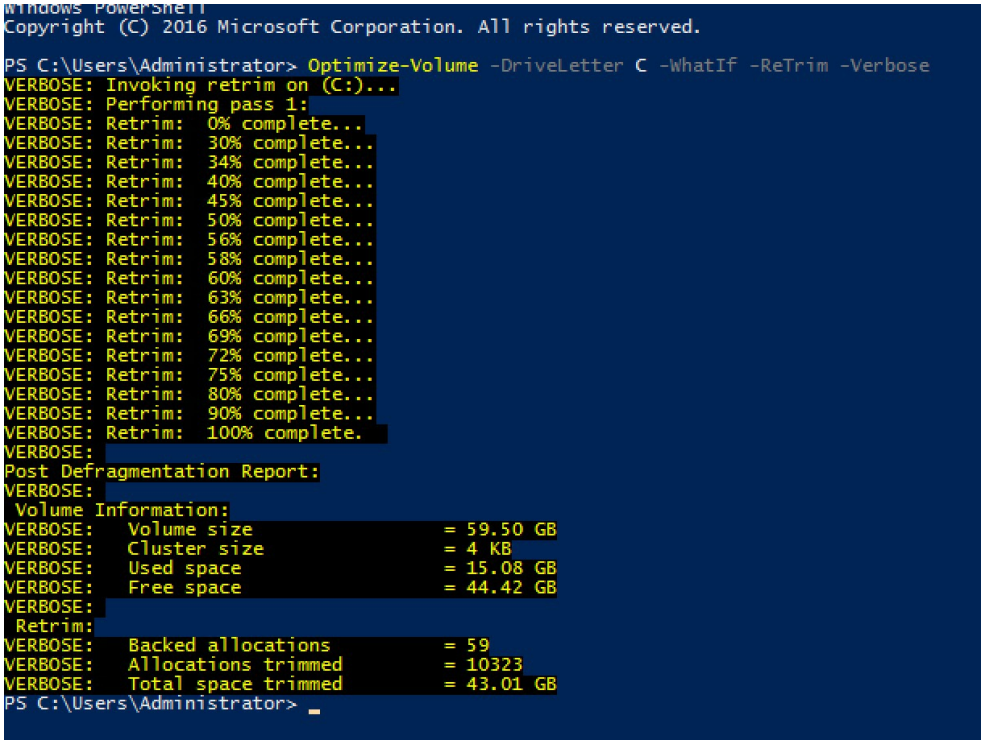
Two methods exist for asynchronously reclaiming space.

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:

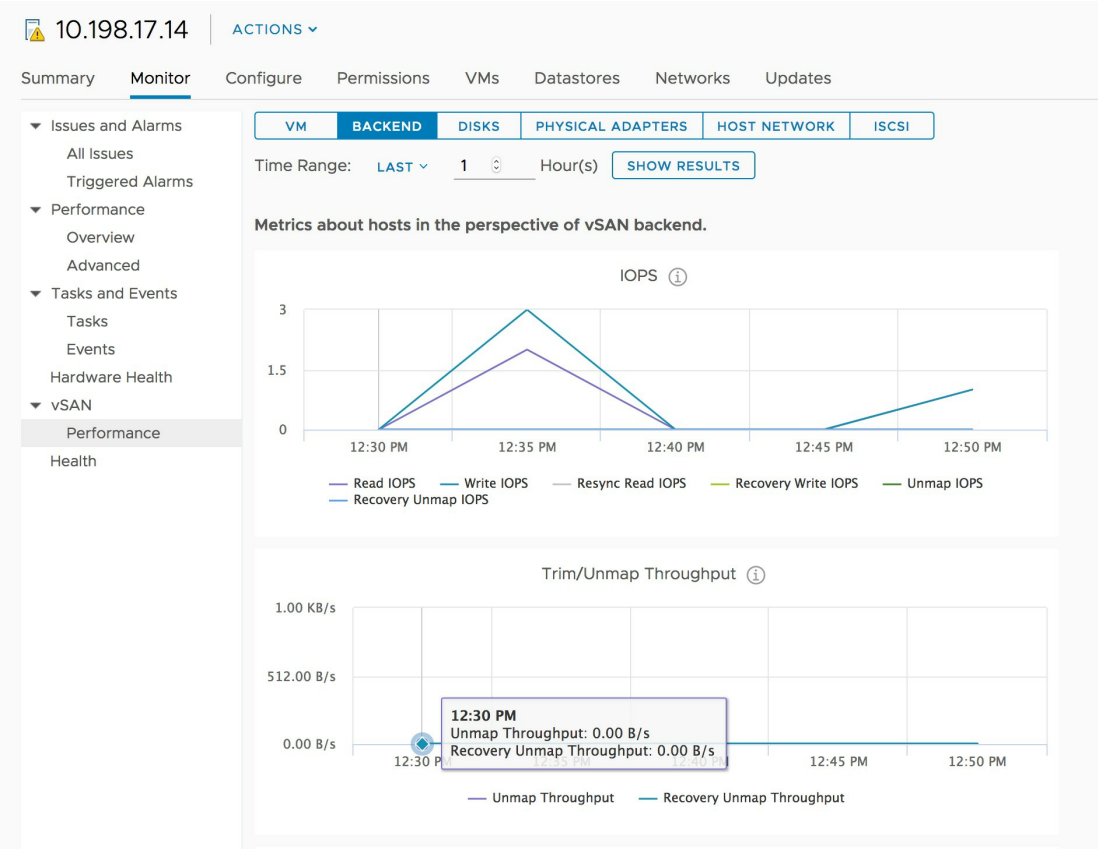
- Windows when using Optimize-Storage or Defrag /L will report sending TRIM commands to all unused blocks. This reporting should not be relied upon up to determine how much space will be reclaimed.
- It is recommended to primarily use the automatic Delete Notification.

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

UNMAP commands will not process through the mirror driver. This means that snapshot consolidation will not commit reclamation to the base disk, and commands will not process when a VM is being storage vMotioned. To compensate for this, the asynchronous running of reclamation performed after the snapshot or migration is performed will be needed to reclaim these unused blocks. This may commonly be seen if using VADP based backup tools that open a snapshot and coordinate log truncation prior to closing the snapshot. One method to clean up before a snapshot is taken is to use the 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.

3. Deduplication

vSAN 6.2 introduces space efficiency features optimized for modern all-flash storage, designed to minimize storage capacity consumption while ensuring availability.

3. 1 Deduplication and Compression

vSAN 6.2 introduces space efficiency features optimized for modern all-flash storage, designed to minimize storage capacity consumption while ensuring availability. One of these features is near-line deduplication and compression. Enabling deduplication and compression can reduce the amount of physical storage consumed by as much as 7x, resulting in a lower total cost of ownership. Environments with highly-redundant data such as full-clone virtual desktops and homogeneous server operating systems will naturally benefit the most from deduplication. Likewise, compression will offer more favorable results with data that compresses well such as text, bitmap, and program files. Data that is already compressed such as certain graphics formats and video files, as well as files that are encrypted, will yield little or no reduction in storage consumption from compression. In other words, deduplication and compression results will vary based on the types of data stored in an all-flash vSAN environment.

Deduplication and compression is a single cluster-wide setting that is disabled by default and can be enabled using a simple drop-down menu. A rolling reformat of all disks in the vSAN cluster is required, which can take a considerable amount of time. However, this process does not incur virtual machine downtime and can be done online, usually during an upgrade. Deduplication and compression are enabled as a unit. It is not possible to enable deduplication or compression individually or for individual virtual machines.

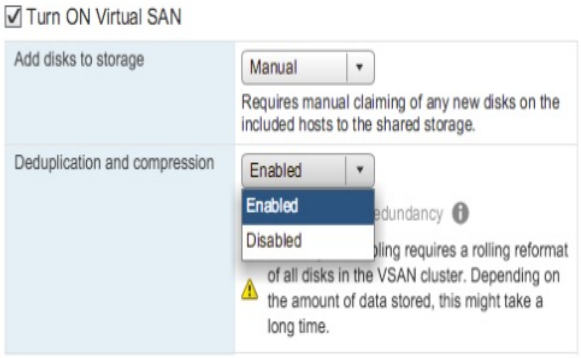


Figure 4. Enabling Deduplication and Compression

Deduplication and compression are implemented after write acknowledgment to minimize impact on performance. Deduplication occurs when data is destaged from the cache tier to the capacity tier of an all-flash vSAN datastore. The deduplication algorithm utilizes a 4K-fixed block size and is performed within each disk group. In other words, redundant copies of a block within the same disk group are reduced to one copy, but

redundant blocks across multiple disk groups are not deduplicated.

The compression algorithm is applied after deduplication has occurred just before the data is written to the capacity tier. Considering the additional compute resource and allocation map overhead of compression, vSAN will only store compressed data if a unique 4K block can be reduced to 2K. Otherwise, the block is written uncompressed to avoid the use of additional resources when compressing and decompressing these blocks which would provide little benefit.

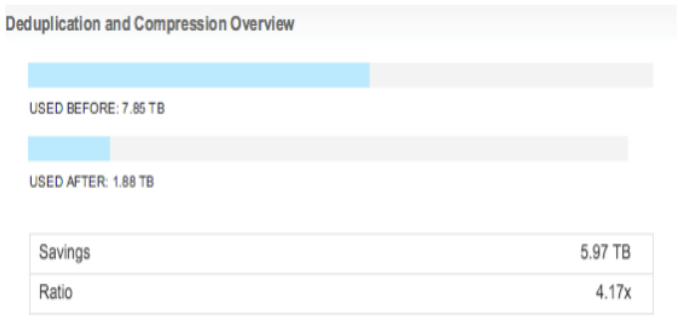


Figure 5. Virtual SAN Deduplication and Compression Overview

Deduplication and compression are applied to data in the capacity tier, commonly accounting for approximately 90% of all data on a vSAN datastore. Storing this data in 4K blocks enables effective deduplication and compression with minimal resource overhead– typically up to 5% - for these operations. Deduplication and compression are not applied to data in the cache tier, which serves as a write buffer in an all-flash vSAN configuration. Naturally, the cache tier is being written to much more frequently than the capacity tier. The resource overhead and performance impact of data deduplication and compression each time a write occurs in the cache tier would far outweigh the benefit for this relatively small amount of data.

The processes of deduplication and compression on any storage platform incur overhead and potentially impact performance in terms of latency and maximum IOPS. vSAN is no exception. However, considering deduplication and compression are only supported in all-flash vSAN configurations, these effects are predictable in the majority of use cases. The extreme performance and low latency of flash devices easily outweigh the additional resource requirements of deduplication and compression.

Enabling deduplication and compression consumes a small amount of capacity for metadata, such as hash, translation, and allocation maps. The space consumed by this metadata is relative to the size of the vSAN datastore and is typically around 5% of the total capacity. Note that the user interface displays the percentage of used capacity, not total capacity (used and free space).

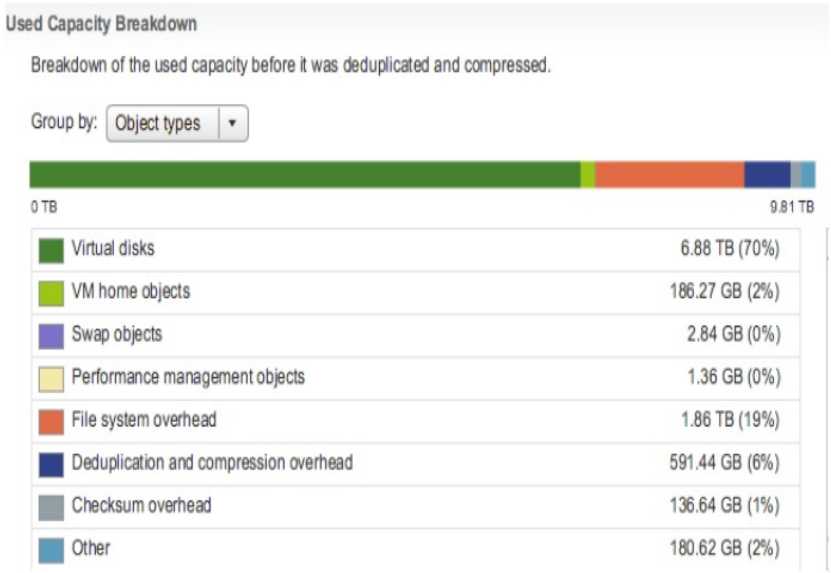


Figure 6. Used Capacity Breakdown

This list provides more details on the object types in the Used Capacity Breakdown view:

- Virtual disks: Virtual disk consumption before deduplication and compression
- VM home objects: VM home object consumption before deduplication and compression
- Swap objects: Capacity utilized by virtual machine swap files
- Performance management objects: When the vSAN performance service is enabled, this is the amount of capacity used to store the performance data
- File system overhead: Capacity required by the vSAN file system metadata
- Deduplication and compression overhead: deduplication and compression metadata, such as hash, translation, and allocation maps
- Checksum overhead: Capacity used to store checksum information
- Other: Virtual machine templates, unregistered virtual machines, ISO files, and so on that are consuming vSAN capacity.

Note : The Object Space Reservation rule affects deduplication and compression space efficiency. If deduplication and compression are enabled, 0% and 100% are the only two values supported for Object Space Reservation. A virtual machine with a storage policy assigned that contains the rule Object Space Reservation = 100% will reserve the entire amount of capacity configured. For example, a 100GB virtual disk will consume 100GB of raw capacity regardless of whether deduplication and compression are enabled or not.

There are pros and cons to using fewer, larger-sized disk groups versus a higher number of smaller-sized disk

groups in a cluster. Since deduplication is performed within each disk group (not across disk groups), fewer and/or larger disk groups will typically yield higher overall deduplication ratios than more/smaller disk groups. However, smaller disk groups have benefits such as the ability to configure more write buffer capacity and less data migration during maintenance operations (e.g. disk replacement). The pros and cons of these configurations should be compared to the workload requirements to determine what option is most appropriate.

Virtual machine snapshots benefit in a few ways with vSAN. Snapshots created on vSAN version 6.0 and higher are created using the vsanSparse format. This method for creating and managing snapshots utilizes thin-provisioned "delta" objects. As the snapshot chain grows, more delta objects are created. This occurs with little or no impact to performance. Furthermore, the base disks and delta disks can be deduplicated and compressed. For more information on vsanSparse snapshots, see vsanSparse-Tech Note for vSAN 6.0 Snapshots .

3. 2 Observations and Recommendations

- Use templates for virtual machine provisioning and apply updates from a central management tool (e.g. Microsoft Windows Server Update Services) to promote uniformity across virtual machines.
- Some database solutions such as Microsoft SQL Server and Oracle Database include native compression. If enabled, this will likely reduce the benefits of vSAN deduplication and compression. For environments where database compression is not enabled, vSAN deduplication and compression should yield more favorable results.
- Online Analytical Processing (OLAP) data warehouses using larger block sizes will most likely show better deduplication and compression results than Online Transaction Processing (OLTP) databases with smaller block sizes.
- Microsoft Exchange Server 2010 and above no longer utilize single instance storage (SIS) to reduced the amount of I/O generated. While this approach increases the size of mailbox databases, it also means deduplication and compression will be more effective versus older versions of Exchange Server.
- Workloads such as TPC benchmarks and streaming media with sustained write throughput tend to fill the write buffer on any storage platform including vSAN. These types of workloads will likely see less benefit from deduplication and compression. Conversely, workloads that produce writes in bursts allow more time for de-staging data from the cache tier to the capacity tier, which is when deduplication and compression are applied in a vSAN environment. Any performance impact from deduplication and compression to these workloads, in nearly all cases, will be less prevalent than the impact to write-intensive workloads.
- Applications that are particularly sensitive to higher latencies and/or a reduction in IOPS such as ERP systems and OLTP applications should be thoroughly tested prior to production implementation to determine if the deduplication and compression space savings is worth a potential impact to performance.
- Generally, read performance will see less of an impact from deduplication and compression. vSAN will first

try to fulfill a read request from the client cache, which resides in host memory. If the data is not available in the client cache, the cache tier of vSAN is queried. Data is not deduplicated and compressed in the client cache or the vSAN cache tier so there is no performance penalty from having to recompile and decompress data before it is delivered. However, reads that come from the vSAN capacity tier will generate a slight amount of resource overhead as the data is recompiled and decompressed.

- Rebuild operations during maintenance or outages will typically require additional compute resources as data is written with deduplication and compression enabled. The impact is less significant in large clusters, or clusters that have multiple disk groups, in combination with larger stripe widths, because writes operations are spread across more write buffers.

4. RAID -5/6

RAID-5/6 erasure coding is a space efficiency feature optimized for all-flash configurations of vSAN 6.2.

4. 1 RAID - 5/6

RAID-5/6 erasure coding is a space efficiency feature optimized for all-flash configurations of vSAN 6.2. Erasure coding provides the same levels of redundancy as mirroring, but with a reduced capacity requirement. In general, erasure coding is a method of taking data, breaking it into multiple pieces and spreading it across multiple devices, while adding parity data so it may be recreated in the event one or more of the pieces is corrupted or lost.

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.

4. 2 Fault Tolerance Method

In vSAN 6.2, a new Storage Policy Based Management (SPBM) rule, Fault Tolerance Method (FTM), has been added. This rule allows the choice of using either mirroring or erasure coding as a data protection method. As mentioned earlier, when using erasure coding data is broken apart and additional redundancy or parity is added for resiliency, and these components are distributed across multiple devices. The choice of FTM rule controls whether data is written in multiple full copies or is spread across multiple partial copies. The level of resiliency and the capacity consumed by an object is a result of the combined FTM rule and Failures to Tolerate (FTT) rule.

Though there are several methods of erasure coding, vSAN 6.2 now supports a RAID-5/6 type data placement and parity pattern as a new method of surviving failures and providing guaranteed space efficiency versus mirroring.

4. 3 Host Requirements

vSAN has historically required a minimum of 3 or more hosts to contribute storage to the distributed datastore in version 5.5 and 6.0 configurations. Version 6.1 added support for as few as 2 hosts contributing storage when using a witness appliance.

Minimum host requirements for mirroring as a failure tolerance method used an algorithm of 2n+1 where an is the number of Failures to Tolerate. For example, an FTT of 1 using the algorithm is 2(1)+1 is 3. This is the basis of the 3 host configuration minimum initially, and is still satisfied with 2 nodes when using a witness appliance.

If a storage policy has the Failure Tolerance Method rule of RAID-5/6 (Erasure Coding), then data is placed in a 3 + 1 arrangement. This can be configured along with an FTT=1 rule or not. This is because the RAID-5/6 (Erasure Coding) rule will default to 3 + 1 data arrangement. This means that either 4 hosts

or 4 fault domains are required for this rule to be satisfied. Any one of the 4 hosts can fail, and all of the data is still present and available, albeit no longer redundant due to the loss of one device.

In the case where the Failure Tolerance Method rule is set to RAID-5/6 (Erasure Coding) combined with FTT=2, data will be placed in a 4 + 2 pattern across hosts. In the event of a host failure, data is still available, and data is still protected from an additional failure. Table 1 details host and fault domain requirements for different Failure Tolerance Methods.

When using the storage policy rule of RAID-5/6 (Erasure Coding) and less than 6 hosts contributing storage are part of a vSAN 6.2 cluster, only a single failure is supported. To be able to protect against dual failures, 6 or more hosts are required.

Failures To Tolerate (FTT)	Fault Tolerance Method (FTM)			
	RAID-1 (Mirroring)		RAID-5/6 (Erasure Coding)	
	Host or Fault Domain Minimum*			
	Required Hosts for policy compliance without failures	Recommended Hosts to allow in-place rebuilds	Required Hosts for policy compliance without failures	Recommended Hosts to allow in-place rebuilds
1	3	4	4	5
2	5	6	6	7
3	7	8		

*When fault domains are not configured hosts behave as individual fault domains

Table 1 Host Requirements Based on Failure Tolerance Method.

Figures 6 and 7 show typical RAID-5/6 (Erasure Coding) object distribution. Like true RAID5 and RAID6 technologies, there is no “parity host” and data parity can land on any capacity device.

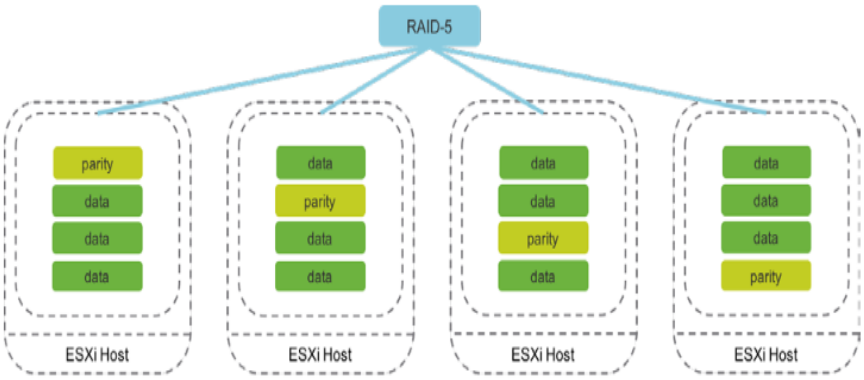


Figure 7. FTT=1 and RAID-5/6 (Erasure Coding)

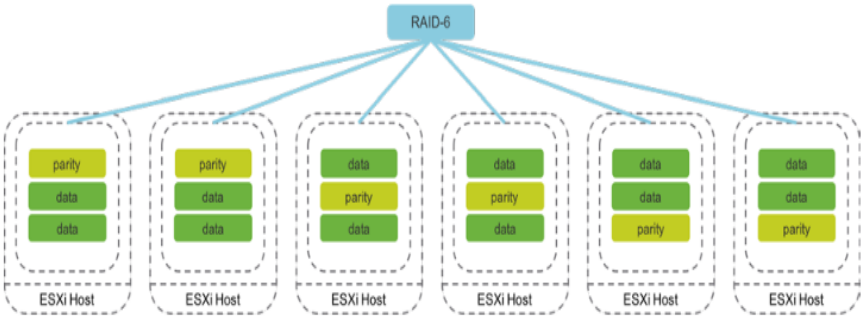


Figure 8. FTT=2 and RAID-5/6 (Erasure Coding)

Due to the host or fault domain requirements of the RAID-5/6 (Erasure Coding) rule, Virtual SAN Stretched Cluster and 2 node configurations are not supported.

Due to the host or fault domain requirements of the RAID-5/6 (Erasure Coding) rule, vSAN Stretched Cluster and 2 node configurations are not supported.

4. 4 Space Savings

Unlike deduplication and compression, which offer variable levels of space efficiency, erasure coding guarantees capacity reduction over a mirroring data protection method at the same failure tolerance. With mirroring, each level of redundancy multiplies the storage requirement by an additional 100%. Surviving from 1 failure requires 2 copies of data at 2x the capacity. vSAN, through the use of the Failures to Tolerate rule, to allow for up to 3

failures.

Erasure coding, because it breaks data apart, does not require the same amount of capacity as mirroring does for the same protection levels. Table 2 shows the amount of space required, depending on the Failure Tolerance Method and Number of Failures to Tolerate.

Use of erasure coding provides significant space savings over mirroring for the same Number of Failures to Tolerate. When allowing for a single failure, RAID-5/6 (Erasure Coding) will consume 33% less capacity than RAID-1 (Mirroring). A 100GB virtual disk, allowing for a single failure will consume 200GB, when mirrored, while the same virtual disk with RAID-5/6 (Erasure Coding) will consume 133GB. The same 100GB virtual disk, allowing for 2 failures, would consume 300GB with the RAID-1 (Mirroring) rule. With the RAID-5/6 (Erasure Coding) rule, that virtual disk would consume 150GB. That is a 50% space savings over mirroring.

Number of Failures to Tolerate (FTT)	Failure Tolerance Method (FTM)		Erasure Coding Space Savings vs. Mirroring
	RAID-1 (Mirroring)	RAID-5/6 (Erasure Coding)	
	Total Capacity Requirement*	Total Capacity Requirement*	
1	2x	1.33x	33% less
2	3x	1.5x	50% less
3	4x	n/a	n/a

Table 2. Space Requirements for FTT and FTM

Regardless of the size of the vSAN datastore, use of erasure coding over mirroring can make a significant impact in the reduction of required storage, while ensuring the same level of data redundancy.

4. 5 Cost of RAID - 5/6

While erasure coding provides significant capacity savings over mirroring, it is important to consider that erasure coding requires additional overhead. This is common among any storage platform today. Because erasure coding is only supported in all-flash vSAN configurations, effects to latency and IOPS are negligible in most use cases due to the inherent performance of flash devices.

Write & Rebuild Overhead

Erasure coding overhead in vSAN is not unlike RAID 5/6 in traditional disk arrays. As new data is written to vSAN, it is sliced up, and distributed to each of the components along with additional parity information. The process of determining an appropriate distribution of data, along with the parity, will consume additional compute

resources. Write latency will also increase somewhat, as whole objects are now being distributed across multiple hosts. In contrast, mirroring always has full copies locally and does no slicing or parity calculation, resulting in less compute overhead and less write latency.

Write operations not only need to be sliced up and distributed with parity, but all the pieces also need to be verified and rewritten with each new write. This is important, because as data is written, it is still necessary to have a uniform distribution of data and parity to account for potential failure and rebuild operations. Writes essentially are a sequence of read and modify, along with recalculation and rewrite of parity. This write overhead occurs during normal operation, and is also present during rebuild operations. As a result, erasure coding rebuild operations will take longer, and require more resources to complete than mirroring.

Converting RAID-5/6 to/from RAID-1

In choosing to convert from a mirroring failure tolerance method, it is necessary to ensure the cluster meets the minimum host or fault domain requirement. The online conversion process adds additional overhead of existing components when applying this policy. VMware recommends to test the impact of converting a smaller set of virtual machines or their objects, before performing the process on larger objects.

4. 6 Observations and Recommendations

Because RAID-5/6 (Erasure Coding) offers guaranteed capacity savings over RAID-1 (Mirroring), any workload is going to see a reduced data footprint. It is importing to consider the impact of erasure coding versus mirroring in particular to performance, and whether the space savings is worth the potential impact to performance.

- Applications that are particularly sensitive to higher latencies and/or a reduction in IOPS such as ERP systems and OLTP applications should be thoroughly tested prior to production implementation.
- Generally, read performance will see less of an impact from erasure coding than writes. vSAN will first try to fulfill a read request from the client cache, which resides in host memory. If the data is not available in the client cache, the capacity tier of vSAN is queried. Reads that come from the vSAN capacity tier will generate a slight amount of resource overhead as the data is recomposed.
- Workloads such as backups, with many simultaneous reads, could see better read performance when erasure coding is used in conjunction with larger stripe count rule in place. This is due to additional read locations, combined with a larger overall combined read IOPS capability. Larger clusters with more hosts and more disk groups can also lessen the perceived overhead.
- Ways to potentially mitigate the effects of the write overhead of erasure coding could include increasing bandwidth between hosts, use of capacity devices that are faster, and using larger/more queue depths. Larger network throughout would allow more data to be moved between hosts and remove the network as a bottleneck.
- Faster capacity devices, capable of larger write IOPS performance, would reduce the amount of time to

handle writes. Additional queue depth space through the use of controllers with larger queue depths, or using multiple controllers, would reduce the likelihood of contention within a host during these operations.

- It is also important to consider that a cluster containing only the minimum number of hosts will not allow for in-place rebuilds during the loss of a host. To support in-place rebuilds, an additional host should be added to the minimum number of hosts.
- It is a common practice to mirror log disks and place configure data disks for RAID5 in database workloads. Because Erasure Coding is a Storage Policy, it can be independently applied to different virtual machine objects, providing simplicity & flexibility to configuring database workloads.

5. Interoperability with Other VMware Solutions

vSAN storage efficiency technologies will yield different levels of benefits on various workloads even within the same product family.

5. 1 Dedupe and Compression Interoperability

vSAN storage efficiency technologies will yield different levels of benefits on various workloads even within the same product family. Deduplication and compression are impressive technologies, but need to be used appropriately for different workloads. By single instancing virtual machine data at the application layer, consumption of compute, disk and network resources can be avoided begin with. There are a few different technologies designed to reduce the storage footprint, such as Linked Clones and now Instant Clones.

VMware Horizon View™

Horizon View has been a top use case for vSAN since the beginning. Customers report amazing performance, great scaling and great cost savings. There are 3 major ways Horizon View is deployed and here is how they will be impacted by these new features.

- Full Clone VDI – Today a legacy choice, and becoming less common as some customers have full clones created (or offloaded through a storage array integration). For these customers, deduplication and compression should yield the most significant savings, but RAID-5/6 should be considered as well to further reduce usage.
- Linked Clone VDI – Commonly used with floating pools, or when recomposes are used for central image updates. Can be combined with profile virtualization (Persona, UEM) to separate the profiles from the desktop so they can survive refreshes to images. They can be more write IO intensive on an ongoing basis than Full Clones due to refreshes and recomposes. Deduplication a compression will yield some benefit (especially on user profile file shares if stored on vSAN) but the single instancing of the base image in the form of the replica, already increases space savings.
- Instant Clone VDI – The next generation of VMware clones. They leverage shared memory and disk for a small foot print more similar to Session Based Computing than Virtual Desktops. Key benefits of instant clones are native space savings, and desktop refreshes that are fast enough for “Just in Time” (JIT) desktops to become an option.
- Space efficiency features will help profile data, but for the desktops themselves testing should be performed to see if the benefits of further space compaction come at the cost of slower desktop refreshes. This feature will enable even denser consolidation ratio’s and object count maximums should be considered before using RAID-5/6.
- iSCSI Volumes - LUNs stored on vSAN exported using the vSAN iSCSI Target Service are fully supported by deduplication and compression.

Other VMware Products

Other VMware solutions that can leverage Linked Clones to reduce the creation of duplicate virtual machine data include VMware vCloud Director, VMware vRealize Automation, and VMware Integrated OpenStack™. In situations where Linked Clones are used, deduplication and compression will result in less overall data efficiency,

than if Full Clones are used.

The VM Encryption feature introduced in vSphere 6.5 is performed at a higher layer than vSAN and as such there will be no expected benefits from VSAN deduplication and compression.

The new vSAN encryption feature introduced in vSAN 6.6 is performed in such a way as to be fully compatible with Deduplication and Compression while still maintaining data at rest encryption. Data is encrypted before it is written to the write buffer device, and decrypted temporarily so that deduplication and compression can run in memory. After this the data is re-encrypted again before being written to the capacity tier.

In situations where vCloud Director, vRealize Automation, and VMware Integrated Openstack workloads are less sensitive to write IO intensive RAID-5/6 (Erasure Coding) can be a viable alternative to RAID-1 (Mirroring) to achieve greater capacity reduction and better overall space utilization.

Be sure to check the VMware Product Interoperability Matrixes for to determine which versions of vSAN are supported with different versions of these, as well as other VMware products.
https://www.vmware.com/resources/compatibility/sim/interop_matrix.php

6. Storage Efficiency Use Case Examples

Storage characteristics vary among different workload types. Some workloads have native data reduction features, while others can be very sensitive to less than predictable conditions.

6. 1 Storage Efficiency Use Case Examples

Storage characteristics vary among different workload types. Some workloads have native data reduction features, while others can be very sensitive to less than predictable conditions.

From a capacity perspective, all workloads are going to see a reduction in space requirements when choosing RAID-5/6 (Erasure Coding) versus RAID-1 (Mirroring). This is a predictable metric given how data is placed on vSAN with the newer Failure Tolerance Method. Again, space savings is not without some tradeoffs around compute resources and the mechanics of erasure coding. Because of this, each Failure Tolerance Method is positioned differently.

If performance is the most important requirement, customers will choose RAID-1 (Mirroring) and if capacity, with some additional resource requirements is important, then customers will choose RAID-5/6 (Erasure Coding). Workloads that already employ some type of data reduction technique may not see a significant difference in total capacity consumed, and therefore customers may choose not to select erasure coding. This is also true when choosing to enable deduplication and compression on a vSAN cluster. Individual experiences may vary and recommendations are general guidelines to follow, that may or may not align with observations, depending on a variety of contributing factors.

While erasure coding provides a predictable amount of space savings, deduplication and compression provide a varying amount of capacity reduction. In workloads that have a significant amount of redundant or compressible data, deduplication and compression can provide excellent data reduction. Some workloads that will see larger deduplication and compression rates can include Full Clone virtual machines, databases without compression or without compression enabled, file servers with common and compressible files, and the like. Deduplication and compression may not be as effective when used with workloads that have native data reduction techniques. Databases with native compression like Microsoft SQL Server and Oracle Database already provide efficient data reduction and therefore may see less additional capacity reduction.

In addition to data reduction, deduplication and compression along with erasure coding will introduce the requirement for additional compute resources and additional time for the data reduction and de-staging process. Workloads that are latency sensitive may not be ideal candidates for both space efficiency techniques being simultaneous or individually leveraged. Alternatively, workloads that are less sensitive to latency may see little or virtually no impact. Results will vary based on the workload requirements, physical configuration, cluster settings, and policies applied.

Table 3 lists some common workload types as well as some general guidelines when used in conjunction with the new space efficiency features of vSAN 6.2.

	Recommendation							
Deduplication & Compression	Disabled		Disabled		Enabled		Enabled	
Failure Tolerance Method	RAID-1 (Mirroring)		RAID-5/6 (Erasure Coding)		RAID-1 (Mirroring)		RAID-5/6 (Erasure Coding)	
○ Least Benefit ◔ Limited Benefit ◑ Some Benefit ◕ More Benefit ● Most Benefit	Performance	Capacity	Performance	Capacity	Performance	Capacity	Performance	Capacity
DB Natively Compressed	●	○	◕	◔	◑	◔	○	◔
DB Uncompressed	●	○	◕	◕	◑	◑	○	●
OLAP	●	○	◕	◑	◑	◑	○	◕
OLAP (Big Block)	●	○	◕	◕	◑	◑	○	●
OLTP (Small Block)	●	○	◕	◔	◑	◑	○	◑
Exchange 2010+	●	○	◕	●	◑	◑	○	●
Real Time Analytics	●	○	◕	◑	◑	◑	○	◑
Read Intensive & Minimal Writes	●	○	◕	◕	◑	◑	○	◕
VDI Full Clone	●	○	◕	●	◑	◑	○	●
VDI Linked Clone	●	○	◕	◑	◑	◑	○	◑
VDI Instant Clone	●	○	◕	◑	◑	◑	○	◑
Hadoop	●	○	◕	◔	◑	◕	○	◑

Table 3 Space Efficiency Use Case General Guidelines

7. Summary

HyperConverged solutions featuring all-flash storage are the future as they continue to decrease in cost and offer dramatically better performance when compared to magnetic disks.

7. 1 Summary

HyperConverged solutions featuring all -flash storage are the future as they continue to decrease in cost and offer dramatically better performance when compared to magnetic disks. vSAN is optimized for modern all -flash storage with efficient near-line deduplication, compression, and erasure coding capabilities that lower TCO. Deduplication and compression enable considerable capacity savings across the entire vSAN cluster — especially in environments where standard OS builds (templates and clones) are used and where there is abundant data commonality such as file shares. RAID-5/6 erasure coding reduces capacity consumption by as much as 50% versus RAID-1 (mirroring) with the same levels of availability for FTT=1 and FTT=2 vSAN rule sets. These rule sets can be assigned to individual virtual machines and virtual disks, which provides precise availability protection in a hyperconverged infrastructure. As with any solution offering space efficiency technologies, there is an overhead cost in the form of storage and compute resources. Potential performance impact from this overhead is minimized by vSAN's unique architecture along with the use of flash devices. However, application testing is always recommended prior to production implementation to verify business requirements are met and the space savings justifies the additional resource cost.

About the Authors

John Nicholson is a Senior Technical Marketing Manager in the Storage and Availability Business Unit. He focuses on delivering technical guidance around VMware vSAN solutions. John previously worked in architecting and implementing enterprise storage and VMware solutions.

Follow John on Twitter: @Lost_Signal

Jase McCarty is a Senior Technical Marketing Architect at VMware with a focus on storage solutions. He has been in the Information Technology field for over 25 years, with roles on both the customer and vendor side. Jase has Co-Authored two books on VMware virtualization, routinely speaks technology focused user group meetings, and has presented at VMworld and EMC World.

Follow Jase on Twitter: @jasemccarty

Jeff Hunter is a Senior Technical Marketing Architect at VMware with a focus on storage and availability solutions. He has been with VMware for more than 8 years, prior to which he spent several years implementing and administering VMware software-defined data centers at two Fortune 500 companies. Jeff has presented at multiple technology conferences including VMworld and EMC World.

Follow Jeff on Twitter: @jhuntervmware