VMware vSAN Design Guide

VMware Storage

Table of contents

VMware vSAN Design Guide

Introduction

VMware® vSAN™ is a hyperconverged, software-defined storage platform fully integrated with VMware vSphere®.

New to the Guide is content for the Express Storage Architecture™ in vSAN 8 and vSAN 8 U2. vSAN ESA is supported on ReadyNodes that can be customized For information on customization of these ESA ReadyNodes see this blog and additional guidance on vSAN ESA Readynode emulated designs. Design Guidance specific to the Original Storage Architecture (OSA) will be marked with an ***(OSA)*** within this guide. For A quick update on what is different with vSAN 8 ESA, see this blog or check out this video:

Overview

VMware® vSAN™ is a hyper-converged, software-defined storage (SDS) platform that is fully integrated with VMware vSphere®. vSAN aggregates locally attached disks of hosts that are members of a vSphere cluster to create a distributed shared storage solution. vSAN activates the rapid provisioning of storage within VMware vCenter™ as part of virtual machine creation and deployment operations. vSAN is the first policy-driven storage product designed for vSphere environments that simplifies and streamlines storage provisioning and management. Using VM-level storage policies, vSAN automatically and dynamically matches requirements with underlying storage resources. With vSAN, many manual storage tasks are automated - delivering a more efficient and cost-effective operational model.

vSAN provides two different configurations: a hybrid configuration that leverages flash-based devices and magnetic disks and an all-flash configuration. The hybrid configuration uses server-based flash devices to provide a cache layer for optimal performance, while magnetic disks provide capacity and persistent data storage. This delivers enterprise performance and a resilient storage platform. The all-flash configuration uses flash for both the caching and capacity layers.

Introduced in vSAN 8, and enhanced in vSAN 8 U1, the Express Storage Architecture uses a powerful new all-NVMe storage design.

This document focuses on helping administrators to correctly design and size a vSAN cluster and answer some of the common questions around the number of hosts, number of disk groups, cache sizing, and number of capacity devices, along with detailed configuration questions to help correctly and successfully deploy vSAN.

There are three ways to build a vSAN cluster:

- Turn-key deployment using appliances such as Dell EMC VxRail, Lenovo ThinkAgile VX Series, and Hitachi UCP-HC
- Certified vSAN ReadyNodes from any of the leading server OEMs This is the only current option for vSAN ESA, using explicitly certified ReadyNodes for ESA.
- Custom-built using components from the VMware Compatibility Guide for vSAN.

A vSAN Ready Node is a validated server configuration in a tested, certified hardware form factor for vSAN deployment, jointly recommended by the server OEM and VMware. vSAN Ready Nodes are ideal as hyper-converged building blocks for larger data center environments looking for automation and needing to customize hardware and software configurations. Select vSAN Ready Node partners provide pre-installed vSAN on ready nodes.

vSAN Design Overview

This document was formerly the vSAN Design and Sizing Guide. With the changes in recent editions of vSAN, it is encouraged that all vSAN sizing go through the vSAN Sizing tool. This tool also includes a "Quick sizer" that allows for reverse sizing calculations. This tool has been updated to cover "reverse sizing" (Working backwards from a bill of materials to deliver usable resources) as well as support vSAN Express Storage Architecture.

When comparing cluster specifications in an OSA cluster versus an ESA cluster, always use the standard ReadyNode Sizer. The QuickSizer will not be able to accurately reflect the lower total cost of ownership that the ESA provides. For more information, see: vSAN 8 Total Cost of Ownership.

Adhere to the VMware Compatibility Guide (VCG)

There are a wide range of options for selecting a host model, storage controller as well as flash devices and magnetic disks, especially for the OSA. It is extremely important that you follow the VMware Compatibility Guide (VCG) to select these hardware components. This on-line tool is regularly updated to ensure customers always have the latest guidance from VMware available to them. A subset of the VCG, the vSAN VCG must also be consulted for for storage devices and controllers. For a list of vSAN ESA compatible storage devices see this vSAN VCG link.

The ESA is much easier to design for, as it is able to offer performance and efficiency capabilities relatively easily. Specifying server configurations through the ReadyNode program for ESA also provides for a simple, yet prescriptive approach to design. Recommendations for optimal performance in the ESA are also much easier than in the OSA.

Hardware, drivers, firmware

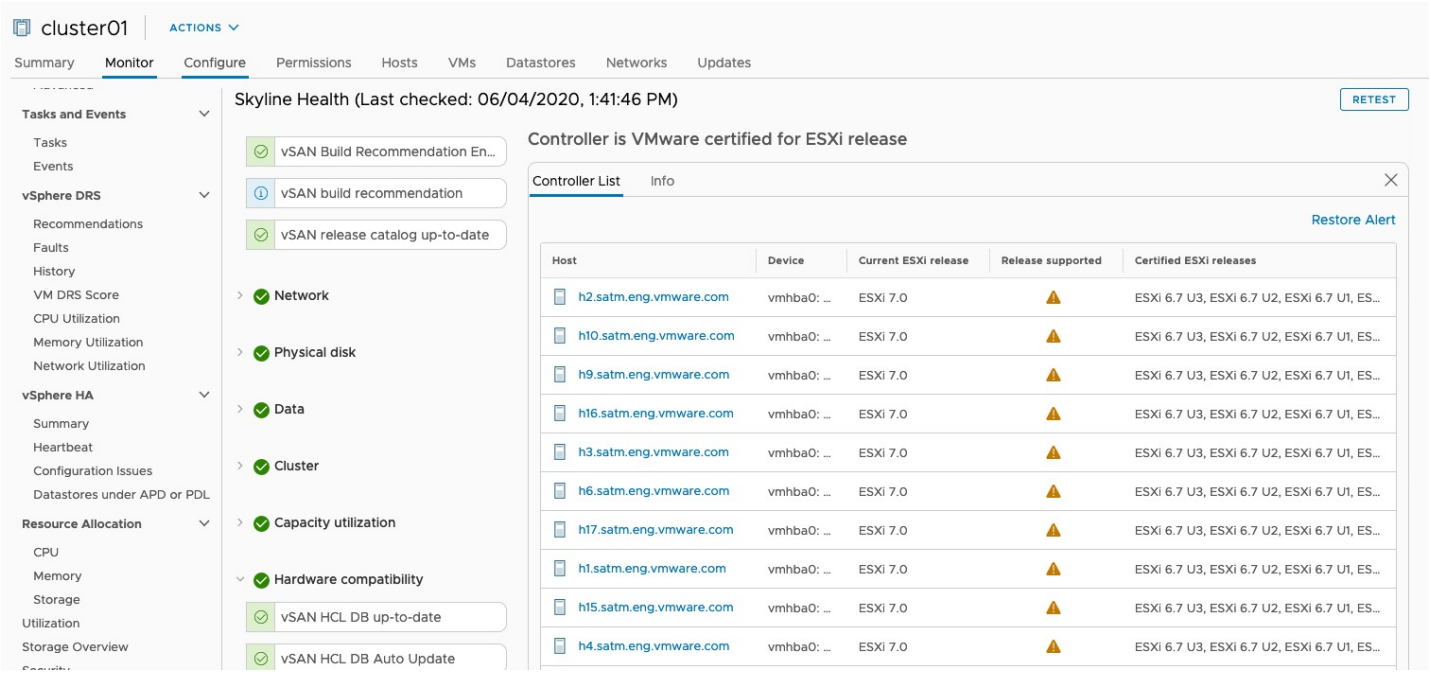
The vSAN VCG makes very specific recommendations on hardware models for storage I/O controllers, solid state drives (SSDs), PCIe flash cards, NVMe storage devices and disk drives. It also specifies which drivers have been fully tested with vSAN, and in many cases – identifies minimum levels of firmware required. For SSDs the minimum version is specified. For Controllers and NVMe drives the exact version supported is specified. Ensure that the hardware components have these levels of firmware, and that any as associated drivers installed on the ESXi hosts in the design have the latest supported driver versions. The vSAN health services will detect new versions of drives and firmware for controllers.

RDMA Support

vSAN 7 Update 2 Introduced support for RoCE v2. For supported NICs please see the vSAN VCG, and confirm that the NIC is explicitly supported for vSAN RDMA usage.

Best practice: Always verify that VMware supports the hardware components that are used in your SAN deployment.

Best Practice: Verify all software, driver and firmware versions are supported by checking the VCG and using the vSAN Health Service. The screen shot below shows an example of a controller driver that is not on the VCG.



Balanced Configurations

Balanced Configurations Considerations

As a recommended practice, VMware recommends deploying ESXi hosts with similar or identical configurations across all cluster members, including similar or identical storage configurations. This will ensure an even balance of virtual machine storage components across the disks and hosts cluster. While hosts that do not contribute storage can still leverage the vSAN datastore, having a cluster with fewer nodes contributing storage increases the capacity and performance impact when a node is lost. For this reason, VMware recommends balanced configurations within a cluster.

VMware vSAN HCI Mesh

HCI Mesh supports the mounting of vSAN clusters by external clusters. This can allow for asymmetric compute scaling, as well as help increase storage utilization between clusters. For more information see the updated HCI Mesh Technote.

Asymmetric Configurations Considerations

If the components are no longer available to purchase try to add equal quantities of larger and faster devices. For example, as 400GB SSDs become more difficult to find, adding a comparable in performance 800GB SSD should not negatively impact performance. To ensure levels of performance consistency, it is ideal, but not required to have hosts using storage devices with similar levels of performance capabilities. Mixing flash devices that using different bus protocols, or interfaces (SATA, SAS, NVMe) within or across hosts that comprise a vSAN cluster will generally only provide performance levels as fast as the slowest devices used (e.g. SATA). If one is transitioning existing hosts to newer, faster devices such as NVMe, these performance gains will likely only be realized when all of the slower devices in the hosts that comprise a cluster have been replaced with newer, faster devices.

New generations of servers can be mixed, but it is recommended to try to keep storage configurations balanced when possible. Be aware that EVC may need to be activated. Mixing of generations can also be used as a process for gradually migrating. Different servers and vendors can be mixed but it may add complexity to the management of the lifecycle on the hosts.

For more information on asymmetric vSAN configurations see the following podcast and blog post: Asymmetrical vSAN Clusters - What is Allowed, and what is Smart.

Host Rebuild Reservation (HRR)

If host Rebuild Reservation is activated, it will base the storage reservation based on the assumption that the largest node within the cluster has failed. For clusters running the vSAN ESA in vSAN 8 U1, where the "Auto-Policy Management" capability is enabled, this may impact the effective storage policy that one can use for a given cluster. For more information, see the post: "Auto-Policy Management Capabilities with the ESA in vSAN 8 U1."

Best practice: Consider alternative solutions for asymmetric demand needs. Single socket servers can help with storage heavy workloads while deploying hosts with empty drive bays activates adding storage later on without the need to add additional nodes. Strategic approaches to purchasing can help.

Increasing cache and capacity within existing nodes

vSAN provides customers with a storage solution that is easily scaled up by adding new or larger disks to the ESXi hosts, and easily scaled out by adding new hosts to the cluster. For clusters running the vSAN OSA, it is important to scale in such a way that there is an adequate amount of cache, as well as capacity, for workloads.

Scaling Out

This allows customers to start with a very small environment and scale it over time, by adding new hosts and/or more disks. In many cases, scaling out by adding additional hosts to a VSAN cluster is preferred over adding or replacing drives in an existing host. Adding a host is non-disruptive as shown in this click-through demonstration: Scale Out by Adding a Host.

Scaling Up

Adding additional drives into the existing hosts in a cluster can be a fast way to scale up capacity within the nodes in a cluster. This can be done by expanding existing disk groups by adding capacity devices, adding new disk groups entirely, or replacing existing devices if additional drive bays are not avalible.

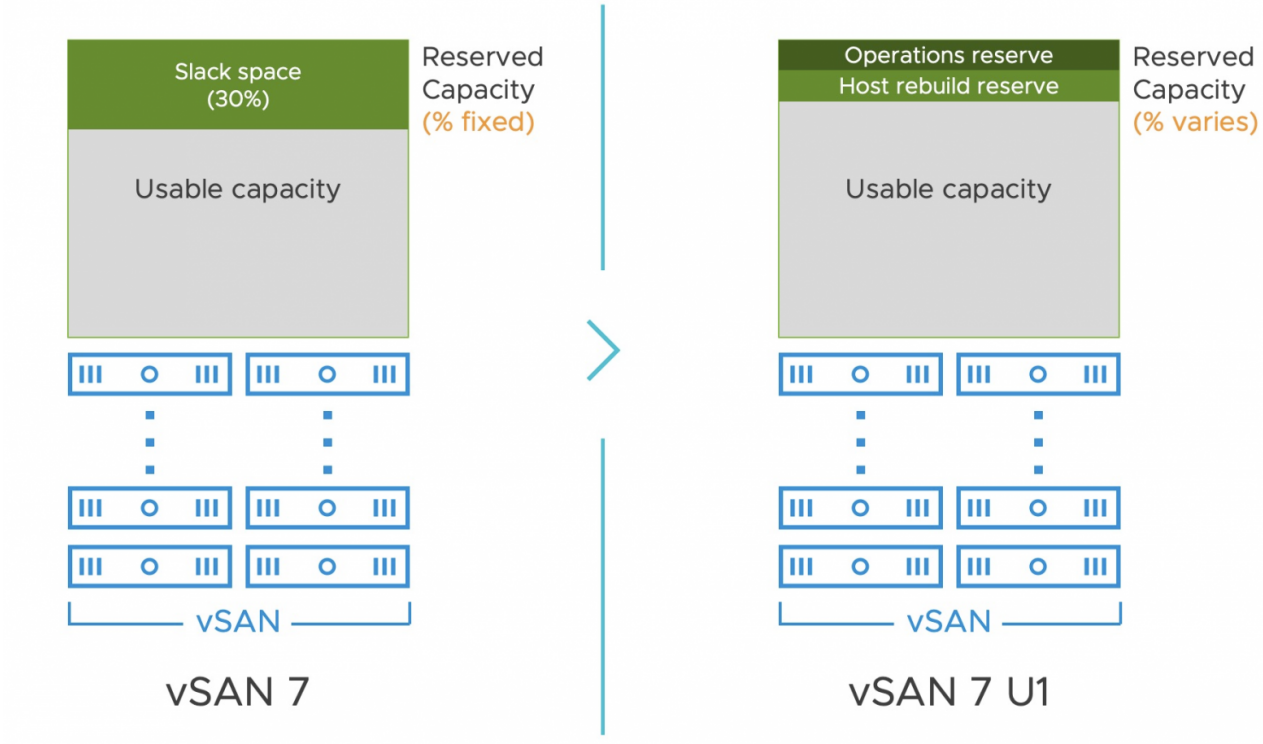
This consideration is covered in depth throughout this guide. In particular, one should consider choosing hosts for a design that have additional disk slots for more capacity, as well as providing an easy way to install additional devices into these slots.

Note that increasing cache capacity requires the removal of the disk group (evacuation of data), replacing the existing cache device with the new one, and then adding the disk group back to the vSAN configuration. This is because a disk group can have only one cache device. In the case of vSAN ESA, adding additional drives to a host will increase both capacity and performance.

Best practice: Scaling out by adding one or more additional hosts to a vSAN cluster is preferred over replacing or adding new disks to existing hosts. This blog has a detailed list of considerations when reviewing Scale Up vs. Scale Out considerations.

Designing for Capacity Maintenance and Availability

Prior to vSAN 7 U1, a 25-30% slack space recommendation was typically used to account for failures as well as maintenance activities. This guidance was based on approximations, and while many clusters may have used less than this on average, there was no way to guarantee proper operation in all circumstances if the recommendation was lowered. vSAN did not have the internal mechanisms to limit usage for these operations, which could lead to cluster full conditions if a user exceeded the free space recommendation.



vSAN 7 U1 made substantial improvements in the free capacity required for vSAN for to handle conditions such as host failures, and internal vSAN operations. Internal optimizations were made to vSAN to dramatically improve its usage of free capacity, and improve the TCO of the cluster. The generalized cluster recommendation of 25-30% of free capacity is replaced with calculations that factor in cluster host count, cluster settings, and other considerations. For the most current guidance on this sizing, size use the vSAN Sizing Calculator.

New Terminology

- Operations reserve: The spaced reserved for vSAN internal operations (resync, rebuilds, etc.)
- Host Rebuild Reserve: The space reserved for rebuilding one host's worth of capacity, should there be a sustained outage of a host.
- Reserved Capacity: The total of operations reserve and host rebuild reserve.
- Operations Usage: The space being used temporarily in the "Operations Reserve" location for the purposes of resyncs, rebuilds, etc.

Private cloud topologies such as stretched clusters, 2-node clusters, and clusters using fault domains are not capable of using the Reserved Capacity mechanism at this time. In those configurations, customers should size for the amount of free capacity based on guidance for previous versions of vSAN.

Example of New Sizing

The amount of free capacity required for both "operations reserve" and "host rebuild reserve" is based on several variables. Host rebuild reserve is based off of N+1, and thus, it and the total reserved capacity will decrease with cluster scale. Here is an example of hosts with 2 disk groups per host, consisting of 10, 1.92 TB capacity devices, DD&C activated, where both Operations Reserve (OR) activated, and Host Rebuild Reserve (HBR) activated.

- 4 node cluster: 10% OR + 25% HRR = Approx 30% total reserved capacity
- 6 node cluster: 10% OR + 17% HRR = Approx 27% total reserved capacity
- 8 node cluster: 10% OR + 13% HRR = Approx 23% total reserved capacity
- 12 node cluster: 10% OR + 8% HRR = Approx 18% total reserved capacity
- 18 node cluster: 10% OR + 6% HRR = Approx 16% total reserved capacity
- 24 node cluster: 10% OR + 4% HRR = Approx 14% total reserved capacity
- 32 node cluster: 10% OR + 3% HRR = Approx 13% total reserved capacity
- 48 node cluster: 10% OR + 2% HRR = Approx 12% total reserved capacity
- 64 node cluster: 10% OR + 2% HRR = Approx 12% total reserved capacity

Note that these are **simply examples based on typical deployment variables**. They **do not** apply to all conditions. **Actual percentages will vary**, and are dependent on the configuration of the cluster. The OR will vary depending on the variables described above, but for this example, they remain the same, and only the host count is changed.

For more information and examples on reserved capacity in vSAN, see the post: "Understanding Reserved Capacity Concepts in vSAN."

Maintenance Considerations

vSAN 7 U1 introduced a method that upon a host entering into maintenance mode (EMM) using the "Ensure Accessibility" option, it will allow vSAN to write all incremental updates to another host in addition to host holding the object replica. This helps ensure the durability of the changed data in the event that the one host holding the updated object replica failed during this maintenance window. In cases of an unrecoverable failure of the host with the current object replica, the changed data can be merged with the replica residing on the host that originally in maintenance mode so that an up-to-date object replica is readily available. This can help drive better efficiency for customers who previously would use FTT=2 to better ensure data durability during maintenance mode events, as FTT=1 using vSAN's enhanced durability is a more efficient alternative to this specific scenario.

vSAN 8 Update 1 extended this behavior for planned maintenance to vSAN Express Storage Architecture.

vSAN 7 Update 2 extended this behavior to unplanned events (Ex. host failure) for vSAN Original Storage Architecture.

Partial Repair Considerations

vSAN uses a concept referred to as "partial repairs." Previous editions of vSAN would only be able to successfully execute a repair effort if there were enough resources to repair all of the degraded or absent components in their entirety. The repair process in vSAN will take a more opportunistic approach to healing by repairing as many degraded, or absent components as possible, even if there are insufficient resources to ensure full compliance. The effective result is that an object might remain non-compliant after a partial repair, but will still gain increased availability from those components that are able to be repaired.

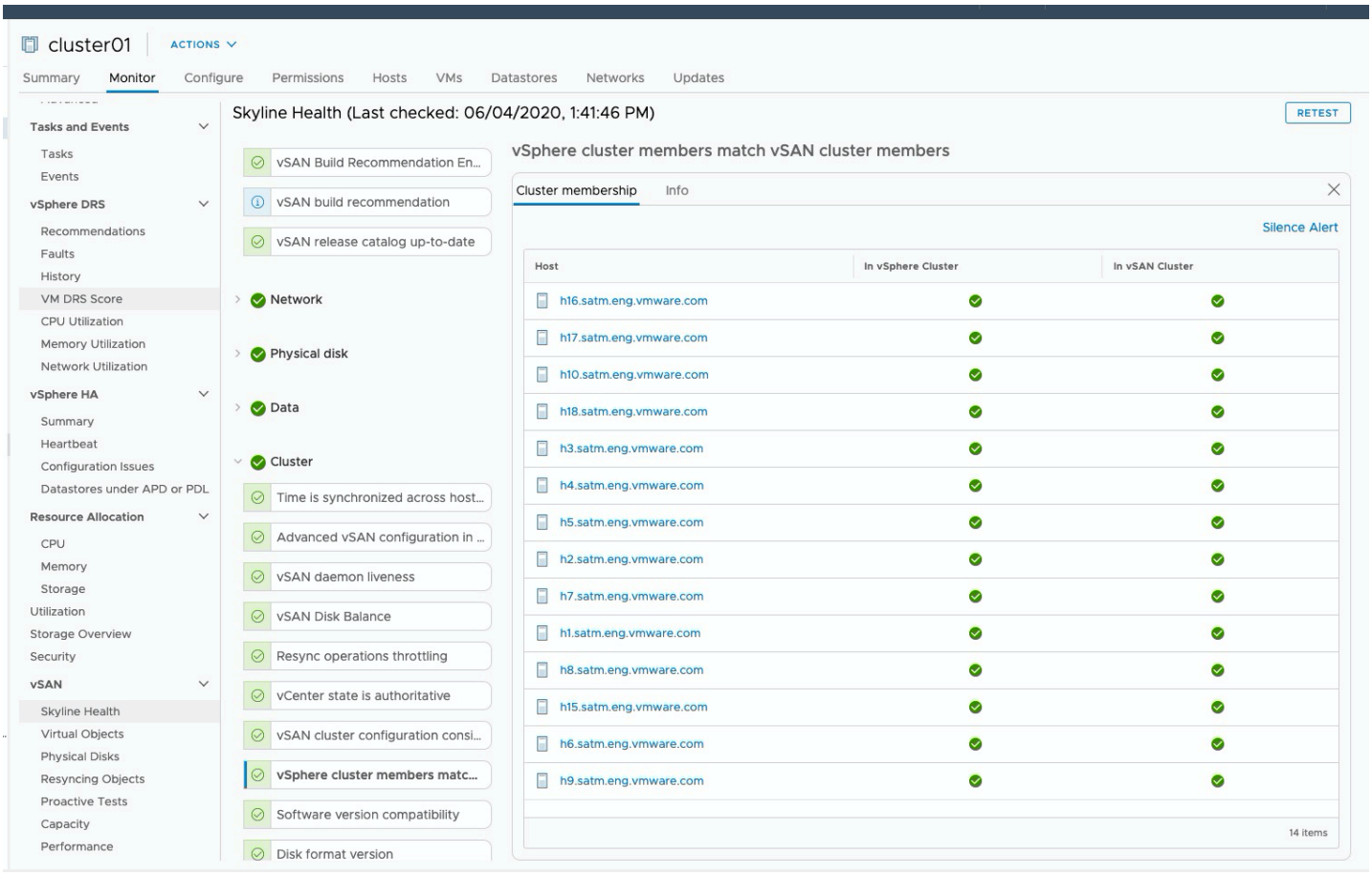
One additional consideration is the size of the capacity layer. Since virtual machines deployed on vSAN are policy-driven, and one of those policy settings (*NumberOfFailuresToTolerate*) will make a mirror copy of the virtual machine data, one needs to consider how much capacity is required to tolerate one or more failures. This design consideration will be discussed in much greater detail shortly.

Design decision: N+1 where N equals the minimum number of hosts or fault domains for compliance offers the option to allow for quick re protection of data. Ensure there is enough storage capacity and fault domains to meet the availability requirements and to allow for a rebuild of the components after a failure.

vSAN Support Insight with Skyline

vSAN includes the vSAN Skyline Health Service. This complimentary feature available to all vSAN license versions regularly checks a range of different health aspects of the vSAN cluster and provides insight into the cause of many potential vSAN issues. Once an issue is detected, the Health Service highlights the problem and, in most cases, directs administrators to the appropriate VMware Knowledge Base article for guidance on correcting the issue. Version 6.6 introduced online health checks that allow for this functionality to be updated without the need to update ESXi, and telemetry to be provided to support staff using vSAN Support Insight phone home system.

Design Decision: Verify the vSAN Health Service is activated. Do not proceed with adding workloads to a vSAN datastore until all items are showing as "Passed" with green check marks.



Use Supported vSphere Software Versions

Verify that the vSphere components in your environment meet the software requirements for using vSAN. To use the full set of vSAN 7 and later capabilities, the ESXi hosts and vCenter Server must be on vSphere 7. VMware continuously fixes issues encountered by customers, so by using the latest version of the software, customers avoid encountering issues that have already been fixed.

Best Practice: Ensure that the latest patch/update level of vSphere is used when doing a new deployment, and consider updating existing deployments to the latest patch versions to address known issues that have been fixed.

All-Flash Considerations

vSAN 8 Express Storage Architecture (ESA)

vSAN 8 introduced support for an all NVMe Express Storage Architecture. Using certified ReadyNodes will ensure that CPU, memory, NVMe device, networking connectivty requirements have already been met. Please use the vSAN ReadyNode selection tool and vSAN sizing tools to identify the correct configuration of hosts to meet your requirements. With vSphere 8 Update 2, feature parity with OSA has now been achieved. It is strongly encouraged for all net new clusters designs to be vSAN ESA instead of OSA. Performance, cost and TCO will be superior.

vSAN Original Storage Architecture (OSA)

All flash vSAN OSA remains for brownfield clusers, and customers unable to deploy vSphere 8 for new clusters at this time. There are some considerable differences with using an all-flash version when compared to the hybrid version which is covered in detail in each applicable section.

All-flash vSAN configuration brings improved, highly predictable and uniform performance regardless of workload as compared to hybrid configurations. All-flash also supports RAID-5/6 erasure coding fault tolerance methods for resiliency. Deduplication and compression can be activated for all-flash configurations to minimize raw capacity consumption. More information can be found regarding these features in the vSAN Space Efficiency Technologies guide.

vSAN All-flash configurations:

- Require a 10Gb network
- Allow a maximum number of 64 nodes/hosts
- Use flash devices for both cache and capacity
- Does not utilize cache devices for reads as these are served directly from the all-flash capacity tier (unless the block has not been destaged yet, in that situation it comes from cache)
- Utilize higher endurance, lower capacity flash devices for the cache tier (write buffer) and lower endurance, higher-capacity flash devices for the capacity tier

Summary of Design Overview Considerations

- For new clusters deploying vSphere 8, vSAN ESA should be preferred
- When using the Express Storage Architecture in vSAN 8 , please make sure proper design principals are followed.
- Ensure that all the hardware used in the design is supported by checking the VMware Compatibility Guide (VCG)
- Ensure that all software, driver and firmware versions used in the design are supported by checking the VCG
- Avoid unbalanced configurations by using similar configurations in a cluster
- Design for growth. Consider initial deployment with capacity in the cluster for future virtual machine deployments, as well as enough flash cache to accommodate future capacity growth.
- When adding capacity to a vSAN cluster, scaling out by adding hosts is the preferred method.
- Design for availability. Consider designing with more than three hosts and additional capacity that activate the cluster to automatically remediate in the event of a failure
- Verify the vSAN Health service is activated. Resolve any issues highlighted by the health service prior to adding workloads to a vSAN datastore.

- Ensure that the latest patch/update level of vSphere is used when doing a new deployment, and consider updating existing deployments to the latest patch versions to address known issues that have been fixed

vSAN Limits

These are vSAN constraints that must be taken into account when designing a vSAN cluster.

ESXi Host and Virtual Machine Limits

vSAN ESA with 8 Update 2 supports up to 500 virtual machines per host. For guidance on if you should consider scaling to this, and why the limit was raised from the 200 limit that exists still for vSAN OSA see this blog.

There are vSAN configuration limits that impact design and sizing. Refer to "Configuration Maximums" in the vSphere documentation on VMware Docs . Note that vSAN stretched clusters have unique limitations. Refer to the vSAN documentation on VMware Docs for more information.

Design decision: vSAN clusters with four or more nodes provide greater flexibility. Consider designing clusters with a minimum of four nodes where possible.

VM Storage Policy Maximums

The VM storage policies impact sizing and are discussed in detail later in the guide.

Design decision: Ensure there are enough physical devices in the capacity layer to accommodate a desired stripe width requirement.

Design decision: Ensure there are enough hosts (and fault domains) in the cluster to accommodate a desired *NumberOfFailuresToTolerate* requirement.

Maximum VMDK Size and Component Counts

The maximum VMDK size on a vSAN datastore is 62TB.

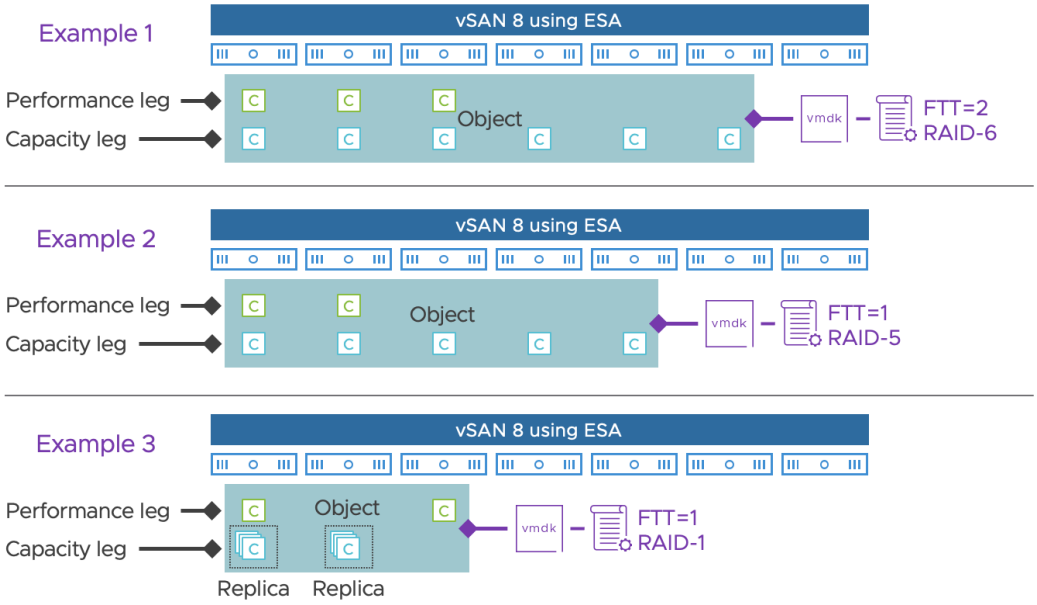
Virtual machines deployed on vSAN are made up of a set of objects. For example, a VMDK is an object, a snapshot is an object, VM swap space is an object, and the VM home namespace (where the .vmx file, log files, etc. are stored) is also an object. Each of these objects is comprised of a set of components, determined by capabilities placed in the VM Storage Policy. For example, if the virtual machine is deployed with a policy to tolerate one failure, then objects will be made up of two replica components. If the policy contains a stripe width, the object will be striped across multiple devices in the capacity layer. Each of the stripes is a component of the object. The concepts of objects and components will be discussed in greater detail later on in this guide, but the limits impacting the sizing are shown in the table below:

As mentioned in the previous section, objects are still striped at 255GB in vSAN 6.x. If an administrator deploys a 62TB object, then there will be approximately 500 components created, assuming a default policy of *NumberOfFailuresToTolerate* = 1. When creating very large VMDKs on vSAN, component maximums need to be considered.

As of vSAN OSA in 7U1 and newer The use of Objects over 2TB will have an effective maximum stripe size of 3 for components past the initial 2TB of capacity. This will lower the risk of exhausting the maximum number of components. Also, The use of Objects over 8TB will marginally increase the operational reserve required. This overhead is factored in by default into the vSAN

Sizer.

vSAN ESA increases the maximum component count to 27,000. The maximum virtual machine count has been raised to 500 virtual machines. for more guidance on if you should deploy this many virtual machines per host, see this blog.



Permanence Leg - For RAID-5, it will be a 2-way mirror and always consist of at least 2 components. For RAID-6 it will be a 3-way mirror, and always consist of at least 3 components.

Capacity leg - Prescribed by selected storage policy. For RAID-6 (4+2) it will always consist of at least 6 components. For RAID-5 (4+1), it will always consist of at least 5, and for RAID-5 (2+1) it will always consist of at least 3 components.

RAID 1/0 - The data structure of a performance/metadata leg and a capacity leg will also exist for objects using RAID-1 mirroring and RAID-0 objects, but the purpose of the new data structure is to eliminate the need for RAID-1 mirroring for all conditions except host constrained conditions (3-node and 2 node clusters).

vSAN ESA Snapshots - Snapshots will no longer be split out as their own discrete objects and components.

Summary of vSAN Limits Design Considerations

- Enabling vSphere HA on a vSAN cluster is highly recommended to minimize downtime due to host failure. For vSphere maximums see Configmax.
- Consider the number of hosts (and fault domains) needed to tolerate failures.
- Consider the number of devices needed in the capacity layer to implement a desired stripe width.
- Consider component count, when deploying very large virtual machines. It is unlikely that many customers will have requirements for deploying multiple 62TB VMDKs per host. Realistically, component count should not be a concern.
- Keep in mind that VMDKs are thinly provisioned by default, so customers should be prepared for future growth in capacity.

Network Design Considerations

The vSAN Network Design Guide provides requirements and best practices. A subset are discussed here

Network Connectivity and Bandwidth

In vSAN hybrid configurations, VMware supports 1Gb, 10Gb, 25Gb, 40Gb, and 100Gb Network Interface Cards (NICs) for vSAN network traffic. If a 1Gb NIC is used, VMware requires that this NIC is dedicated to vSAN traffic. If a 10Gb or higher bandwidth NICs are used, these can be shared with other network traffic types. While VMware has successfully run smaller hybrid vSAN deployments over 1Gb, the best practice is to use 10Gb links.

vSAN Express Storage Architecture (ESA) are supports 10Gbps for ESA-RN-0 profiles and ESA-2 and above profiles will 25Gbps or higher speed requirements. ESA ReadyNodes configured for vSAN ESA will be configured with 25/50/100Gbps NICs. The 10Gbps requirement for the ESA-RN-0 profiles should be considered for brownfield deployments, but it is strongly advised to still purchase 25Gbps NICs for future proofing as the SFP28 interface will be backwards compatible with a 10Gbps SFP+ switch interface.

vSAN OSA all-flash configurations are only supported with a 10Gb or higher connections. One reason for this is that the improved performance with an all-flash configuration may consume more network bandwidth between the hosts to gain higher throughput.

vSAN max has it's own networking speed requirements. See vSAN max documentation for configuration networking minimums. For large hosts with 200TB+ per host, 100Gbps will be required.

Consideration needs to be given to how much replication and communication traffic is going between the ESXi hosts, which is directly related to the number of virtual machines in the cluster, how many replicas as per virtual machine, and how I/O intensive are the applications running in the virtual machines.

Network switches and the interface cards that connect to them are what glues everything together. Networking plays a particularly important role with HCI as much of the I/O activity may have to extend beyond the local host. Unfortunately, the industry practice of referring to a switch specification simply by its maximum port speed dismisses all of the important details about the switchs. Switch capabilities are dependent on other factors such as the back plane bandwidth, the amount of port buffering available on the switch, and if the ASICs are powerful and plentiful enough to keep up with the "packets per second" processing requirements of the environment. When helping others trying to pinpoint the cause of their performance issues, and I ask about their switchgear, unfortunately, the responses often are not much longer than, "10 Gigabit." Another challenge with switches is that the typical life in production is longer than other assets in the data center. Longer life means that you have to be more aware of your future demands and invest in them perhaps sooner than you may wish. When using 10 and 25Gbps switching consider deeper switch buffers and more modern switch ASICs.

Recommendation: Strongly consider 25Gbps for new clusters and hosts, and consider 100Gbps Ethernet. While these connections can be shared with other traffic types, Network I/O Control is recommended to prioritize vSAN traffic.

Additional Content:

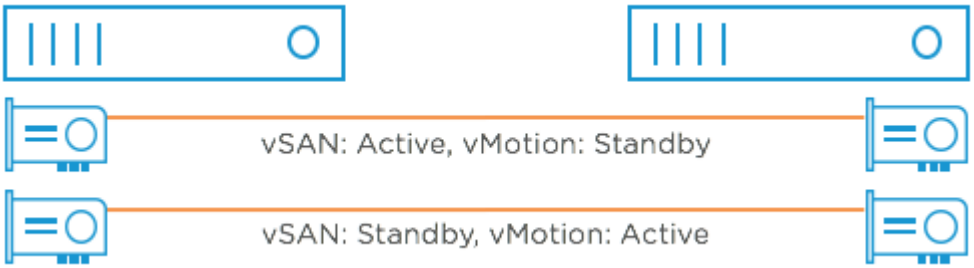
- Watch the HCI1845 VMworld session.
- Read "vSAN Design Considerations: Fast Storage devices vs. Fast Networking"
- Read "Common vSAN networking questions"

NIC Teaming for Redundancy

vSAN network traffic uses a single VMkernel port for most configurations. While some load balancing may occur when using LACP/LAG configurations, care should be applied when selecting and configuring an advanced hash to maximize path selection. Do note that the hash and configuration will need to be applied to the vDS as well as the physical switch. Basic IP hash selection policy on a standards witch is not advised as it will not use a dynamic LAG, and will "fail closed". In addition it will not yield significant performance gains vs. more advanced hashes that can balance TCP streams.

Please consult with Vmware Cloud Foundation or your cloud providers documentation if they support LAG/LACP configurations.

Multiple VMkernel ports can be used, but this configuration is not supported with datastore sharing, vSAN Max, or stretched clusters.



MTU and Jumbo Frames Considerations

vSAN supports jumbo frames. VMware testing finds that using jumbo frames can reduce CPU utilization and improve throughput. The gains are minimal because vSphere already uses TCP Segmentation Offload (TSO) and Large Receive Offload (LRO) to deliver similar benefits.

In data centers where jumbo frames are already activated in the network infrastructure, jumbo frames are recommended for vSAN deployment. Otherwise, jumbo frames are not recommended as the operational cost of configuring jumbo frames throughout the network infrastructure typically outweighs the benefits.

Recommendation: Consider jumbo frames for vSAN if the existing network environment is already configured to use jumbo frames.

Multicast Considerations (Legacy Pre 6.6 vSAN)

vSAN 6.6 and beyond no longer uses Multicast.

Network QoS Via Network I/O Control

Quality of Service (QoS) can be implemented using Network I/O Control (NIOC). This feature activates a dedicated amount of network bandwidth to be allocated to vSAN traffic. By using NIOC, it ensures that no other traffic will impact vSAN network performance through the use of a "shares" mechanism.

NIOC requires a distributed switch (VDS). NIOC is not available on a standard switch (VSS). Virtual Distributed Switches are included with vSAN licensing. This means NIOC can be configured with any edition of vSphere. vSAN supports the use of both VDS and VSS.

Summary of Network Design Considerations

- 25Gbps or faster (50/100Gbps) are required for vSAN Express Storage Architecture. 10 Gbps is supported for ESA-AF-0 ReadyNode profiles to support legacy/brownfield enviroments, but 25Gbps NICs should be used for these configurations.
- For ReadyNodes hosting 200TB+ of capacity (Common to vSAN Max), 100Gbps is required for the storage cluster.
- 10Gb networks at a minimum are required for vSAN Original Storage Architecture (OSA) all-flash configurations. 25Gbps or faster are recommended for best performance.
- NIC teaming is recommended for availability/redundancy
- Jumbo frames can provide benefits in a vSAN environment.
- Consider vDS with NIOC to provide QoS for vSAN traffic.
- The VMware vSAN Networking Design Guide reviews design options, best practices, and configuration details, including:
- vSphere Teaming Considerations – IP Hash vs other vSphere teaming algorithms
- Physical Topology/switch Considerations – Leaf Spine topology is preferred to legacy 3 tier designs or use of fabric extension.

- vSAN Network Design for High Availability – Design considerations to achieve a highly available vSAN network
- Load Balancing Considerations – How to achieve aggregated bandwidth via multiple physical uplinks for vSAN traffic in combination with other traffic types
- vSAN with other Traffic Types – Detailed architectural examples and test results of using Network I/O Control with vSAN and other traffic types

Storage Design Considerations

Before storage can be correctly sized for a vSAN, an understanding of key vSAN concepts is required. This understanding will help with the overall storage design of vSAN

A number of sub headings are marked with (OSA). This denotes design decisions that are only relevent for the vSAN Original Storage Architecture. For more information comparing vSAN Express Storage Architecture (ESA) to OSA please see this blog.

Disk Groups (OSA)

Disk Groups are only used for vSAN Original Storage Architecture. vSAN Express Storage Architecture uses storage pools. For a comparison of these options see this blog.

Disk groups can be thought of as storage containers on vSAN; they contain a maximum of one flash cache device and up to seven capacity devices: either magnetic disks or flash devices used as capacity in an all-flash configuration. To put it simply, a disk group assigns a cache device to provide the cache for a given capacity device. This gives a degree of control over performance as the cache to capacity ratio is based on disk group configuration.

If the desired cache to capacity ratio is very high, it may require multiple flash devices per host. In this case, multiple disk groups must be created to accommodate this since there is a limit of one flash device per disk group. There are significant performance advantages to using multiple disk groups. They typically provide more IOPS and also reduce the failure domains within a host.

The more cache capacity, then the more cache is available to virtual machines for accelerated performance. However, this leads to additional costs.

Design decision : For all but the lowest performance needs, start with two disk groups.

Cache Sizing Overview (OSA)

Cache devices are only used for vSAN Original Storage Architecture. vSAN Express Storage Architecture uses storage pools. For a comparison of these options see this blog.

Customers should size the cache requirement in vSAN based on the active working set of their virtual machines. Ideally the cache size should be big enough to hold the repeatedly used blocks in the workload. We call this the active working set. However, it is not easy to obtain the active working set of the workload because typical workloads show variations with respect to time, changing the working set and associated cache requirements.

As a guideline, VMware recommends having at least a 10% flash cache to consumed capacity ratio in Hybrid vSAN configurations. Previous to 6.5 All flash maintained the same recommendation. While this is still supported, new guidance is available on sizing based on target performance metrics and the read/write ratio of the workload. See this post for more information.

Flash Devices in vSAN Original Storage Architecture (OSA)

In vSAN hybrid configurations, the flash device serve two purposes; a read cache and a write buffer.es are used for the capacity layer. Both configurations dramatically improve the performance of virtual machines running on vSAN. More information can be found in An Overview of vSAN Caching Algorithms.

Purpose of read cache

The read cache, which is only relevant on hybrid configurations, keeps a collection of recently read disk blocks. This reduces the I/O read latency in the event of a cache hit, i.e. the disk block can be fetched from cache rather than magnetic disk.

For a given virtual machine data block, vSAN always reads from the same replica/mirror. However, when there are multiple replicas (to tolerate failures), vSAN divides up the caching of the data blocks evenly between the replica copies.

If the block being read from the first replica is not in cache, the directory service is referenced to find if the block is in the cache of another mirror (on another host) in the cluster. If it is found there, the data is retrieved from there. If it is not in cache on the other host, then there is a read cache miss. In that case, the data is retrieved directly from the magnetic disk.

Purpose of write cache

The write cache, found on both hybrid and all-flash configurations, behaves as a non-volatile write buffer. This greatly improves performance in both hybrid and all-flash configurations and also extends the life of flash capacity devices in all-flash

configurations.

When writes are written to flash, vSAN ensures that a copy of the data is written elsewhere in the cluster. All virtual machines deployed to vSAN have a default availability policy setting that ensures at least one additional copy of the virtual machine data is available. This includes making sure that writes end up in multiple write caches in the cluster.

Once a write is initiated by the application running inside of the Guest OS, the write is duplicated to the write cache on the hosts which contain replica copies of the storage objects. This means that in the event of a host failure, we also have a copy of the in-cache data and no loss of data can occur; the virtual machine will simply reuse the replicated copy of the cache as well as the replicated capacity data.

Client Cache

The Client Cache, introduced in vSAN 6.2, used on both hybrid and all-flash vSAN configurations, leverages DRAM memory local to the virtual machine to accelerate read performance. The amount of memory allocated is 0.4% up to 1GB per host.

As the cache is local to the virtual machine, it can properly leverage the latency of memory by avoiding having to reach out across the network for the data. In testing of read cache-friendly workloads, it was able to significantly reduce read latency.

This technology is complementary to CBRC and will activate the caching of VMDK's other than the read-only replica's that CBRC is limited to.

Caching Algorithm (OSA)

All the same considerations for sizing the capacity layer in hybrid configurations also apply to all-flash vSAN configurations. For example, one will need to take into account the number of virtual machines, the size of the VMDKs, the number of snapshots that are taken concurrently, and of course the number of replica copies that will be created based on the *NumberOfFailuresToTolerate* requirement in the VM storage policy.

With all-flash configurations, the caching algorithms are different than hybrid model. Read requests no longer need a cache tier to enhance performance. By removing the read cache in all-flash configurations, the entire device is devoted to write buffering and protecting the endurance of the capacity tier. This means that endurance and performance now become a consideration for the capacity layer in all-flash configurations.

In vSAN 5.5, which was available as a hybrid configuration only with a mixture of flash and spinning disk, cache behaved as both a write buffer (30%) and read cache (70%). If the cache did not satisfy a read request, in other words there was a read cache miss, then the data block was retrieved from the capacity layer. This was an expensive operation, especially in terms of latency, so the guideline was to keep your working set in cache as much as possible. Since the majority of virtualized applications have a working set somewhere in the region of 10%, this was where the cache size recommendation of 10% came from. With hybrid, there is regular destaging of data blocks from write cache to spinning disk. This is a proximal algorithm, which looks to destage data blocks that are contiguous (adjacent to one another). This speeds up the destaging operations.

All-Flash vSAN still has a write cache, and all VM writes hit this cache device. The major algorithm change, apart from the lack of read cache, is how the write cache is used. The write cache is now used to hold "hot" blocks of data (data that is in a state of change). Only when the blocks become "cold" (no longer updated/written) are they are moved to the capacity layer.

In all-flash configurations, having a high endurance flash cache device, or write intensive flash device is critical to the write latency consistency of the cache tier. If the working sets of the application running in the virtual machine fits mostly in the flash write cache, then there is a reduction in the number of writes to the flash capacity tier.

Flash Cache Sizing for Hybrid Configurations (OSA)

The general recommendation for sizing flash capacity for vSAN is to use 10% of the expected consumed storage capacity before the *NumberOfFailuresToTolerate* is considered. For example, a user plans to provision 1,000 virtual machines, each with 100GB of logical address space, thin provisioned. However, they anticipate that over time, the consumed storage capacity per virtual machine will be an average of 20GB.

So, in aggregate, the anticipated consumed storage, before replication, is 1,000 x 20GB = 20TB. If the virtual machine's availability factor is defined to support *NumberOfFailuresToTolerate* = 1 (FTT=1), this configuration results in creating two replicas for each virtual machine. That is, a little more than 40TB of consumed capacity,including replicated data. However, the flash sizing for this case is 10% x 20TB = 2TB of aggregate flash capacity in the cluster where the virtual machines are provisioned.

The optimal value of the target flash capacity percentage is based upon actual workload characteristics, such as the size of the working set of the data on disk. 10% is a general guideline to use as the initial basis for further refinement.

VMware recommends that cache be sized to be at least 10% of the capacity consumed by virtual machine storage (i.e. VMDK) For the majority of virtualized applications, approximately 10% of the data is being frequently accessed. The objective is to try to keep this data (active working set) in cache as much as possible for the best performance.

In addition, there are considerations regarding what happens in the event of a host failure or flash cache device failure, or in the event of a host in a vSAN cluster being placed in maintenance mode. If the wish is for vSAN to rebuild the components of the virtual machines impacted by a failure or maintenance mode, and the policy contains a setting for read cache reservation, this amount of read flash cache must be available after the failure for the virtual machine to be reconfigured.

The *FlashReadCacheReservation* policy setting is only relevant on hybrid clusters. All-flash arrays do not have a read cache. Reads come directly from the flash capacity layer unless the data block is already in the write cache.

This consideration is discussed in detail in the VM Storage Policies section later on in this guide.

Working example –hybrid configuration

A customer plans to deploy 100 virtual machines on a 4-node vSAN cluster. Assume that each VMDK is 100GB, but the estimate is that only 50% of each VMDK will be physically consumed.

The requirement is to have '*NumberOfFailuresToTolerate*' capability set to 1 in the policy used by these virtual machines.

Note: Although the '*NumberOfFailuresToTolerate*' capability set to 1 in the policy will double the amount of disk space consumed by these VMs, it does not enter into the calculation for cache sizing.

Therefore the amount of estimated consumed capacity will be 100 x 50GB = 5TB.

Cache should therefore be sized to 10% of 5TB = 500GB of flash is required. With a 4-node cluster, this would mean a flash device that is at least 125GB in size in each host.

However, as previously mentioned, considering designing with a larger cache configuration that will allow for seamless future capacity growth.In this example, if VMDKs eventually consume 70% vs. the estimate of 50%, the cache configuration would be undersized, and performance may be impacted.

Best practice: Ensure there is enough cache to meet the design requirements. The recommendation for cache is 10% of of the anticipated consumed storage capacity before the *NumberOfFailuresToTolerate* is considered.

Design consideration: Design for growth. Consider purchasing large enough flash devices that allow the capacity layer to be scaled simply over time.

Flash Cache Sizing for All-Flash Configurations (OSA)

All-flash vSAN configurations use the flash tier for write caching only, prior to 6.5 the all flash guidance was the same 10% and this will continue to be supported for existing deployments. Beyond this though guidance has shifted to be performance based.

Here are a table showing endurance classes and the total write buffer needed per host.

Note: while 1 to 5 disk groups per host are supported, we recommend a minimum of 2. Adding more disk groups can improve performance. Note, drives must still be certified specifically for All Flash Write Cache usage.

Assumptions

- Fault Tolerance Method = RAID5 / RAID6
- Accounted for 30% future performance increase & impact of resync/rebuild
- While assuming max sustained throughput, IOPS decreases proportionately if block size increases
- Ready Node profile details: https:// www.vmware.com/resources/compatibility/vsan_profile.html IOPS are assuming 4KB size. Large blocks divide accordingly.
- * The testing for these results was with 2 Disk groups for the AF-8 and AF-6. So in the case of the AF-8 the 100% sequential test was done with 400, 600 and 800 GB SSD drives. Note using more Disk groups could improve performance further.

For further information on this see the following blog post.

Best practice : Check the VCG and ensure that the flash devices are (a) supported and (b) provide the endurance characteristics that are required for the vSAN design.

Flash Endurance Considerations

Check the VCG and ensure that the flash devices are (a) supported and (b) provide the endurance characteristics that are required for the vSAN design.

For vSAN Express Storage Architecture (ESA) devices are explicitly certified and selected using the vSAN ESA ReadyNode sizer tool and will be selected as part of using the ReadyNode tool. For a list of all ESA capacity drives see this link. Note, that Read Intensive drives have been added to the ESA VCG, and are supported on all but the vSAN ESA-8 profile.

For vSAN Original Storage Architecture (OSA) Flash Devices are certified for use with Hybrid Cache, All-Flash Cache, All Flash Capacity.

Scale up Capacity, Ensure Adequate Cache (OSA)

One of the attractive features of vSAN is the ability to scale up as well as scale out.

The same is true if both cache and capacity are being scaled up at the same time through the addition of a new disk group. An administrator can simply add one new tier-1 flash device for cache, and at least one additional magnetic disk or flash devices for the capacity tier and build a new disk group. However, if the intent is to scale up the capacity of the vSAN datastore (adding more capacity per server), then it is important to ensure that there is sufficient cache. One consideration would be to provide a higher cache to capacity ratio initially, which will allow the capacity layer to grow with impacting future flash to capacity ratios.

It is relatively easy to scale up both cache and capacity together with the introduction of new disk groups. It is also easy to add additional capacity by inserting new magnetic disks to a disk group in hybrid (or flash devices for all-flash). But it could be much more difficult to add additional cache capacity. This is especially true if there is a need to swap out the current cache device and replace it with a newer larger one. Of course, this approach is also much more expensive. It is far easier to overcommit on flash resources to begin with rather than trying to increase it once vSAN is in production.

Design decision: Design with additional flash cache to allow easier scale up of the capacity layer. Alternatively scaling up cache and capacity at the same time through the addition of new disks groups is also an easier approach than trying to simply update the existing flash cache device in an existing disk group.

SATA, SAS, PCI-E and NVMe flash devices

vSAN Express Storage Architecture explicitly requires ReadyNodes configured with ESA compatible NVMe devices.

vSAN OSA Device Considerations

There are a number of considerations when deciding to choose SATA, SAS, PCI-E or NVMe flash devices. The considerations fall into three categories; cost, performance & capacity.

Performance deflation has been observed mixing SATA and SAS especially in larger drive configurations. This is most pronounced in cases where SAS expanders are in use.

Another useful performance consideration is that by using a PCI-E or NVMe caching device, it decreases the load on the storage controller used for SAS or SATA capacity devices paired with NVMe Cache.

For sustained workloads that will exceed the size of the write buffer, consider faster SAS or NVMe capacity tier devices. Slower SATA flash devices can become a bottleneck for large sustained write workloads.

NVMe offers low latency, higher performance, and lower CPU overhead for IO operations.

The cost difference for comparable performance SAS and NVMe is often negligible. Write endurance consideration is another important consideration; the higher the endurance, the higher the cost.

When sizing, ensure that there is sufficient tier-1 flash cache versus capacity (whether the capacity layer is magnetic disk or flash).

NVMe namespaces are not supported at this time.

Design consideration: Consider NVMe for Cache devices for demanding workloads.

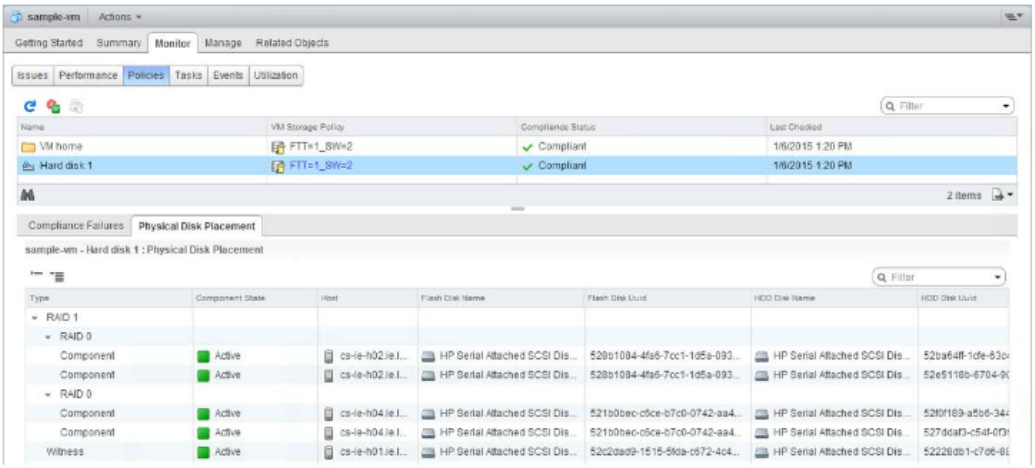
Design consideration: Consider if a design should have one large disk group with one large flash device, or multiple disk groups with multiple smaller flash devices. The latter design reduces the failure domain, and may also improve performance.

Magnetic Disks (OSA)

Magnetic disks make up the capacity of the vSAN datastore in hybrid configurations.

The number of magnetic disks is also a factor for stripe width. When stripe width is specified in the VM Storage policy, components making up the stripe will be placed on separate disks. If a particular stripe width is required, then there must be the required number of disks available across hosts in the cluster to meet the requirement. If the virtual machine also has a failure to tolerate requirement in its policy, then additional disks will be required on separate hosts, as each of the stripe components will need to be replicated.

In the screenshot below, we can see such a configuration. There is a stripe width requirement of two (RAID 0) and a failure to tolerate of one (RAID 1). Note that all components are placed on unique disks by observing the HDD Disk Uuid column:



Note that HDD refers to the capacity device. In hybrid configurations, this is a magnetic disk. In all-flash configurations, this is a flash device.

Magnetic disk performance –NL SAS, SAS or SATA

When configuring vSAN in hybrid mode, the capacity layer is made up of magnetic disks. A number of options are available to vSAN designers, and one needs to consider reliability, performance, capacity and price. There are three magnetic disk types supported for vSAN:

- Serial Attached SCSI (SAS)
- Near Line Serial Attached SCSI (NL-SAS)

- Serial Advanced Technology Attachment (SATA)

NL-SAS can be thought of as enterprise SATA drives but with a SAS interface. The best results can be obtained with SAS and NL-SAS. SATA magnetic disks are not certified for vSAN 6.0 and onward.

Magnetic disk capacity - SAS or NL-SAS

NL-SAS drives provide greater capacity than SAS drives for hybrid vSAN configurations. There is definitely a trade-off between the numbers of magnetic disks required for the capacity layer, and how well the capacity layer will perform.

vSAN 6.7 introduces the following new features and enhancements: 4Kn drive support. Please see the vSAN VCG list for supported 4Kn drives.

Magnetic disk performance

SAS disks tend to be more reliable and offer more performance, but at a cost. These are usually available at speeds up to 15K RPM (revolutions per minute). The VCG lists the RPM (drive speeds) of supported drives. This allows the designer to choose the level of performance required at the capacity layer when configuring a hybrid vSAN. While there is no need to check drivers/firmware of the magnetic disks, the SAS or SATA drives must be checked to ensure that they are supported.

Since SAS drives can perform much better than SATA, for performance at the magnetic disk layer in hybrid configurations, serious consideration should be given to the faster SAS drives.

Cache-friendly workloads are less sensitive to disk performance than cache-unfriendly workloads. However, since application performance profiles may change over time, it is usually a good practice to be conservative on required disk drive performance, with 10K RPM drives being a generally accepted standard for most workload mixes.

Number of magnetic disks matter in hybrid configurations

While having adequate amounts of flash cache is important, so are having enough magnetic disk spindles. In hybrid configurations, all virtual machines write operations go to flash, and at some point later, these blocks are destaged to a spinning magnetic disk. Having multiple magnetic disk spindles can speed up the destaging process.

Similarly, hybrid vSAN configurations target a 90% read cache hit rate. That means 10% of reads are going to be read cache misses, and these blocks will have to be retrieved from the spinning disks in the capacity layer. Once again, having multiple disk spindles can speed up these read operations.

Design decision: The number of magnetic disks matters in hybrid configurations, so choose them wisely. Having more, smaller magnetic disks will often give better performance than fewer, larger ones in hybrid configurations.

Using different magnetic disks models/types for capacity

VMware recommends against mixing different disks types in the same host and across different hosts. The reason for this is that performance of a component will depend on which individual disk type to which a component gets deployed, potentially leading to unpredictable performance results. VMware strongly recommends using a uniform disk model across all hosts in the cluster.

Design decision : Choose a standard disk model/type across all nodes in the cluster. Do not mix drive models/types.

How Much Capacity do I need?

When determining the amount of capacity required for a vSAN design, the '*NumberOfFailuresToTolerate*' policy setting plays an important role in this consideration. There is a direct relationship between the NumberOfFailuresToTolerate and the number of replicas created. For example, if the *NumberOfFailuresToTolerate* is set to 1 in the virtual machine storage policy, then there is another replica of the VMDK created on the capacity layer on another host (two copies of the data). If the *NumberOfFailuresToTolerate* is set to two, then there are two replica copies of the VMDK across the cluster (three copies of the data).

At this point, capacity is being sized for failure. However, there may be a desire to have enough capacity so that, in the event of a failure, vSAN can rebuild the missing/failed components on the remaining capacity in the cluster. In addition, there may be a desire to have full availability of the virtual machines when a host is taken out of the cluster for maintenance.

Another fundamental question is whether or not the design should allow vSAN to migrate and re-protect components during maintenance (or rebuild components during a failure) elsewhere in the cluster. If a host is placed in maintenance mode, and the storage objects are not rebuilt, a device failure during this time may cause data loss – an important consideration. Note that this

will only be possible if there are more than 3 nodes in the cluster. If it is a 3-node cluster only, then vSAN will not be able to rebuild components in the event of a failure. Note however that vSAN will handle the failure and I/O will continue, but the failure needs to be resolved before vSAN can rebuild the components and become fully protected again. If the cluster contains more than 3 nodes, and the requirement is to have the components rebuilt in the event of a failure or during a maintenance activity, then a certain amount of additional disk space needs to be reserved for this purpose. One should consider leaving one host worth of free storage available as that is the maximum amount of data that will need to be rebuilt if one failure occurs. If the design needs to tolerate two failures, then 2 additional nodes worth of free storage is required. This is the same for 16, 32 or 64 node configurations. The deciding factor on how much additional capacity is required depends on the *NumberOfFailuresToTolerate* setting.

Design decision: Always include the *NumberOfFailuresToTolerate* setting when designing vSAN capacity.

Design decision: If the requirement is to rebuild components after a failure, the design should be sized so that there is a free host worth of capacity to tolerate each failure. To rebuild components after one failure or during maintenance, there needs to be one full host worth of capacity free. To rebuild components after a second failure, there needs to be two full host worth of capacity free.

Formatting Overhead Considerations

The vSAN datastore capacity is determined by aggregating the device capacity layer from all ESXi hosts that are members of the cluster. In hybrid configurations, disk groups consist of a flash-based device and one or more magnetic disks pooled together, but only the usable capacity of the magnetic disks counts toward the total capacity of the vSAN datastore. For all flash configurations, only the flash devices marked as capacity are included in calculating the vSAN datastore capacity.

All the disks in a disk group are formatted with an on-disk file system. If the on-disk format is version 1, formatting consumes a total of 750 MB to 1GB of capacity per disk. In vSAN 6.x, administrators can use either v1 (VMFS-L) or v2 (VirstoFS). Formatting overhead is the same for on-disk format v1, but overhead for on-disk format v2 is different and is typically 1% of the drive's capacity. This needs to be considered when designing vSAN capacity requirements. The following table provides an estimation on the overhead required.

There is no support for the v2 on-disk format with vSAN version 5.5. The v2 format is only supported starting with vSAN 6.0. This overhead for v2 is very much dependent on how fragmented the user data is on the file system. In practice what has been observed is that the metadata overhead is typically less than 1% of the physical disk capacity. The v3 format introduces deduplication. Metadata overhead is highly variable and will depend on your data set.

Checksum Overhead: 5 bytes for every 4KB data are allocated for Checksum usage. Without deduplication this will use .12% of raw capacity and with deduplication will use up to 1.2%

Design decision: Include formatting overhead in capacity calculations.

Design consideration: There are other considerations to take into account apart from *NumberOfFailuresToTolerate* and formatting overhead. These include whether or virtual machine snapshots are planned. We will visit these when we look at some design examples.

Snapshot Cache Sizing Considerations

In vSAN 5.5, administrators who wished to use virtual machine snapshots needed to consider all of the same restrictions when compared to using virtual machine snapshots on VMFS or NFS datastores. As per VMware KB 1025279, VMware recommended using no single snapshot for more than 24-72 hours, and although 32 snap shots were supported in a chain, VMware recommended that only 2-3 snapshots in a chain were used.

In vSAN 6.x, and on-disk format (v2), there have been major enhancements to the snapshot mechanism, making virtual machine snapshots far superior than before. vSAN 6.x fully supports 32 snapshots per VMDK with the v2 on-disk format. The new snapshot mechanism on v2 uses a new "vsanSparse" format. However, while these new snapshots outperform the earlier version, there are

still some design and sizing concerns to consider.

When sizing cache for vSAN 6.x hybrid configurations, a design must take into account potential heavy usage of snapshots. Creating multiple, active snapshots may exhaust cache resources quickly, potentially impacting performance. The standard guidance of sizing cache to be 10% of consumed capacity may need to be increased to 15% or greater, especially with demanding snapshot usage.

Cache usage for virtual machine snapshots is not a concern for vSAN 6.x all-flash configurations. If the on-disk format is not upgraded to v2 when vSAN has been upgraded from version 5.5 to 6.0, and the on-disk format remains at v1, then the older (redo log) snapshot format is used, and the considerations in VMware KB 1025279 continue to apply.

Design consideration: If virtual machine snapshots are used heavily in a hybrid design, consider increasing the cache-to-capacity ratio from 10% to 15%.

Choosing a Storage I/O Controller

The most important aspect of storage design is ensuring that the components that are selected appear in the VMware Compatibility Guide (VCG). A VCG check will ensure that VMware supports the storage I/O controller and solid-state disk or PCIe flash device. Some design considerations for the storage hardware are listed here.

What kind of Controller to pick

RAID controllers add extra cost, and complexity. Higher performance and more consistent operations are experienced with pure pass-through HBA's. NVMe devices do not use a SAS controller and contain embedded controllers in the drives.

Design Decision: VMware recommends going forward to choose HBA over RAID controllers when using SAS/SATA drives

Multiple controllers and SAS Expanders

vSAN supports multiple controllers per ESXi host. The maximum number of disks per host is 35 (7 disks per disk group, 5 disk groups per host). Some controllers support 16 ports and therefore up to 16 disks can be placed behind one controller. The use of two such controllers in one host will get close to the maximums. However, some controllers only support 8 ports, so a total of 4 or 5 controllers would be needed to reach the maximum.

SAS expanders are sometimes considered to extend the number of storage devices that can be configured with a single storage I/O controller. VMware has not extensively tested SAS expanders with vSAN, and thus does not encourage their use. In addition to potential compatibility issues, the use of SAS expanders may impact performance and increase the impact of a failed disk group. SAS expanders have been tested in limited cases with Ready Nodes on a case by case. These ready nodes may have an “up to” maximum on the number of drives that have been certified with the expander. Refer to the vSAN VCG to see what SAS expanders have been certified and will be supported.

Intel VMD

Intel Volume Management Device is supported on ReadyNodes that have validated it. A list can be found here on the vSAN VCG. Intel VMD adds serviceability capabilities including hot plug support, and drive light support.

NVMe Hotplug Support

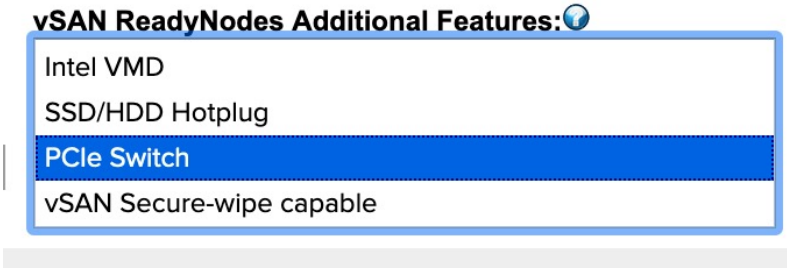
A vSAN VCG ReadyNodes list containing servers with vSphere native NVMe hot plug support can be found here.

Tri-mode controllers

Discrete RAID controllers that support SATA/SAS/NVMe are often known as "Tri-mode controllers". While some of these devices may become certified on the vSAN VCG, they are not supported for use with NVMe devices attached to them and passing IO through them. Tri-Mode controllers may only be used with SAS and SATA devices. NVMe drives are expected to connect to PCI-E without passing through a RAID controller. In cases where additional PCI-E lanes are needed to support dense server configurations, PCI Switches are a supported alternative.

PCIe Switch

Similar to a SAS expander, some servers will contain a PCIe switch that allows for oversubscription of PCIe channels to NVMe drives. The support policy for these is the same as SAS expanders in that this will only be supported on server platforms that have a ReadyNode certified with one included. For examples on the VCG you may specify a search includes a PCIe switch.



Multiple Controllers Versus Single Controllers

The difference between configuring ESXi hosts with multiple storage controllers and a single controller is that the former will allow potentially achieve higher performance as well as isolate a controller failure to a smaller subset of disk groups.

With a single controller, all devices in the host will be behind the same controller, even if there are multiple disks groups deployed on the host. Therefore a failure of the controller will impact all storage on this host.

If there are multiple controllers, some devices may be placed behind one controller and other devices behind another controller. Not only does this reduce the failure domain should a single controller fail, but this configuration also improves performance.

Design decision : Multiple storage I/O controllers per host can reduce the failure domain, and can also improve performance.

Storage Controller Queue Depth

There are two important items displayed by the VCG for storage I/O controllers that should be noted. The first of these is "features" and the second is queue depth.

Queue depth is extremely important, as issues have been observed with controllers that have very small queue depths. In particular, controllers with small queue depths (less than 256) can impact virtual machine I/O performance when vSAN is rebuilding components, either due to a failure or when requested to do so when entering maintenance mode.

Design Decision : Choose storage I/O controllers that have as large a queue depth as possible. While 256 are the minimum, the recommendation would be to choose a controller with a much larger queue depth where possible.

RAID-0 versus pass-through

The second important item is the "feature" column that displays how vSAN supports physical disk presentation to vSAN. There are entries referring to RAID 0 and pass-through. Pass-through means that this controller can work in a mode that will present the magnetic disks directly to the ESXi host. RAID 0 implies that each of the magnetic disks will have to be configured as a RAID 0 volume before the ESXi host can see them. There are additional considerations with RAID 0. For example, an administrator may have to take additional manual steps replacing a failed drive. These steps include rebuilding a new RAID 0 volume rather than simply plugging in a replacement empty disk into the host and allowing vSAN to claim it.

Design decision : Storage I/O controllers that offer RAID-0 mode typically take longer to install and replace than pass-thru drives from an operations perspective. When Possible use pass-through controllers.

Storage controller cache considerations

VMware's recommendation is to deactivate the cache on the controller if possible. vSAN is already caching data at the storage layer – there is no need to do this again at the controller layer. If this cannot be done due to restrictions on the storage controller, the recommendation is to set the cache to 100% read.

Advanced controller features

Some controller vendors provide third-party features for acceleration. For example, HP has a feature called Smart Path and LSI has a feature called Fast Path. VMware recommends disabling advanced features for acceleration when controllers are used in vSAN environments.

Design decision: When choosing a storage I/O controller, verify that it is on the VCG, ensure cache is deactivated, and ensure any third-party acceleration features are deactivated. If the controller offers both RAID 0 and pass-through support, consider using pass-through as this makes maintenance tasks such as disk replacement much easier.

Knowledge base related controller issues

The vSAN Online health service can identify controller configuration settings that are needed, as well as identify driver and firmware versions.

A search of kb.vmware.com should be performed for known configuration issues for a given controller.

In the case of the Dell H730 family of controllers (H730, H730p, H730 mini) see VMware KB 2109665 .

Disk Group Design

While vSAN requires at least one disk group per host contributing storage in a cluster, you should consider using more than one disk group per host.

Disk groups as a storage failure domain

A disk group can be thought of as a storage failure domain in vSAN. Should the flash cache device or storage I/O controller associated with a disk group fail, this will impact all the devices contributing towards capacity in the same disk group, and thus all the virtual machine components using that storage. All of the components residing in that disk group will be rebuilt elsewhere in the cluster, assuming there are enough resources available.

No other virtual machines that have their components in other hosts or in other disk groups, or attached to a different storage I/O controller are impacted.

Therefore, having one very large disk group with a large flash device and lots of capacity might mean that a considerable amount of data needs to be rebuilt in the event of a failure. This rebuild traffic could impact the performance of the virtual machine traffic. The length of time to rebuild the components is also a concern because virtual machines that have components that are being rebuilt are exposed to another failure occurring during this time.

By using multiple smaller disk groups, performance can be improved and the failure domain reduced in the event of storage I/O controller or flash device failure. The tradeoff once again is that this design requires multiple flash devices and/or storage I/O controllers, which consumes extra disk slots and may be an additional expense and needs consideration

Often times the cost of implementing multiple disk groups is not higher. If the cost of 2 x 400GB solid-state devices is compared to 1 x 800GB solid-state device, the price is very often similar. Also worth considering is that two cache devices in two disk groups on the same host can provide significantly higher IOPS than one cache device in one disk group.

Design decision : Multiple disk groups typically mean better performance and smaller fault domains.

Small Disk Drive Capacity Considerations

When using small capacity devices, and deploying virtual machines with large VMDK sizes, a VMDK object may be split into multiple components across multiple disks to accommodate the large VMDK size. This is shown as a RAID-0 configuration for the VMDK object. However, when vSAN splits an object in this way, multiple components may reside on the same physical disk, a configuration that is not allowed when *NumberOfDiskStripesPerObject* is specified in the policy.

This is not necessarily an issue, and vSAN is designed to handle this quite well. But it can explain why objects are getting striped when there was no stripe width request placed in the policy.

Very Large VMDK Considerations

Starting with vSAN 6.0, virtual machine disk sizes of 62TB are now supported. However, consideration should be given as to whether an application actually requires this size of a VMDK. As previously mentioned, the maximum component size on vSAN is 255GB. When creating very large VMDKs, the object will be split (striped) into multiple 255GB components. This may quickly consume the component count of the hosts, and this is especially true when *NumberOfFailuresToTolerate* is taken into account. A single 62TB VMDK with *NumberOfFailuresToTolerate* = 1 will require 500 or so components in the cluster (though many of these components can reside on the same physical devices).

One other consideration is that although vSAN might have the aggregate space available on the cluster to accommodate this large size VMDK object, it will depend on where this space is available and whether or not this space can be used to meet the requirements in the VM storage policy.

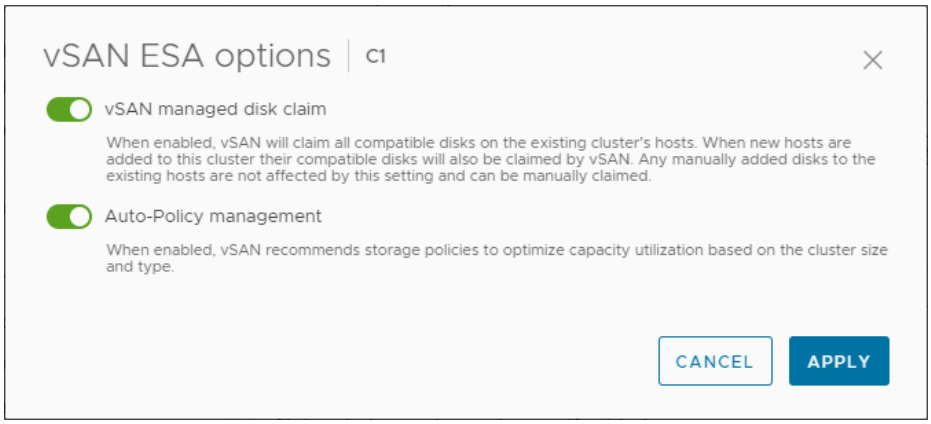
For example, in a 3-node cluster which has 200TB of free space, one could conceivably believe that this should accommodate a VMDK with 62TB that has a *NumberOfFailuresToTolerate*=1 (2 x 62TB = 124TB). However if one host has 100TB free, host two

has 50TB free and host three has 50TB free, then this vSAN will not be able to accommodate this request.

Disk Replacement/Upgrade Ergonomics

Ergonomics of device maintenance is an important consideration. One consideration is the ease of replacing a failed component on the host. One simple question regarding the host is whether the disk bays are located in the front of the server, or does the operator need to slide the enclosure out of the rack to gain access. A similar consideration applies to PCIe devices, should they need to be replaced.

vSAN 8 ESA introduced a managed disk claim option.



Further improvements in vSAN 8U2 ESA allow prescriptive disk reclaim and replacement. This new approach uses a cluster-specific definition residing on the vCenter Server that manages the cluster. It is where one can set a variety of attributes that will denote device characteristics such as a disk vendor, disk capacity, and the number of disks per hosts. Many of the attributes are an optional specification to provide as much flexibility as possible.

For example, let's suppose the ESA prescriptive disk claim for this cluster specifies that 6 devices in each host should be claimed for vSAN even though there are 8 eligible devices in the host. vSAN will claim no more than 6 eligible devices in each host across the cluster. If a host is added to the cluster, it will apply this same desired state to the new host. If additional storage devices are added to each host in increase capacity, no additional devices are claimed until the desired state configuration is adjusted to say otherwise. Any type of non-compliance such as a change in definition, or a change in a host configuration will trigger a "vSAN Managed disk claim" health finding to identify configuration drift.

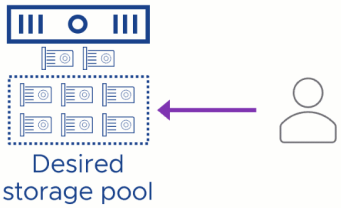


Figure 2. vSAN ESA Prescriptive Disk Claim process.

There is another consideration is around hot plug/host swap support. If a drive fails, vSAN 6.x provides administrators with the capability of lighting the LED on the drive for identification purposes. Once the drive is located in the server/rack, it can be removed from the disk group via the UI (which includes a disk evacuation option in version 6.x) and then the drive can be ejected and replaced with a new one.

vSAN OSA Considerations

Certain controllers, especially when they are using RAID 0 mode rather than pass-through mode, require additional steps to get the drive discovered when the original is ejected and a new drive inserted. This operation needs to be as seamless as possible, so it is important to consider whether or not the controller chosen for the vSAN design can support plug-n-play operations.

Large Capacity Device Conslideration s

A limitation of vSAN up to this point had been around the logical capacity limits per storage device in the capacity tier. The result was 16TB of addressable space per storage device, which meant that for clusters running deduplication and compression, the effective ratios could be impacted by the device capacity.

New disk groups created will leverage a new logical space format if:

1. Deduplication and compression is turned on, and devices larger than 4TB are used.

2. Deduplication and compression is turned off, and devices larger than 16TB are used.

3. If larger devices are added to a disk group after creation, a health check will prompt for an optional upgrade.

As capacity per host capabilities increase, so should the considerations around network demands from accessing potentially more data, or host evacuation and rebuild scenarios. Customers looking at configurations with significantly higher densities should take network speed and configuration options into consideration, to ensure that the network can sufficiently meet their expectations. For additional information on this topic see this blog.

Summary of Storage Design Considerations

- Consider whether an all-flash solution or a hybrid solution is preferable for the vSAN design. All-flash, while possibly more expensive, can offer higher performance and low latency.
- Keep in mind the 10% flash to capacity ratio for the hybrid vSAN. Spend some time determining how large the capacity layer will be over time, and use the formula provided to extrapolate the flash cache size. For All flash consult the performance table when sizing the cache devices as it has replaced the 10% rule.
- Determine the endurance required for the flash cache, and the flash capacity requirement for all-flash solution designs.
- Consider whether PCI-E flash devices or SSDs are best for the design.
- Consider if a design should have one large disk group with one large flash device or multiple disk groups with multiple smaller flash devices.
- Determine the best magnetic disk for any hybrid solution design.
- Design with one additional host with enough capacity to facilitate remediation on disk failure, which will allow for another failure in the cluster to occur while providing full virtual machine availability.
- Remember to include file system overhead when sizing the capacity layer.
- Consider, if possible, multiple storage I/O controllers per host for performance and redundancy.
- Consider the benefits of pass-through over RAID-0 and ensure that the desired mode is supported by the controller.
- Deactivate cache on controllers, or if not possible, set cache to 100% read.
- Deactivate advanced features of the storage I/O controllers.
- When designing disk groups, consider disk groups not only as a failure domain, but a way of increasing performance.
- Consider the limitations around using very small physical drives.

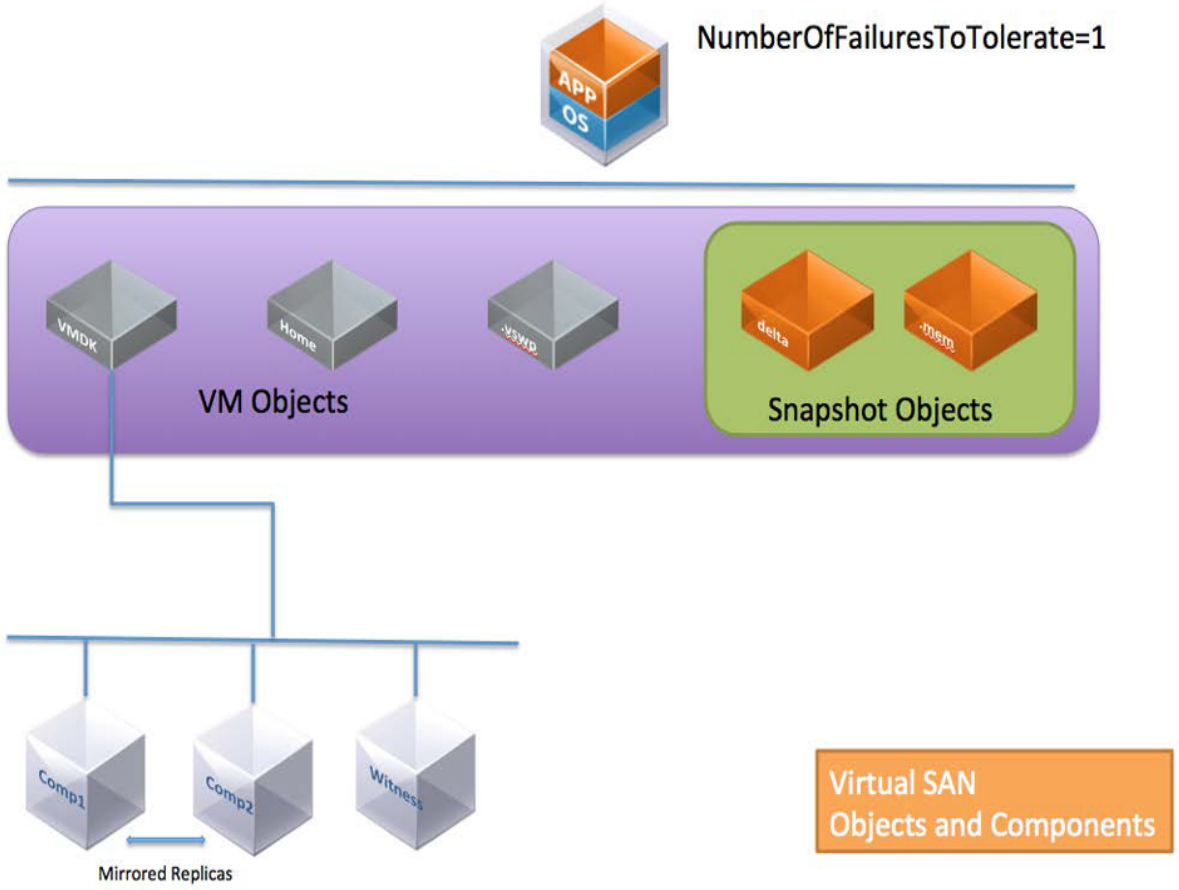
- Consider the limitations when deploying very large virtual machine disks on vSAN.
- Consider a design that will facilitate easy replacement of failed components.
- Although not required, it is best to create vSAN disk groups using the same drive configuration (number of drives, cache drive and capacity drive types and sizes) across all hosts in the cluster. This helps ensure consistent performance and provides the most flexibility for data placement particularly in cases where there has been a host or drive failure.

VM Storage Policy Design Considerations

It is important to have an understanding of the VM Storage Policy mechanism as part vSAN. VM Storage Policies define the requirements of the application running in the virtual machine from an availability, sizing and performance perspective.

Objects and Components

A virtual machine deployed on a vSAN datastore is comprised of a set of objects. These are the VM Home Namespace, the VMDK, VM Swap (when the virtual machine is powered on) and in the case of a snapshot, there is the delta VMDKs and the virtual machine memory snapshot (when this is captured as part of the snapshot):



Each of these objects is comprised of a set of components, determined by capabilities placed in the VM Storage Policy. For example, if *NumberOfFailuresToTolerate=1* is set in the VM Storage Policy, then the VMDK object would be mirrored/replicated, with each replica being comprised of at least one component. If *NumberOfDiskStripesPerObject* is greater than one in the VM Storage Policy, then the object is striped across multiple disks and each stripe is said to be a component of the object.

In vSAN OSA, if the component is built on the capacity layer that has been upgraded to the v2 on-disk format, it is an additional 4MB.

This appreciation of the relationship between virtual machines, objects and components will help with understanding the various vSAN failure scenarios.

Design consideration: Realistically, the metadata overhead incurred by creating components on vSAN is negligible and doesn't need to be included in the overall capacity. It is built into the vSAN Sizer and that tool should be used to address overhead concerns in sizing.

Witness and Replicas

In vSAN OSA witnesses are an integral component of RAID 1 storage objects, as long as the object is configured to tolerate at least one failure. They are components that do not contain data, only metadata. Their purpose is to serve as tiebreakers when availability decisions are made to meet the failures to tolerate policy setting. They are used when determining if a quorum of components exists in the cluster. A witness consumes about 2MB of space for metadata on the vSAN datastore.

For VSAN quorum is computed on a weighted vote system. Each component has a number of votes, which maybe 1 or more. Quorum is calculated based on the rule that "more than 50% of votes" is required. It then becomes a possibility that components are distributed in such a way that vSAN can still guarantee failures to tolerate without the use of witnesses. For vSAN OSA RAID 1 a witness may still be common, but for vSAN ESA the combination of performance and capacity leg components removes the need for a witness.

Replicas make up virtual machine storage objects. Replicas are instantiated when an availability capability (*NumberOfFailuresToTolerate*) is specified for the virtual machine. The availability capability dictates how many replicas are created. It activates virtual machines to continue running with a full complement of data when there are host, network or disk failures in the cluster.

Quorum improvements in 7 Update 3 for Witness with Stretched Cluster and 2-Node cluster

For an object to be accessible in vSAN 6.x, more than 50% of its votes must be accessible. A special case exists in vSphere 7 Update 3 now where in cases of using a witness appliance (either for stretched, or two node clusters) votes can be re-assigned after one of the two primary fault domains has failed. This process allows a witness to recognize that it's votes should be transferred to the remaining fault domain, thus avoiding an outage should a witness fail at a time after which one of the sites, or nodes within a 2 node cluster has failed. For more information see this blog.

Design consideration: Realistically, the overhead incurred by creating witnesses on vSAN is negligible and does not need to be included in the overall capacity.

Virtual Machine Snapshot Considerations

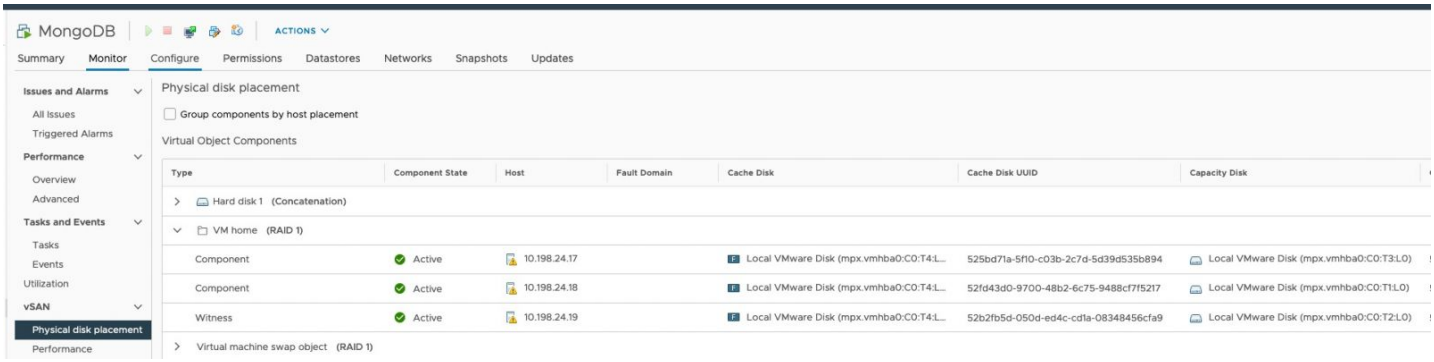
vSAN OSA uses it's own SparseSE snapshotting system. This requires IO to merge on snapshot deletion and while improved there can be performance impacts from long term running with many snapshots.

vSAN ESA uses a native file system snapshot system. For more information see this blog or this video demo.

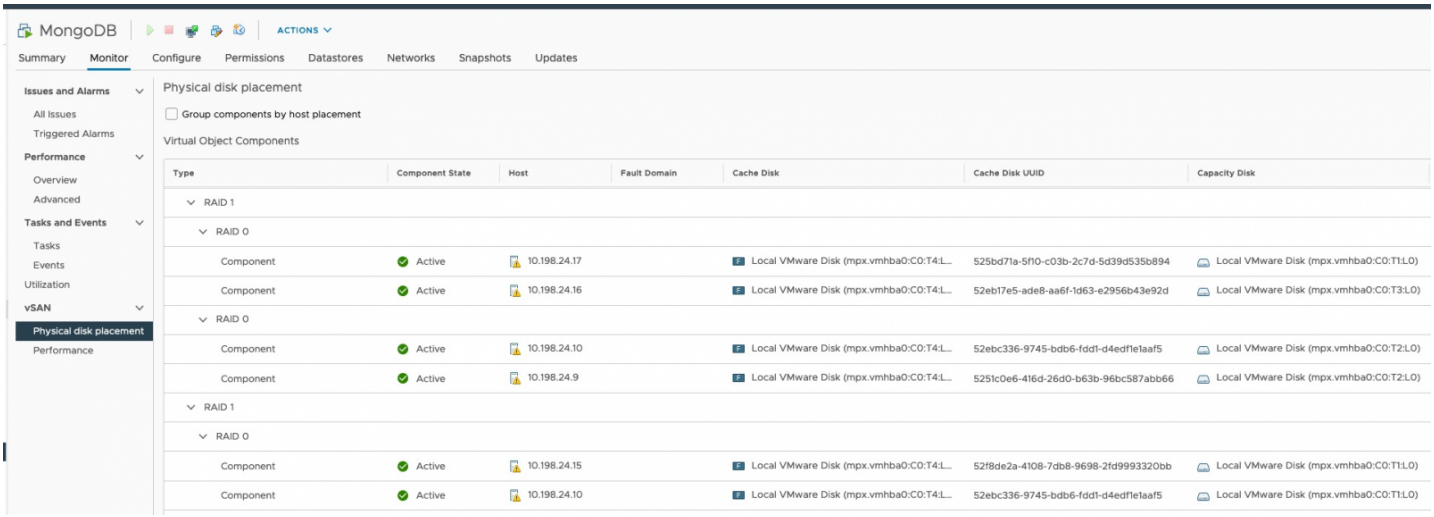
Design consideration: The virtual machine memory snapshot size needs to be considered when sizing the vSAN datastore, if there is a desire to use virtual machine snapshots and capture the virtual machine's memory in the snapshot.

Reviewing Object Layout from UI

The vSphere web client provides a way of examining the layout of an object on vSAN. Below, the VM Home namespace object and VMDK object are displayed when a virtual machine has been deployed with a policy settings of *NumberOfFailuresToTolerate = 1* and *NumberOfDiskStripesPerObject = 2*. The first screenshot is from the VM home. This does not implement the stripe width setting, but it does implement the failures to tolerate policy setting. There is a RAID 1 containing two components (replicas) and a third witness component for quorum. Both the components and witness must be on different hosts.



This next screenshot is taken from the VMDK – Hard disk 1. It implements both the stripe width (RAID 0) and the failures to tolerate (RAID 1) requirements. There are a total of 5 components making up this object; two components are striped, and then mirrored to another two-way stripe. Finally, the object also contains a witness component for quorum decisions.



Note: The location of the Physical Disk Placement view has changed between versions 5.5 and 6.0. In 5.5, it is located under the Manage tab. In 6.0, it is under the Monitor tab.

Policy Design Decisions

Administrators must understand how these storage capabilities affect the consumption of storage capacity in vSAN.

vSAN ESA - Auto Policy Management

vSAN 8 U1 introduces a way in the ESA to ensure data using a default storage policy is stored in the most optimal, resilient way. The cluster service can be enabled by highlighting the cluster, clicking on Configure > vSAN Services > Storage, then clicking on "Edit" and enabling "Auto-Policy management."

The policy settings the optimized storage policy uses are based on the **type of cluster**, the **number of hosts in a cluster**, and if the **Host Rebuild Reserve** (HRR) capacity management feature is enabled on the cluster. A change to any one of the three will result in vSAN making a suggested adjustment to the cluster-specific, optimized storage policy.

- Standard vSAN clusters (with Host Rebuild Reserve **turned off**):
  - 3 hosts without HRR : FTT=1 using RAID-1
  - 4 hosts without HRR: FTT=1 using RAID-5 (2+1)
  - 5 hosts without HRR: FTT=1 using RAID-5 (2+1)
  - 6 or more hosts without HRR: FTT=2 using RAID-6 (4+2)
- Standard vSAN clusters (with Host Rebuild Reserve **enabled**)
  - 3 hosts with HRR: (HRR not supported with 3 hosts)
  - 4 hosts with HRR: FTT=1 using RAID-1

  - 5 hosts with HRR: FTT=1 using RAID-5 (2+1)
  - 6 hosts with HRR: FTT=1 using RAID-5 (4+1)
  - 7 or more hosts with HRR: FTT=2 using RAID-6 (4+2)

- vSAN Stretched clusters:
  - 3 data hosts at each site: Site level mirroring with FTT=1 using RAID-1 mirroring for a secondary level of resilience
  - 4 hosts at each site: Site level mirroring with FTT=1 using RAID-5 (2+1) for secondary level of resilience.
  - 5 hosts at each site: Site level mirroring with FTT=1 using RAID-5 (2+1) for secondary level of resilience.
  - 6 or more hosts at each site: Site level mirroring with FTT=2 using RAID-6 (4+2) for a secondary level of resilience.
- vSAN 2-Node clusters:
  - 2 data hosts: Host level mirroring using RAID-1

For more information about this policy see this blog.

Number of Disk Stripes Per Object/Stripe Width

Number Of Disk Stripes Per Object , commonly referred to as stripe width, is the setting that defines the minimum number of capacity devices across which each replica of a storage object is distributed. vSAN may actually create more stripes than the number specified in the policy.

Striping may help performance if certain virtual machines are I/O intensive and others are not. With striping, the data of a virtual machine is spread across more drives which all contribute to the overall storage performance experienced by that virtual machine. In the case of hybrid, this striping would be across magnetic disks. In the case of all-flash, the striping would be across whatever flash devices are making up the capacity layer.

vSAN 7 Update 1 makes a number of changes to how striping is handled:

- For objects over 2TB, the components beyond the first 2TB will reduce the maximum stripe width to 3.
- RAID 5 is now considered to be 4 stripe width by default, and RAID 6 a stripe width of 6.
- If possible, the cluster level object manager will strive to place the stripes of an R0 leaf on different disk groups. This helps avoid a scenario in which under a deduplication activated cluster, the stripes were placed on the same disk group, potentially offering no material benefit

However, striping may not help performance if an application is not especially I/O intensive, or a virtual machine's data is spread across devices that are already busy servicing other I/O intensive virtual machines.

However, for the most part, VMware recommends leaving striping at the default value of 1 unless performance issues that might be alleviated by striping are observed. The default value for the stripe width is 1 whereas the maximum value is 12.

Stripe Width – Sizing Consideration (OSA)

For information about stripe width and vSAN ESA see this blog.

There are two main sizing considerations when it comes to stripe width. The first of these considerations is if there are enough physical devices in the various hosts and across the cluster to accommodate the requested stripe width, especially when there is also a *NumberOfFailuresToTolerate* value to accommodate.

The second consideration is whether the value chosen for stripe width is going to require a significant number of components and consume the host component count. Both of these should be considered as part of any vSAN design, although considering the increase in the maximum component count in 6.0 with on-disk format v2, this is not a major concern anymore. Later, some working examples will be looked at which will show how to take these factors into consideration when designing a vSAN cluster.

Flash Read Cache Reservation

Previously we mentioned the 10% rule for flash cache sizing. This is used as a read cache and write buffer in hybrid configurations, and as a write buffer only for all-flash configurations, and is distributed fairly among all virtual machines. However, through the use of VM Storage Policy setting *Flash Read Cache Reservation* , it is possible to dedicate a portion of the read cache to one or more virtual machines.

Note : This policy setting is only relevant to hybrid configurations. It is not supported or relevant in all-flash configurations due to changes in the caching mechanisms and the fact that there is no read cache in an all-flash configuration.

For hybrid configurations, this setting defines how much read flash capacity should be reserved for a storage object. It is specified as a percentage of the logical size of the virtual machine disk object. It should only be used for addressing specifically identified read performance issues. Other virtual machine objects do not use this reserved flash cache capacity.

Unreserved flash is shared fairly between all objects, so, for this reason, VMware recommends not changing the flash reservation unless a specific performance issue is observed. The default value is 0%, implying the object has no read cache reserved but shares it with other virtual machines. The maximum value is 100%, meaning that the amount of reserved read cache is the same size as the storage object (VMDK).

Flash Read Cache Reservation – sizing considerations

Care must be taken when setting a read cache reservation requirement in the VM Storage Policy. What might appear to be small *Flash Read Cache Reservation* numbers to users can easily exhaust all SSD resources, especially if thin provisioning is being used (Note that in VM Storage Policy terminology, thin provisioning is referred to as Object Space Reservation).

Flash Read Cache Reservation configuration example

In this hybrid vSAN example, the customer has set the VM Storage Policy – *Flash Read Cache Reservation* to 5% for all the virtual machine disks. Remember that 70% of flash is set aside for read cache in hybrid configurations.

With thin provisioning, customers can overprovision and have more logical address space than real space. In this example, the customer has thin provisioned twice as much logical space than physical space (200%).

If the *Flash Read Cache Reservation* requested by the administrator is calculated and compared to the total flash read cache available on the host, it reveals the following:

Total disk space consumed by VMs: X

- Total available flash read cache: (70% of 10% of X) = 7% of X
- Requested flash read cache reservation: (5% of 200% of X) = 10% of X

=> 10% of X is greater than 7% of X

Therefore if thin provisioning is being used to over-commit storage space, great care must be taken to ensure this does not negatively impact cache reservation settings. If cache reservation uses up all of the read cache, it can negatively impact performance.

Design consideration : Use *Flash Read Cache Reservation* with caution. A misconfiguration or miscalculation can very easily over-allocate read cache to some virtual machines while starving others.

Force Provisioning

The Force provisioning policy allows vSAN to violate the *NumberOfFailuresToTolerate (FTT)* , *NumberOfDiskStripesPerObject* (SW) and *FlashReadCacheReservation* (FRCR) policy settings during the initial deployment of a virtual machine.

vSAN will attempt to find a placement that meets all requirements. If it cannot, it will attempt a much simpler placement with requirements reduced to FTT=0, SW=1, FRCR=0. This means vSAN will attempt to create an object with just a single mirror. Any *ObjectSpaceReservation (OSR)* policy setting is still honored.

vSAN does not gracefully try to find a placement for an object that simply reduces the requirements that cannot be met. For example, if an object asks for FTT=2, if that cannot be met, vSAN will not try FTT=1, but instead immediately tries FTT=0.

Similarly, if the requirement was FTT=1, SW=10, but vSAN does not have enough capacity devices to accommodate SW=10, then it will fall back to FTT=0, SW=1, even though a policy of FTT=1, SW=1 may have succeeded.

There is another consideration. Force Provisioning can lead to capacity issues if its behavior is not well understood by administrators. If a number of virtual machines have been force provisioned, but only one replica copy of an object is currently instantiated due to lack of resources, as soon as those resources become available through the addition of new hosts or new disks,

vSAN will consume them on behalf of those virtual machines.

Administrators who use this option to force provision virtual machines need to be aware that once additional resources become available in the cluster, vSAN may immediately consume these resources to try to satisfy the policy settings of virtual machines.

Caution : Another special consideration relates to entering Maintenance Mode in full data migration mode, as well as disk/disk group removal with data migration that was introduced in vSAN 6.0. If an object is currently non-compliant due to force provisioning (either because initial placement or policy reconfiguration could not satisfy the policy requirements), then "Full data evacuation" of such an object will actually behave like "Ensure Accessibility", i.e. the evacuation will allow the object to have reduced availability, exposing it a higher risk. This is an important consideration when using force provisioning, and only applies for non-compliant objects.

Best practice: Check if any virtual machines are non-compliant due to a lack of resources before adding new resources. This will explain why new resources are being consumed immediately by vSAN. Also check if there are non-compliant VMs due to force provisioning before doing a full data migration.

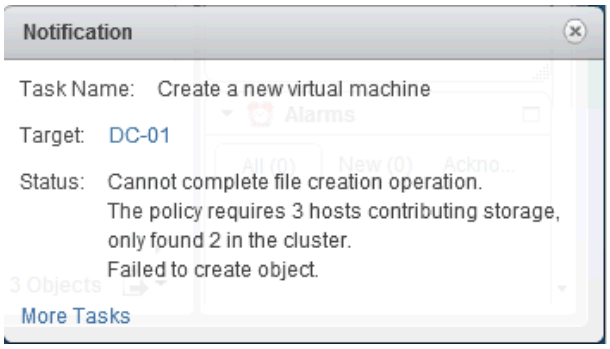
Object Space Reservation

An administrator should always be aware of over-committing storage on vSAN, just as one needs to monitor over-commitment on a traditional SAN or NAS array.

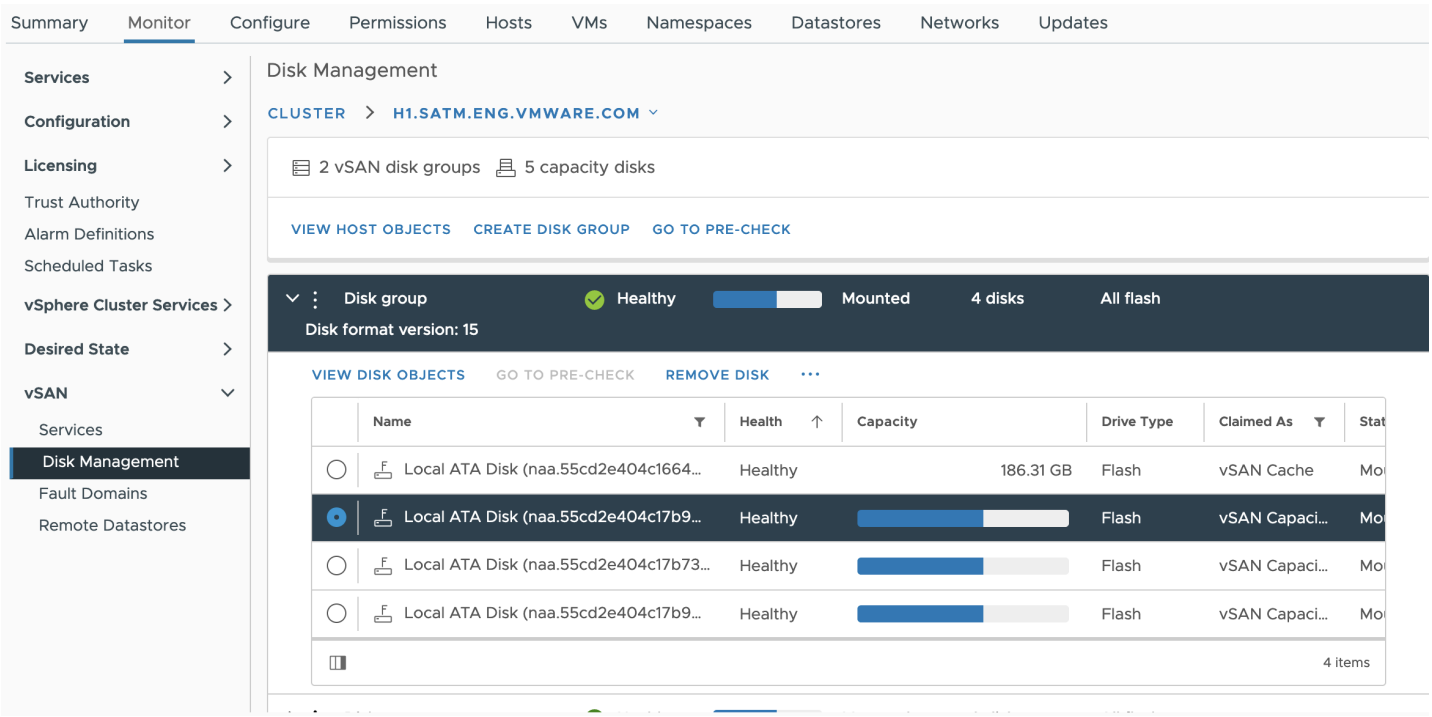
By default, virtual machine storage objects deployed on vSAN are thinly provisioned. This capability, *ObjectSpaceReservation (OSR*), specifies the percentage of the logical size of the storage object that should be reserved (thick provisioned) when the virtual machine is being provisioned. The rest of the storage object will remain thin provisioned. The default value is 0%, implying the object is deployed as thin. The maximum value is 100%, meaning the space for the object is fully reserved, which can be thought of as full, thick provisioned. Since the default is 0%, all virtual machines deployed on vSAN are provisioned as thin disks unless one explicitly states a requirement for *ObjectSpaceReservation* in the policy. If *ObjectSpaceReservation* is specified, a portion of the storage object associated with that policy is reserved.

There is no eager-zeroed thick format on vSAN. OSR, when used, behaves similarly to lazy-zeroed thick.

There are a number of safeguards that will prevent over-commitment. For instance, if there is not enough storage capacity across the required number of hosts in the cluster to satisfy a replica or stripe width policy setting, then the following warning is displayed.



The Monitor > vSAN > Disk Management view will display the amount of used capacity in the cluster.



Design consideration : While the creation of replicas is taken into account when the capacity of the vSAN datastore is calculated, thin provisioning over-commitment is something that should be considered in the sizing calculations when provisioning virtual machines on a vSAN.

IOP Limit For Object

There are cases where an administrator will want to limit the maximum amount of IOPS that are available to an object or virtual machine. There are two key use cases for this functionality

Preventing noisy neighbor workloads from impacting other workloads that need more performance available.

Create artificial standards of service as part of a tiered service offering using the same pool of resources.

By default, vSAN seeks to dynamically adjust performance based on demand and provide a fair weighting of resources available. This capability *IopLimitForObject* limits the amount of performance available to an object. This is normalized to a 32KB block size. A virtual machine reading or writing at 16KB would be treated the same as one performing 32 KB-sized operations. A 64KB read or write however would be treated as two separate operations, leading to half of the configured IOP limit being the number of operations performed.

Deactivate Object Checksum

This policy only works with vSAN OSA.

Object Checksums were introduced as a policy in vSAN 6.2. They activate the detection of corruption caused by hardware/software components including memory, drives, etc during the read or write operations.

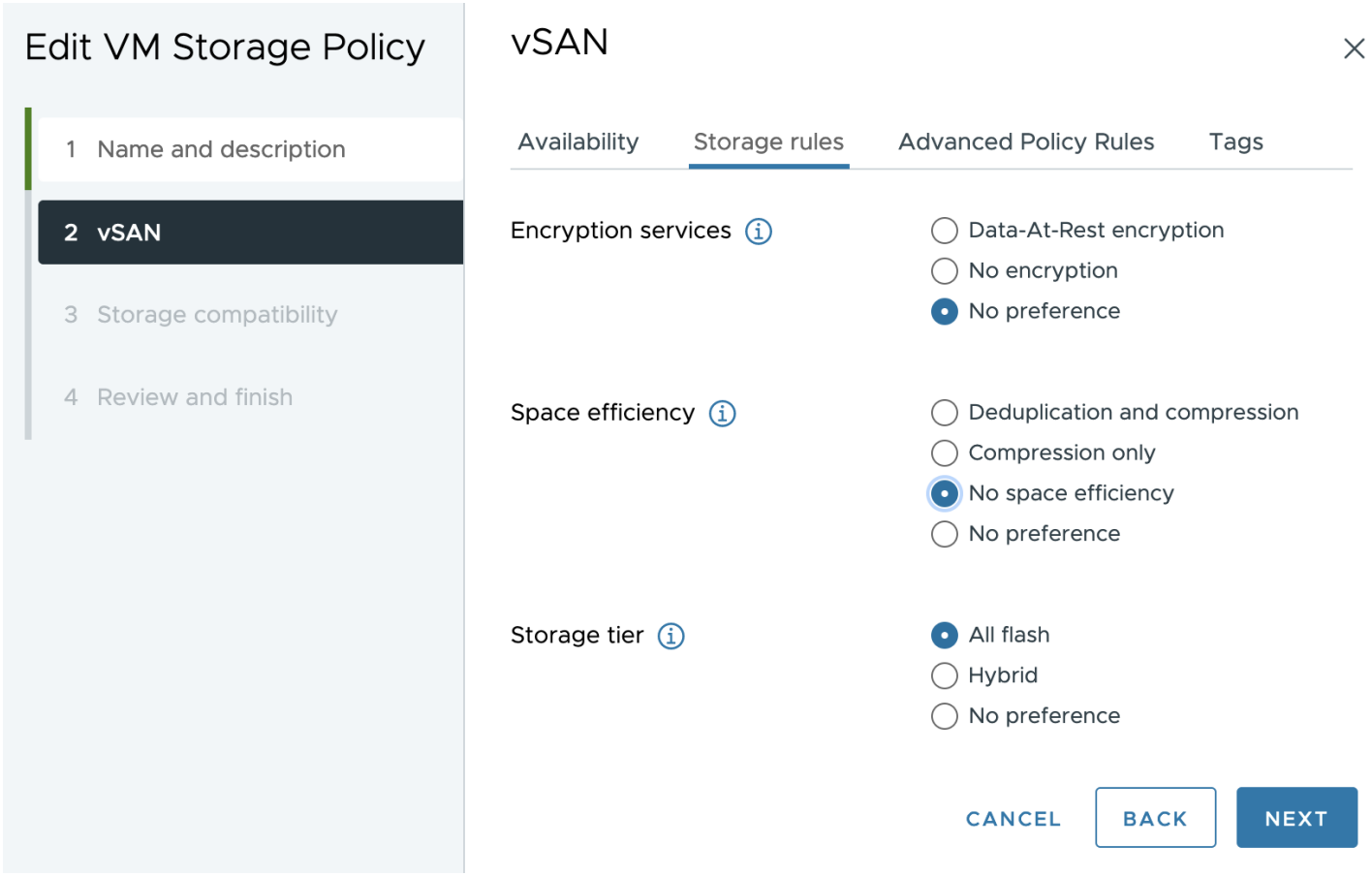
Object Checksums are activated by default for objects residing on vSAN file system version 3. They will verify checksums on all reads, as well as a scrubber will scan all blocks that have not been read within one year. The scrubber schedule can be adjusted to run more often but note this may increase background disk IO. In vSAN 7 U2, this disk scrubbing frequency will occur once every two weeks, with negligible impact on performance during the activity (no more than 2% performance hit while the task is running.

Object checksums carry a small disk IO, memory, and compute overhead and can be deactivated on a per-object basis using the checksum storage policy.

vSAN ESA Compression

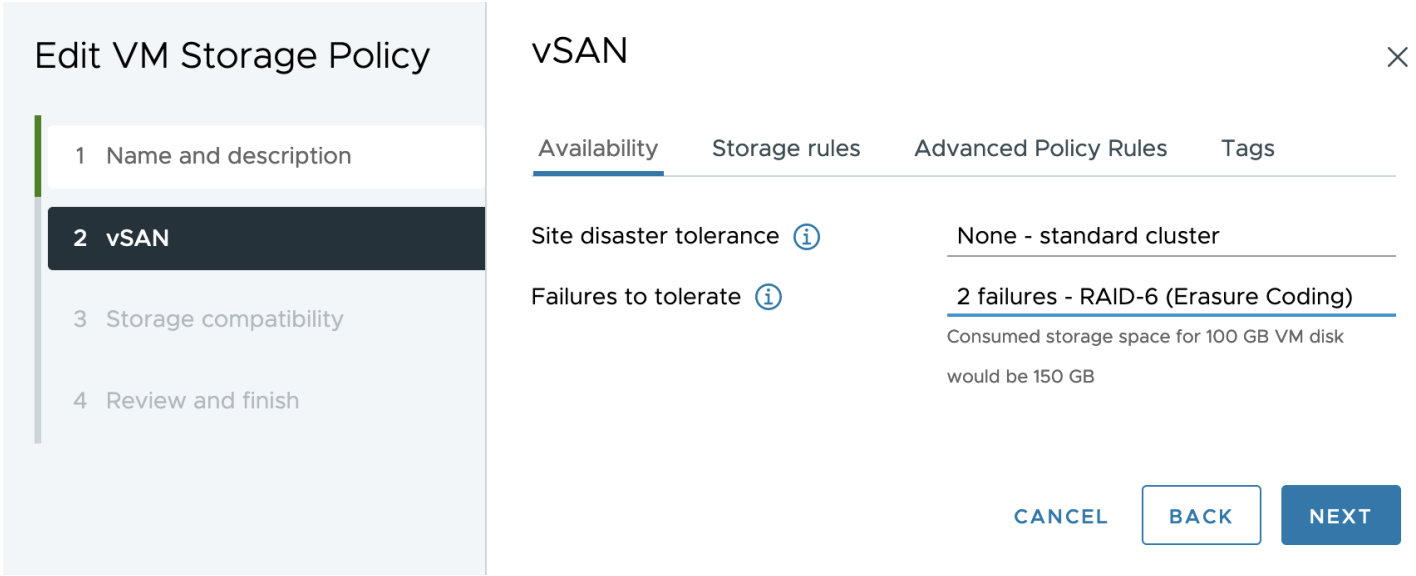
VSAN 8 Express Storage Architecture by default enables compression only on all virtual machines created. You can optionally disable compression for new writes on virtual machines using SPBM policies. This new compression system is up to 4x more efficient than the old compression system allowing up to 8x compression. It is advised to leave compression enabled as it will improve performance. Data is now compressed before it is sent over the network which will result in potentially higher throughput

and less networking contention. Disabling compression on a virtual machine does not re-write the data but simply stops compressing new writes. To disable vSAN ESA compression, select the "No space efficiency" storage rule SPBM policy (See image below).



Failure Tolerance Method (vSAN 8)

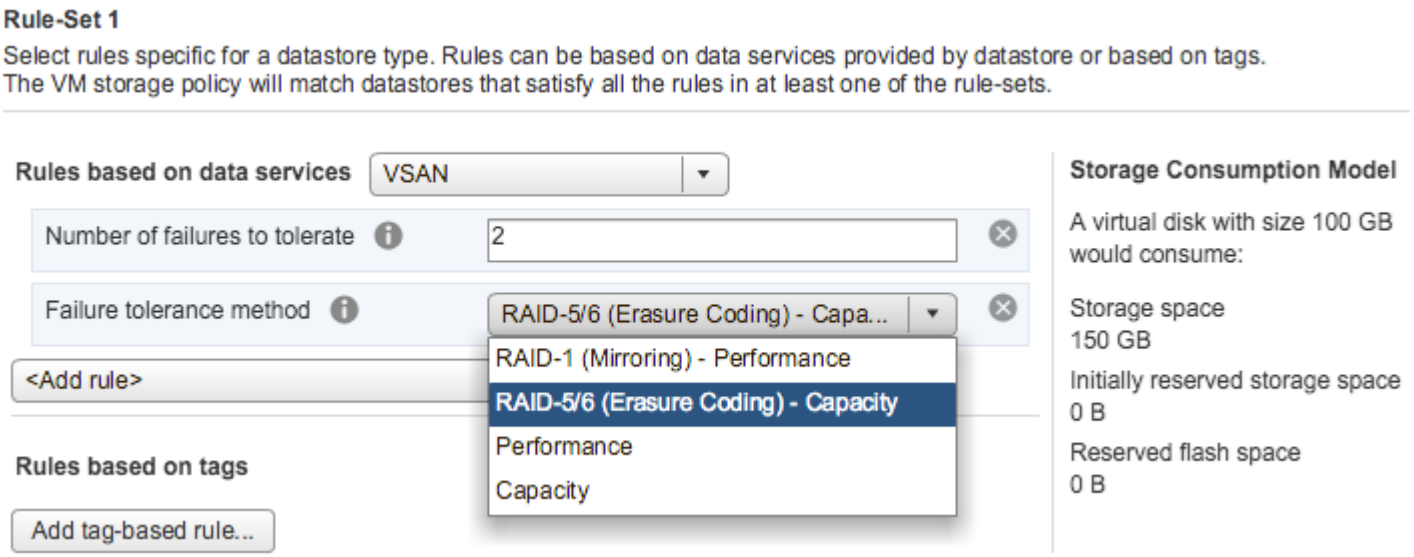
The policies "Failure Tolerance Method" and Failures to Tolerate have been unified into a single option that selects the resiliency option.



Failure Tolerance Method (vSAN OSA Prior to vSphere 8) and Number of Failures to Tolerate Policies

RAID-1 (Mirroring) was used as the failure tolerance method. vSAN 6.2 adds RAID-5/6 (Erasure Coding) to all-flash configurations. While mirroring techniques excel in workloads where performance is the most important factor, they are expensive in terms of capacity required. RAID-5/6 (Erasure Coding) data layout can be configured to help ensure the same levels of availability while consuming less capacity than RAID-1 (Mirroring)

RAID-5/6 (Erasure Coding) is configured as a storage policy rule and can be applied to individual virtual disks or an entire virtual machine. Note that the failure tolerance method in the ruleset must be set to RAID5/6 (Erasure Coding).



Number of Failures To Tolerate (OSA)

The *NumberOfFailuresToTolerate* policy setting is an availability capability that can be applied to all virtual machines or individual VMDKs. This policy plays an important role when planning and sizing storage capacity for vSAN. Based on the availability requirements of a virtual machine, the setting defined in a virtual machine storage policy can lead to the consumption of as many as four times the capacity of the virtual machine.

For “n” failures tolerated, "n+1" copies of the object are created and "2n+1" hosts contributing storage are required. The default value for *NumberOfFailuresToTolerate* is 1. This means that even if a policy is not chosen when deploying a virtual machine, there will still be one replica copy of the virtual machine’s data. The maximum value for *NumberOfFailuresToTolerate* is 3.

vSAN 6.0 introduced the concept of fault domains. This allows vSAN to tolerate not just host failures, but also environmental failures such as rack, switch and power supply failures by locating replica copies of data in different locations. When working with fault domains, to tolerate “n” number of failures, "n+1" copies of the object are once again created but now "2n+1" fault domains are required. Each fault domain must contain at least one host contributing storage. Fault domains will be discussed in more detail shortly.

Failures to Tolerate sizing consideration

If the *NumberOfFailuresToTolerate* is set to 1, two replica mirror copies of the virtual machine or individual VMDKs are created across the cluster. If the number is set to 2, three mirror copies are created; if the number is set to 3, four copies are created.

For FTT=1 the storage overhead will be 1.33X rather than 2X. In this case, a 20GB VMDK would use on 27GB instead of the 40GB traditionally used by RAID-1.

For FTT=2 the storage overhead will be 2X rather than 3X. In this case, as 20GB VMDK will use 40GB instead of 60GB.

	Tolerated Failures	RAID-1 (Mirroring)		RAID-5/6 (Erasure Coding)		Erasure Coding Space Savings vs. Mirroring
		Minimum Hosts Required	Total Capacity Requirement*	Minimum Hosts Required	Total Capacity Requirement*	
FTT=0	0	3	1x	n/a	n/a	n/a
FTT=1	1	3	2x	4	1.33x	33% less
FTT=2	2	5	3x	6	1.5x	50% less
FTT=3	3	7	4x	n/a	n/a	n/a

*Without Deduplication/Compression taken into account.

Erasure coding can provide significant capacity savings over mirroring, but it is important to consider that erasure coding incurs 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.

For more guidance on which workloads will benefit from erasure coding, see VMware vSAN Space Efficiency Technologies .

Virtual Machine Namespace & Swap Considerations

Virtual machines on vSAN datastore consist of objects. vSAN creates a virtual machine namespace (VM home) object when a virtual machine is deployed. When the virtual machine is powered on, a VM swap object is also instantiated whilst the virtual machine remains powered on. Neither the VM home namespace nor the VM swap inherits all of the settings from the VM Storage Policy. These have special policy settings that have significance when sizing a vSAN cluster.

VM Home

As of vSAN 8 Update 1 the namespace object size can be increased using powerCLI from the previous limit of 255GB to allow administrators to store ISOs and Content libraries more easily. This capability should not be confused with vSAN file services which should be used for general purpose file storage. See the vSAN operations guide for more information.

The following syntax example shows creating a directory, querying the size of the directory, increasing the size of the directory and deleting the directory.

The VM home namespace on vSAN is by default a 255 GB thinly provisioned object. Each virtual machine has its own VM home namespace. If certain policy settings are allocated to the VM home namespace, such as *Object Space Reservation* and *Flash Read Cache Reservation* , much of the storage capacity and flash resources could be wasted unnecessarily. The VM home namespace would not benefit from these settings. To that end, the VM home namespace overrides certain capabilities of the user provided VM storage policy.

Number of Disk Stripes Per Object: 1
Flash Read Cache Reservation: 0%
Number of Failures To Tolerate: (inherited from policy)
Force Provisioning: (inherited from policy)
Object Space Reservation: 0% (thin)

The VM Home object has the following characteristics.

Type	Component State	Host	Fault Domain
> Hard disk 1 (RAID 1)			
> Hard disk 2 (Concatenation)			
> Hard disk 3 (RAID 1)			
v VM home (RAID 1)			
Component	Active	h17.satm.eng.vmware.com	
Component	Active	h7.satm.eng.vmware.com	
Witness	Active	h1.satm.eng.vmware.com	
> Virtual machine swap object (RAID 1)			

The RAID 1 is the availability aspect. There is a mirror copy of the VM home object which is comprised of two replica components, implying that this virtual machine was deployed with a *NumberOfFailuresToTolerate* = 1. The VM home inherits this policy setting. The components are located on different hosts. The witness serves as the tiebreaker when availability decisions are made in the vSAN cluster in the event of, for example, a network partition. The witness resides on a completely separate host from the replicas. This is why a minimum of three hosts with local storage is required for vSAN.

The VM Home Namespace inherits the policy setting *NumberOfFailuresToTolerate* . This means that if a policy is created which includes a *NumberOfFailuresToTolerate* = 2 policy setting, the VM home namespace object will use this policy setting. It ignores most of the other policy settings and overrides those with its default values.

VM Swap Object (vSAN 6.7)

Starting in vSAN 6.7 the swap object will inherit the VM home object policy. This provides benefits that FTT values above one can be chosen, as well that the object will be thin by default which will provide significant space savings.

VM Swap Object (Pre-6.7)

Prior to vSAN 6.7 The virtual machine swap object also has its own default policy, which is to tolerate a single failure. It has a default stripe width value, is thickly provisioned, and has no read cache reservation.

The virtual machine swap object did not inherit any of the settings in the VM Storage Policy. With one exception it always uses the following settings:

- Number of Disk Stripes Per Object: 1 (i.e. no striping)
- Flash Read Cache Reservation: 0%
- Number of Failures To Tolerate: 1
- Force Provisioning: Activated
- Object Space Reservation: 100% (thick)

Starting in 6.2 a new advanced configuration parameter activates the deactivation of object space reservation for VM Swap. *Swap Thick Provision* deactivated if set to 1 will make the swap object Thin. As virtual machines are powered on this setting will be changed. For memory dense environments using linked clones such as Horizon View, this should yield significant capacity savings. For additional guidance see this explanation of how to set and change this setting.

Deltas Disks Created for Snapshots

Delta disks, which are created when a snapshot is taken of the VMDK object, inherit the same policy settings as the base disk VMDK.

Note that delta disks are also not visible in the UI when VM Storage Policies are examined. However, the VMDK base disk is visible and one can deduce the policy setting for the snapshot delta disk from the policy of the base VMDK disk. This will also be an important consideration when correctly designing and sizing vSAN deployments.

Snapshot memory

In vSAN 5.5, snapshots of virtual machines that included memory snapshots would store the memory image in the VM home namespace. Since the VM home namespace is of finite size (255GB), it means that snapshots of virtual machines that also captured memory could only be done if the memory size was small enough to be saved in the VM home namespace.

In 6.x, memory snapshots are instantiated as objects on the vSAN datastore in their own right, and are no longer limited in size. However, if the plan is to take snapshots that include memory, this is an important sizing consideration.

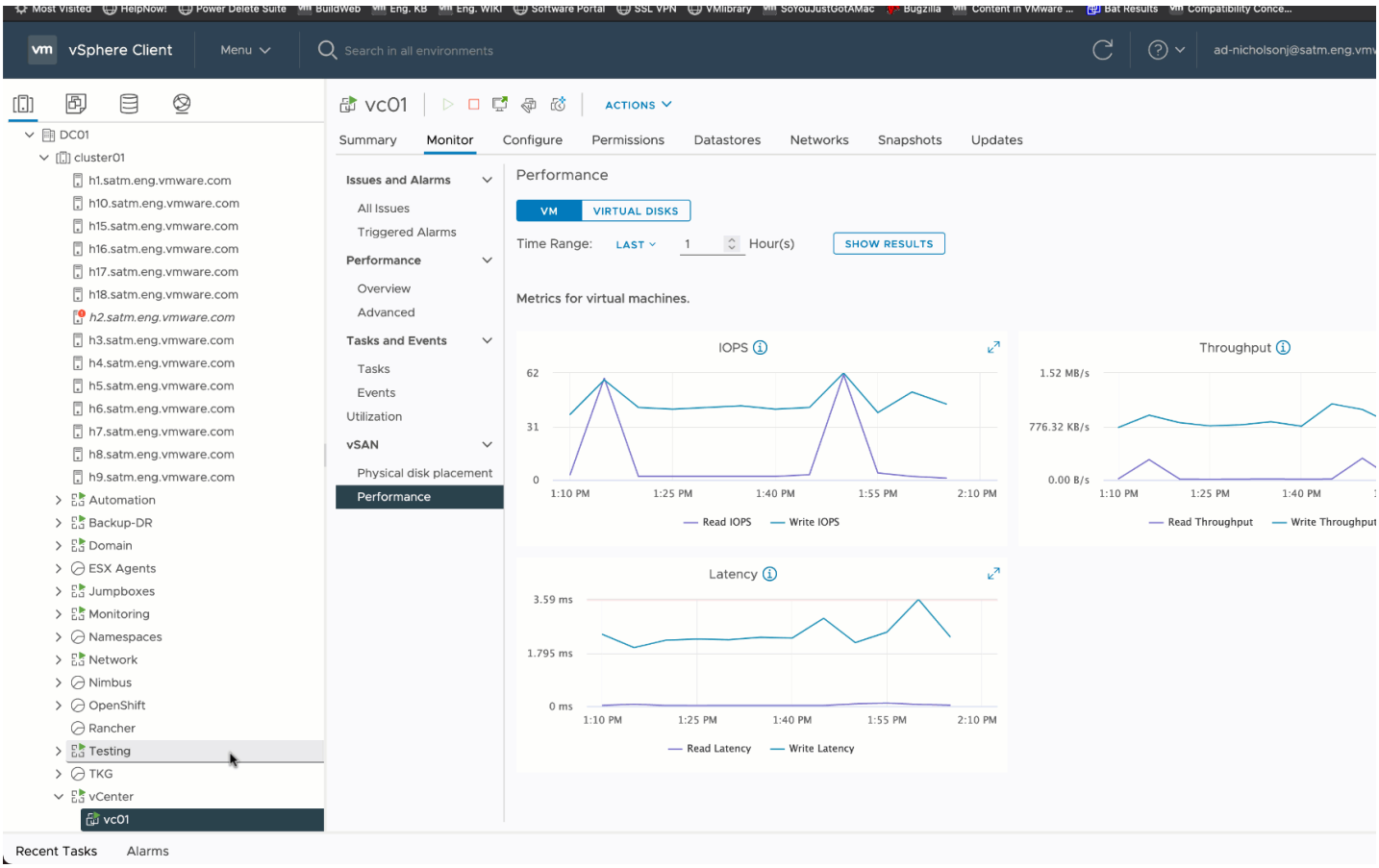
Shortly, a number of capacity sizing examples will be looked at in detail, and will take the considerations discussed here into account.

Changing a VM Storage Policy Dynamically

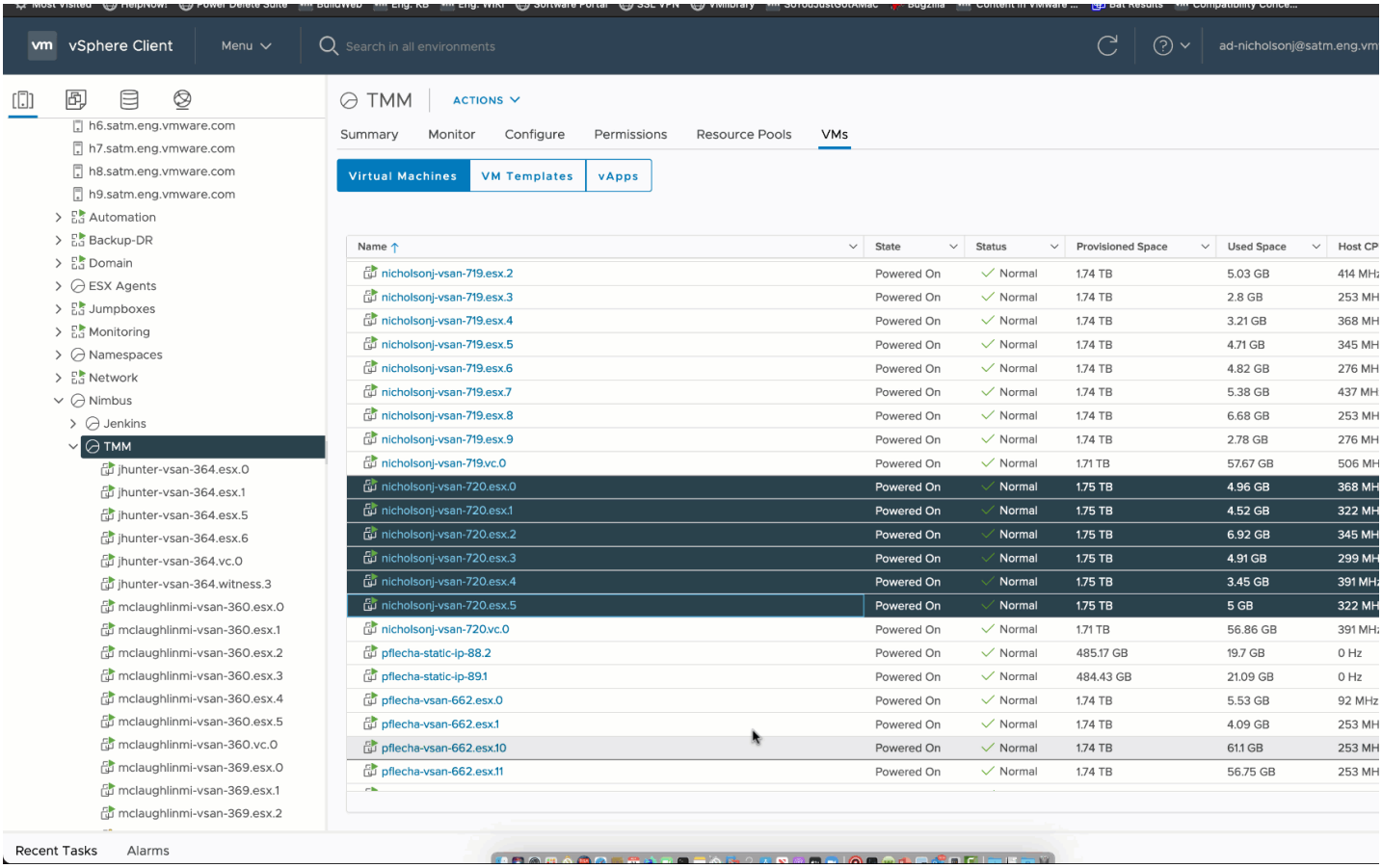
Dynamically changing the policy associated with a virtual machine in a non-disruptive manner has been a core feature of vSAN from the earliest version.

The following examples show how this process works:

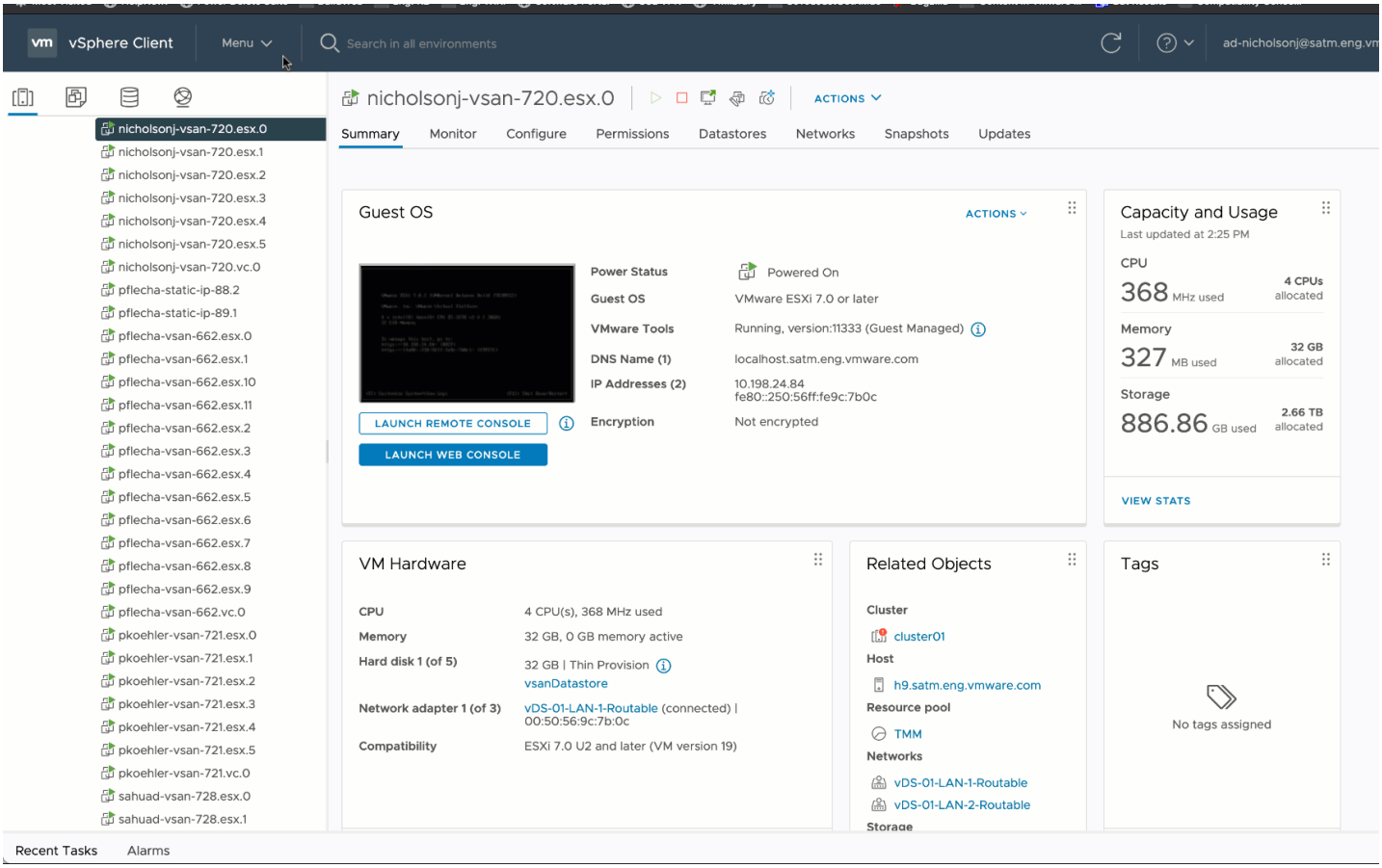
1. Changing the policy associated with a specific virtual machine.



2. Changing the policy associated with multiple Virtual Machines



3. Changing the RAID level associated with a default management policy



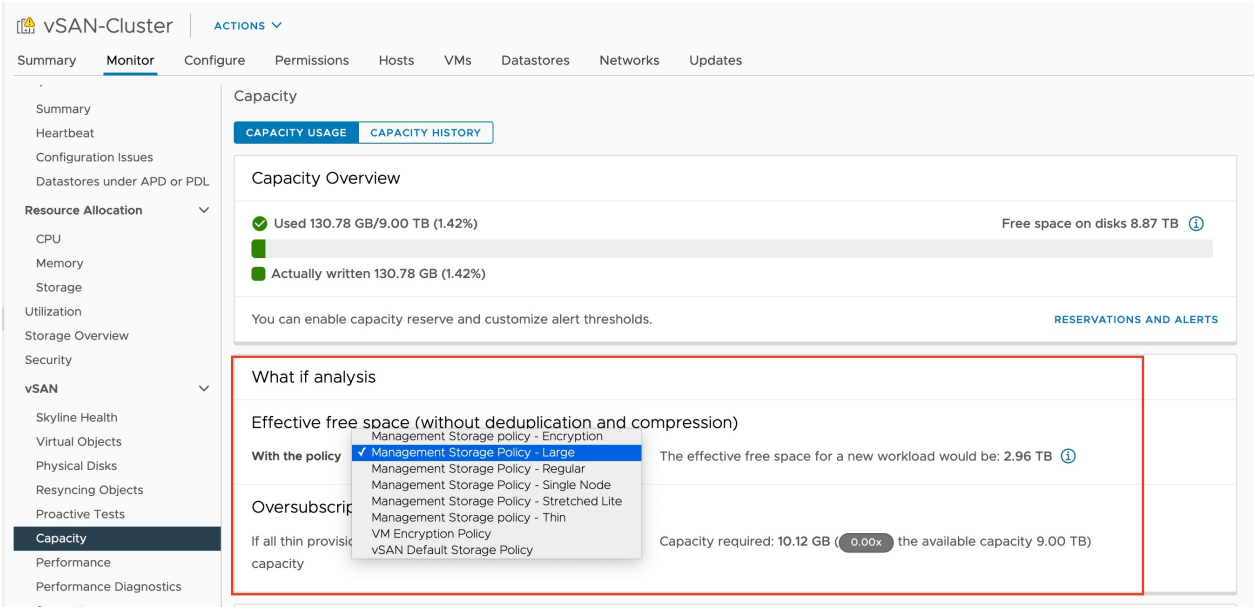
Moreover, it is ***not imperative*** you get these policies right on the first try because changing policies in a running vSAN environment does not require reformatting disks.

Capacity considerations of policy changes

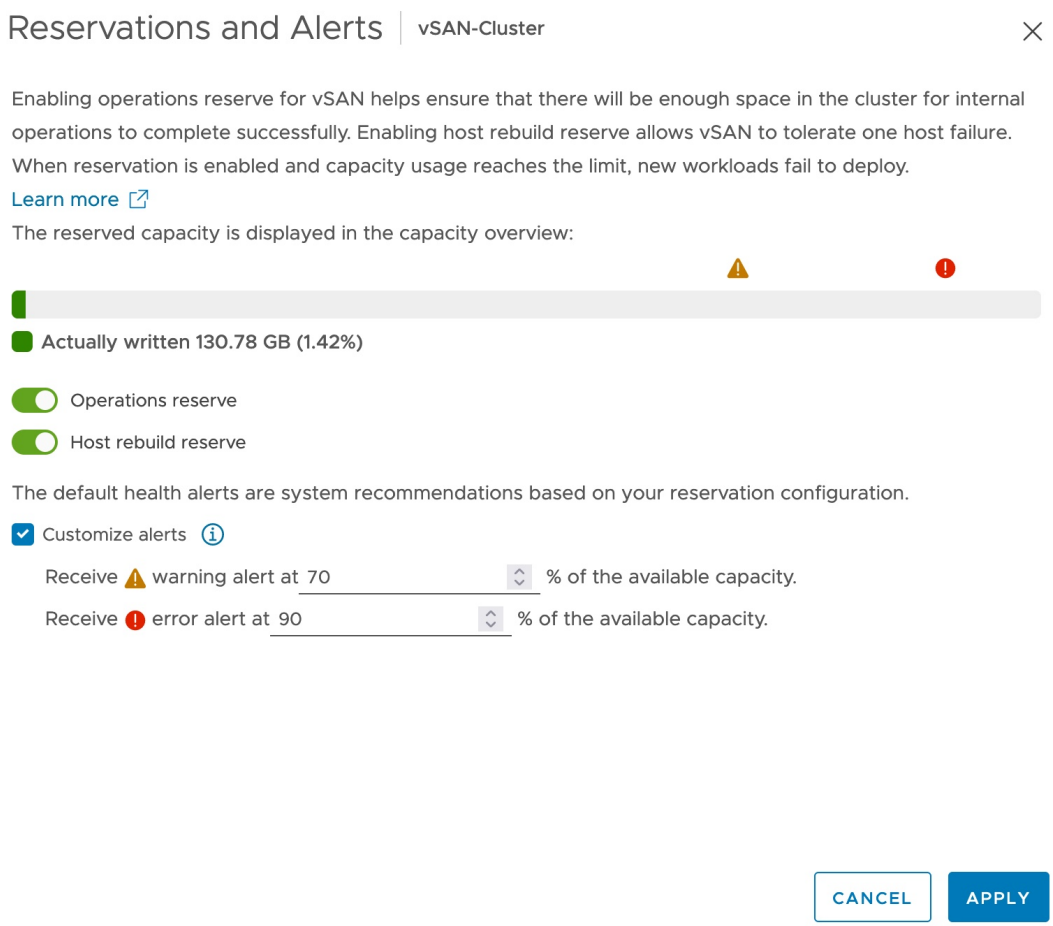
It is important for vSAN administrators to be aware of how vSAN changes a VM Storage Policy dynamically, especially when it comes to sizing. Administrators need to be aware that changing policies dynamically may lead to an increase in the amount of space consumed on the vSAN datastore.

When administrators make a change to a VM Storage Policy and then apply this to a virtual machine to make the change, vSAN will attempt to find a new placement for a replica with the new configuration. If vSAN fails to find a new placement, the reconfiguration will fail. In some cases existing parts of the current configuration can be reused and the configuration just needs to be updated or extended. For example, if an object currently uses NumberOfFailuresToTolerate=1, and the user asks for NumberOfFailuresToTolerate =2, if there are additional hosts to use, vSAN can simply add another mirror (and witnesses).

The vSAN Capacity Overview allows an administrator to model what free space on a cluster will look like with a different policy assumed for new workloads.



vSAN 7 Update 1 introduced changes to how rebalances are handled. vSAN resyncs get paused when disks groups reach a configurable resync pause fullness threshold. This is to avoid filling up the disks with resync I/O. If the disks reach this threshold, vSAN stops reconfiguration workflows, such as EMM, repairs, rebalance, and policy change. Space for rebalance operations is reserved using the vSAN operational reserve configuration.



Provisioning a Policy that Cannot be Implemented

Another consideration related to VM Storage Policy requirements is that even though there may appear to be enough space in the vSAN cluster, a virtual machine will not provision with certain policy settings.

While it might be obvious that a certain number of spindles is needed to satisfy a stripe width requirement, and that the number of spindles required increases as a *NumberOfFailuresToTolerate* requirement is added to the policy, vSAN does not consolidate current configurations to accommodate newly deployed virtual machines.

For example, vSAN will not move components around hosts or disks groups to allow for the provisioning of a new replica, even though this might free enough space to allow the new virtual machine to be provisioned. . Therefore, even though there may be enough free space overall in the cluster, most of the free space may be on one node, and there may not be enough space on the remaining nodes to satisfy the replica copies for *NumberOfFailuresToTolerate*.

A well balanced cluster, with uniform storage and flash configurations, will mitiage this issue significantly.

Provisioning with the Default Policy

With vSAN 5.5, VM Storage Policies should always be used. Failure to select a policy will not deploy the VM's disks as thin. Rather it will use the default policy which implements the virtual machine Provisioning wizard's default VMDK provisioning format, which is Lazy-Zero-Thick. vSAN 6.x has a default VM Storage Policy that avoids this scenario.

Best practice: In vSAN 5.5, always deploy virtual machines with a policy. Do not use the default policy if at all possible. This is not a concern for vSAN 6.x where the default policy has settings for all capabilities.

Summary of Storage Policy Design Considerations

- Any policies settings should be considered in the context of the number of components that might result from said policy.
- *StripeWidth* may or may not improve performance for hybrid configurations; it will have little to offer for all-flash configurations.
- *FlashReadCacheReservation* should be used with caution, and only when a specific performance issue has been identified.
- *NumberOfFailuresToTolerate* needs to take into account how much additional capacity will be consumed, as this policy setting is incremented.
- When configuring *NumberOfFailuresToTolerate*, consideration needs to be given to the number of hosts contributing storage, and if using fault domains, the number of fault domains that contain hosts contributing storage.
- *ForceProvisioning* will allow non-compliant VMs to be deployed, but once additional resources/capacity become available, these VMs will consume them to become compliant.
- VM's that have been force provisioned have an impact on the way that maintenance mode does full data migrations, using "Ensure accessibility" rather than "Full data migration".
- All virtual machines deployed on vSAN (with a policy) will be thin provisioned. This may lead to over-commitment that the administrator will need to monitor.

How to create a policy can be found here Creating and Assigning a Storage Policy.

Host Design Considerations

The following are a list of questions and considerations that will need to be included in the configuration design in order to adequately design a vSAN Cluster.

CPU Considerations (OSA)

AMD EPYC™ deliver best results when used with vSAN 7 Update 2 or later. Changes were made in the hypervisor to better accommodate the architecture of AMD-based chipsets.

When selecting a CPU consider the following metrics:

- Desired sockets per host
- Desired cores per socket
- Desired number of VMs and thus how many virtual CPUs (vCPUs) required
- Desired vCPU-to-core ratio
- Provide for a 10% CPU overhead for vSAN

vSAN File Services currently allocated 4 vCPU per host, however, this allocation is not "reserved" and is elastic based on the DRS group created for this purpose.

When using vSAN encryption, and Data In Transit (DIT) encryption, note that newer CPU generations have improved encryption offload capabilities. Consult with your CPU vendor, if specific SKUs of CPU may offer superior encryption offload capabilities.

Network offload considerations

Network interface cards that have CPU offload features (LRO/TSO, VxLAN, NUMA aware drivers, vSAN RDMA Support) can be leveraged to lower the CPU requirements to transport network traffic. For network cards compatible with vSAN RDMA please see the vSAN VCG.

Memory Considerations (OSA)

- Desired memory for VMs
- A minimum of 32GB is required per ESXi host for full vSAN functionality (5 disk groups, 7 disks per disk group)
- For more information see "Understanding vSAN 6.x Memory Considerations"

Existing vSAN memory overhead can be found within the UI under **Monitor** > **vSAN** > **Support** > **Performance for Support**. From the Performance Dashboards drop-down, under More Dashboards, select **Memory** > **vSAN Memory**

vSAN File Services currently allocated 4GB of memory per host, however this allocation is not "reserved" and is elastic based on the DRS group created for this purpose.

Host Storage Requirement

For best results in estimating storage requirements use the VSAN ReadyNode Sizer.

- Number of VMs, associated VMDKs, size of each virtual machine and thus how much capacity is needed for virtual machine storage
- Memory consumed by each VM, as swap objects will be created on the vSAN datastore when the virtual machine is powered on
- Desired *NumberOfFailuresToTolerate* setting, as this directly impacts the amount of space required for virtual machine disks
- Snapshots per VM, and how long maintained
- Estimated space consumption per snapshot

Boot Device Considerations

Boot Devices

Boot device requirements have changed as of vSphere 7. See this KB for more information. It is strongly recommended to avoid USB and SD card boot devices and instead chose more resilient and performant boot media.

M.2 form factor SSD devices are supported and activate significant partition size for local logs, traces, core dumps. It is recommended to not run virtual machines from these boot devices.

vSAN 6.0 introduced SATA DOM as a supported ESXi boot device

For information on using a controller for boot devices and vSAN See KB2129050

ESXi boot from SAN can be used with vSAN.

Note: ESXi does not support boot devices configured using software RAID.

Auto Deploy

vSAN supports Stateful AutoDeploy. Stateless AutoDeploy is currently not supported.

Core Dump

On upgrade to 6.7, the active core dump partition will get increased according to the conditions of the host.
Previously for embedded installs, only a 100MB crash dump partition was created.
While this could be resized using KB2147881, it will now be automatically increased if space is available.
https://kb.vmware.com/s/article/2147881

Without vSAN activated:
For every 1 TB of DRAM, there should be a core dump size partition of 2.5 GB

With vSAN activated:
In addition to the core dump size, the physical size of the size of caching tier SSD(s) in GB will be used as the basis of calculation the additional core dump size requirements
The base requirement for vSAN is 4GB
For every 100GB cache tier, 0.181GB of space is required
Every disk group needs a base requirement of 1.32 GB
Data will be compressed by 75%

Logs and Traces

If still using deprocated USB and SD devices, the logs and traces reside in RAM disks which are not persisted during reboots.

- Consider redirecting logging and traces to persistent storage when these devices are used as boot devices
- VMware does not support storing logs and traces on the vSAN datastore. These logs may not be retrievable if vSAN has an issue which impacts access to the vSAN datastore. This will hamper any troubleshooting effort.
- VMware KB 1033696 has details on how to redirect scratch to a persistent datastore.
- To redirect vSAN traces to a persistent datastore, esxcli vsan trace set command can be used. Refer to the vSphere command-line documentation for further information.
- vSAN traces are written directly to SATADOMs devices; there is no RAM disk used when SATA DOM is the boot device. Therefore, the recommendation is to use an SLC class device for performance and more importantly endurance.

TPM Hardware Consideration

TPM Devices

vSAN as of vSphere 7 Update 3 Supports the use of TPM 2.0 devices to cache encryption keys used by vSAN Encryption wither both the vSphere Native Key Provider (NKP), as well as external key management servers (KMS). TPM 1.2 devices have been deprecated and should not be used in hosts. TPM devices can also be used for host attestation and configuration encryption. For more information see the following documentation. This functionality requires UEFI used for server boot.

Considerations for Compute-Only Hosts

The following example will provide some background as to why VMware recommends uniformly configured hosts and not using compute-only nodes in the cluster.

Assume a six-node cluster, and that there are 100 virtual machines running per ESXi host in a cluster, and overall they consume 2,000 components each. In vSAN 5.5, there is a limit of 3,000 components that a host can produce. If all hosts in the cluster were to equally consume components, all hosts would consume ~2,000 components to have 100 running VMs in the above example. This will not give rise to any issues.

Now assume that in the same six-node vSAN cluster, only three hosts has disks contributing to the vSAN datastore and that the other three hosts are compute-only. Assuming vSAN achieves perfect balance, every host contributing storage would now need to produce 4,000 components for such a configuration to work. This is not achievable in vSAN 5.5, so care must be taken when deploying virtual machines to vSAN clusters in which not all hosts contribute storage.

While the number of components per host has been raised to 9,000 in vSAN 6.0, the use of compute-only hosts can lead to unbalanced configurations, and the inability to provision the maximum number of virtual machines supported by vSAN.

vSAN 7 Update 1 with HCI mesh activates sharing of storage between clusters. As of vSAN 7 Update 2 clusters participating in this VMware vSAN HCI no longer require a local vSAN datastore.

Best practice: Use uniformly configured hosts for vSAN deployments. While compute-only hosts can exist in a vSAN environment, and consume storage from other hosts in the cluster, VMware does not recommend having unbalanced cluster configurations.

Note: All hosts in a vSAN cluster must be licensed for vSAN--even if they are not contributing capacity, i.e. compute-only nodes. Clusters with no local vSAN datastore mounting remote vSAN clusters do not need vSAN licensing.

Maintenance Mode Considerations

When doing remedial operations on a vSAN Cluster, it may be necessary from time to time to place the ESXi host into maintenance mode. Maintenance Mode offers the administrator various options, one of which is a full data migration. There are a few items to consider with this approach:

1. Consider the number of hosts needed in the cluster to meet the *NumberOfFailuresToTolerate* policy requirements
2. Consider the number of capacity devices left on the remaining hosts to handle stripe width policy requirement when one host is in maintenance mode
3. Consider if there is enough capacity on the remaining hosts to handle the amount of data that must be migrated off of the host being placed into maintenance mode
4. *(Hybrid clusters only)* Consider if there is enough flash cache capacity on the remaining hosts to handle any flash read cache reservations in a hybrid configurations

Blade System Considerations

While vSAN will work perfectly well and is fully supported with blade systems there is an inherent issue with blade configurations in that they are not scalable from a local storage capacity perspective; there are simply not enough disk slots in the hosts. However, with the introduction of support for attached storage enclosures, blade systems can now scale the local storage capacity, and become an interesting solution for vSAN deployments. Introduced with VSAN 7 Update 2, HCI Mesh Compute Clusters support, offers a way to leverage existing blade investments. For more information see the updated HCI Mesh tech note.

External Storage Enclosure Considerations (OSA)

VMware is supporting limited external storage enclosure configurations starting in vSAN 6.0. This will be of interest to those customers who wish to use blade systems and are limited by the number of disk slots available on the hosts. The same is true for rack mount hosts that are limited by disk slots by the way.

Once again, if the plan is to use external storage enclosures with vSAN, ensure the VCG is adhered to with regards to versioning for these devices.

Processor Power Management Considerations

While not specific to vSAN, processor power management settings can have an impact on overall performance. Certain applications that are very sensitive to processing speed latencies may show less than expected performance when processor power management features are activated. A best practice is to select a 'balanced' mode and avoid extreme power-saving modes. There are further details found in VMware KB 1018206.

Cluster Design Considerations

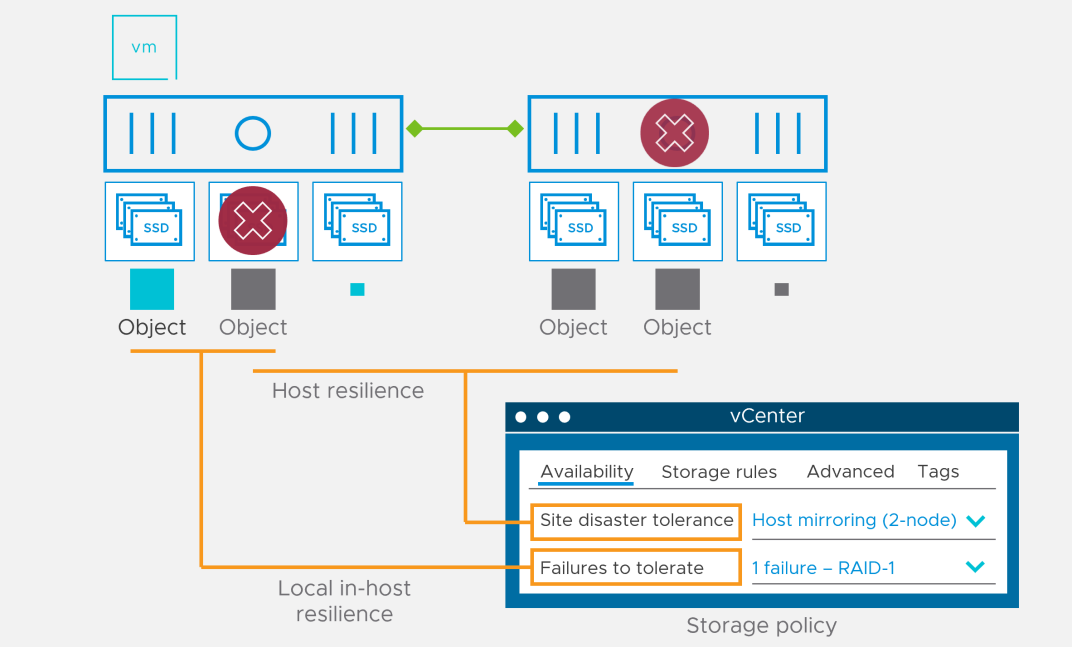
This section of the guide looks at cluster specific design considerations.

Small Cluster Configurations

While vSAN fully supports 2-node and 3-node configurations, these configurations can behave differently than configurations with 4 or greater nodes. In particular, in the event of some kinda of failure, there may not be resources to fully rebuild components on another hosts in the cluster to tolerate another failure.

2-Node Considerations

vSAN 2-Node supports the ability to mirror data within a host. Using 3 disk groups RAID 1 can be used, or with 4 disk groups RAID 5 can be used. For additional information on this capability see this blog or the vSAN 2-Node Cluster Guide.



3 - Node Considerations

vSAN with 3 hosts will use RAID-1 (OSA) or optionally a 2+1 RAID 5 stripe with the new vSAN Express Storage Architecture (ESA). Making sure support agreements and operational staff can replace failed components in a timely manner is critically important in clusters that are not "N+1" of the RAID protection level.

Design decision: Consider 4 or more nodes for the vSAN cluster design for maximum availability. Always use maintenance mode before rebooting a host to maintain availability. This will invoke vSAN's ability to capture missing writes from absent components (new in 7 Update 1).

vSphere HA considerations

vSAN, in conjunction with vSphere HA, provides a highly available solution for virtual machine workloads. If the host that fails is not running any virtual machine compute, then there is no impact to the virtual machine workloads. If the host that fails is running virtual machine compute, vSphere HA will restart those VMs on the remaining hosts in the cluster.

In the case of network partitioning, vSphere HA has been extended to understand vSAN objects. That means that vSphere HA would restart a virtual machine on a partition that still has access to a quorum of the VM's components if the virtual machine previously ran on a partition that lost access due to the partition.

There are a number of requirements for vSAN to operate with vSphere HA.

- vSphere HA uses the vSAN network for communication
- vSphere HA does not use the vSAN datastore as a "datastore heart beating" location. Note external datastores can still be

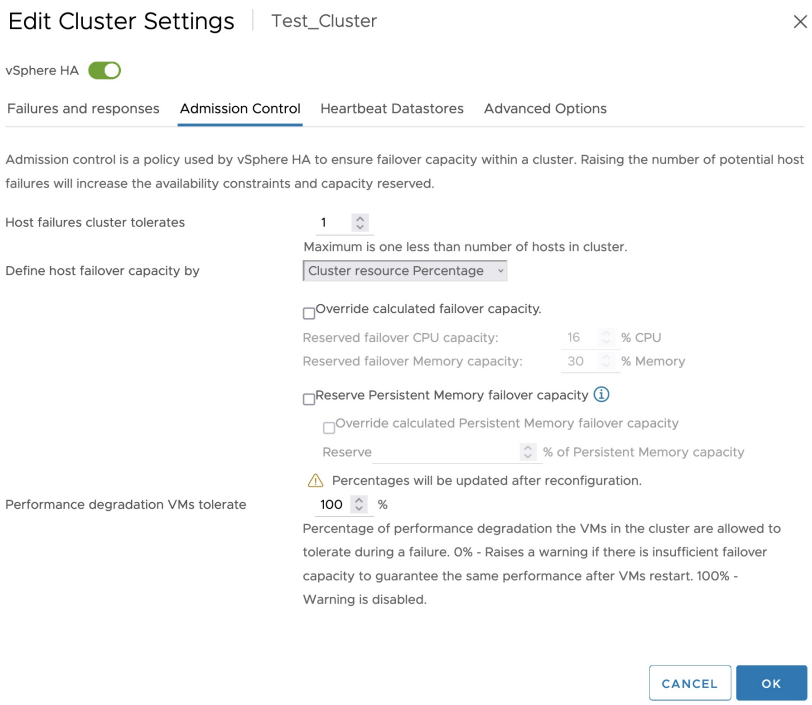
used with this functionality if they exist.

- vSphere HA needs to be deactivated before configuring vSAN on a cluster; vSphere HA may only be activated after the vSAN cluster is configured.

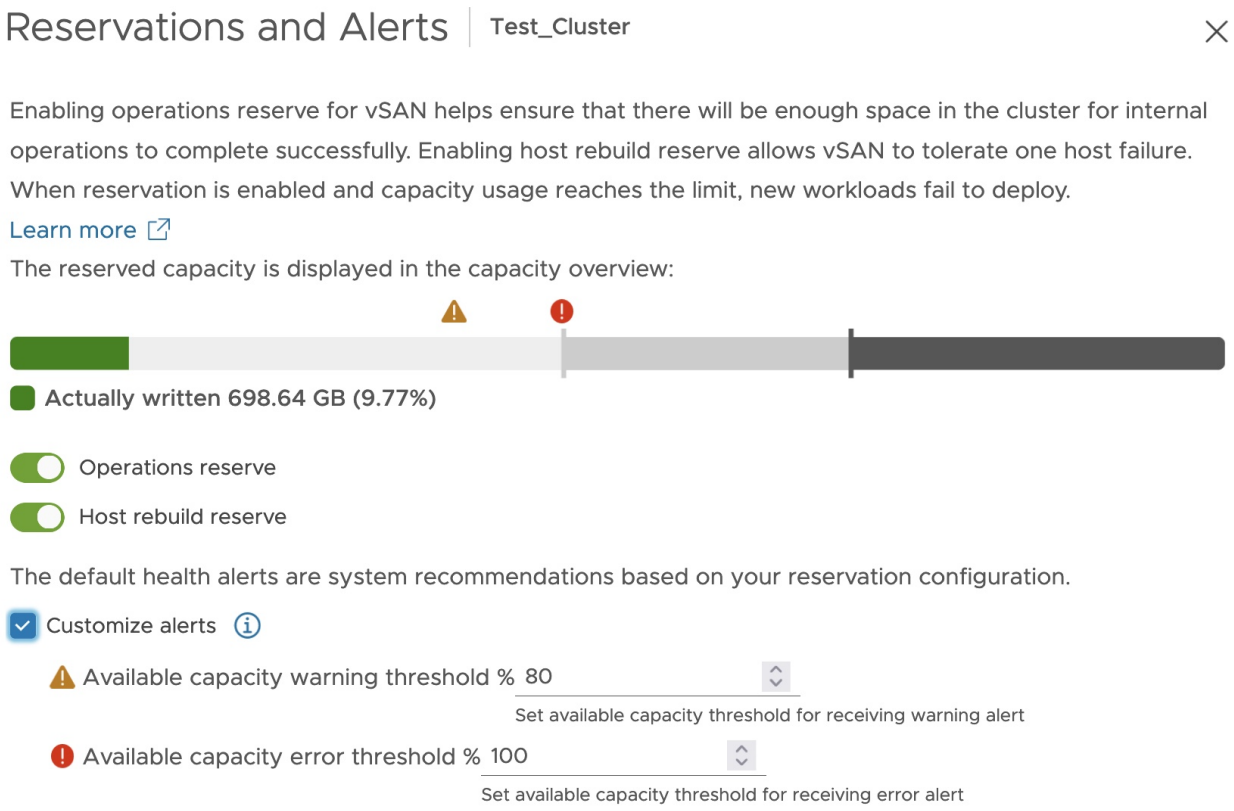
One major sizing consideration with vSAN is interoperability with vSphere HA. Current users of vSphere HA are aware that the NumberOfFailuresToTolerate setting will reserve a set amount of CPU & memory resources on all hosts in the cluster so that in the event of a host failure, there are enough free resources on the remaining hosts in the cluster for virtual machines to restart.

HA Admistion Control and Host Rebuild Reserve

vSphere HA uses admission control to ensure that sufficient resources are reserved for virtual machine recovery when a host fails.



While vSphere Admision Control does not reserve storage capacity, the vSAN host rebuild reserve can be configured to reserve capacity in the event of host failure.



Also note that although VM Component Protection (VMCP), which provides the ability to respond to an APD or PDL scenario, can be configured on a vSAN Cluster it does not have any impact on VMs running on a vSAN Datastore. VMCP only applies to traditional storage or vSAN HCI Mesh usage at the moment.

Heartbeat Datastore Recommendation

Heartbeat datastores are not necessary for a vSAN cluster, but like in a non-vSAN cluster, if available, they can provide additional benefits. VMware recommends provisioning Heartbeat datastores when the benefits they provide are sufficient to warrant any additional provisioning costs.

Host Isolation Addresses Recommendations

The HA agent on a host declares a host isolated if it observes no HA agent to agent network traffic, and if attempts to ping the configured isolation addresses fail, and when no leader election traffic has been observed and it has declared itself as leader. Thus, isolation addresses prevent an HA agent from declaring its host isolated if, for some reason, the HA agent cannot communicate with other HA agents, such as the other hosts having failed. HA allows you to set 10 isolation addresses.

- When using vSAN and vSphere HA configure an isolation addresses that will allow all hosts to determine if they have lost access to the vSAN network. For example: utilize the default gateway(s) of the vSAN network(s). If the vSAN network is non-routable and a single-host partition is possible, then provide pingable isolation addresses on the vSAN subnet. Isolation addresses are set using the vSphere HA advanced option ***das.isolationAddressX.***
- Configure HA not to use the default management network's default gateway. This is done using the vSphere HA advanced option ***das.useDefaultIsolationAddress=false***
- If isolation and partitions are possible, ensure one set of isolation addresses is accessible by the hosts in each segment during a partition.
- For the isolation address, consider using a Switch Virtual Interface (SVI) on the vSAN subnet. Make sure to isolate it from the routing tables and or use Access Control Lists or a VRF to prevent this allowing routing into the vSAN subnet.

Isolation Response Recommendations

The HA isolation response configuration for a VM can be used to ensure the following during a host isolation event:

- To avoid VM MAC address collisions if independent heartbeat datastores are not used. Note: These can be caused by the FDM leader restarting an isolated VM resulting in 2 instances of the same VM on the network.

- To minimize the likelihood that the memory state of the VM is not lost when its host becomes isolated.

The isolation response selection to use depend on a number of factors. These are summarized in the tables below. The tables include recommendations for vSAN and non vSAN virtual machines since clusters may contain a mixture of both.

Decisions Table 1

Type of VM	Host will retain access to a VM storage	VMs will retain access to VM network	Recommended Isolation Policy	Rationale
Non-VSAN	Yes	Yes	Leave Powered On	VM is running fine, why power it off?
Non-VSAN	Yes	No	Leave Powered On or Shutdown	Shutdown if VM network access is important to allow FDM master to restart the VM
VSAN and non-VSAN	No	Yes	Power Off or Leave Powered On	See Table Below
VSAN and non-VSAN	No	No	Power Off or Leave Powered On	See Table Below

Decisions Table 2

Is it likely that all hosts will be isolated when one is	Are there heartbeat datastores and if so, will hosts still have access to them when isolated	Is VM memory state important	VMs will retain access to VM network?	Recommended Isolation Policy	Rationale
No	Yes	Yes	N/A	Leave Powered On	Memory state is important so leave VM powered on. HB datastores will ensure that HA does not start a second copy of the VM
No	No	N/A	Yes	Power Off	The FDM master will likely restart the isolated VMs if it does, there will be two instances of each VM on the network. "Power off" prevents this from occurring.
No	No	Yes	No	Leave Powered On	It is possible that the FDM master may not be able to restart the isolated VMs (e.g., there is no capacity), and there is no side effect of leaving the original VM powered on until isolation is resolved. If a second instance is not restarted, when the isolation ends, the original VM will regain access to its storage.
Yes	N/A	N/A	N/A	Leave Powered On	No point in powering off VMs since there will be no FDM master available to restart them. If it does, the original VM will be terminated when the isolation response ends if a 2nd instance of the VM was restarted.

Note: “Shutdown” may also be used anytime “power off” is mentioned if it is likely that a VM will retain access to some of its storage but not all during host isolation. (This is unlikely in the case of a vSAN datastore.) However, note that in such a situation

some of its virtual disks may be updated while others are not, which may cause inconsistencies when the VM is restarted. Further, a shutdown can take longer than power off.

Best practice: activate HA with vSAN 6.x for the highest possible level of availability. However, any design will need to include additional capacity for rebuilding components

Fault Domains

The idea behind fault domains is that we want to be able to tolerate groups of hosts (chassis or racks) failing without requiring additional data copies. The implementation allows vSAN to save replica copies of the virtual machine data in different domains, for example, different racks of compute.

Design decision: Although it is not required, it is best to build fault domains with the same number of hosts per fault domain. Hardware configurations should be consistent across the entire cluster. Do not activate this feature without careful review of the considerations it has for impacting host rebuild and rebalance data placement.

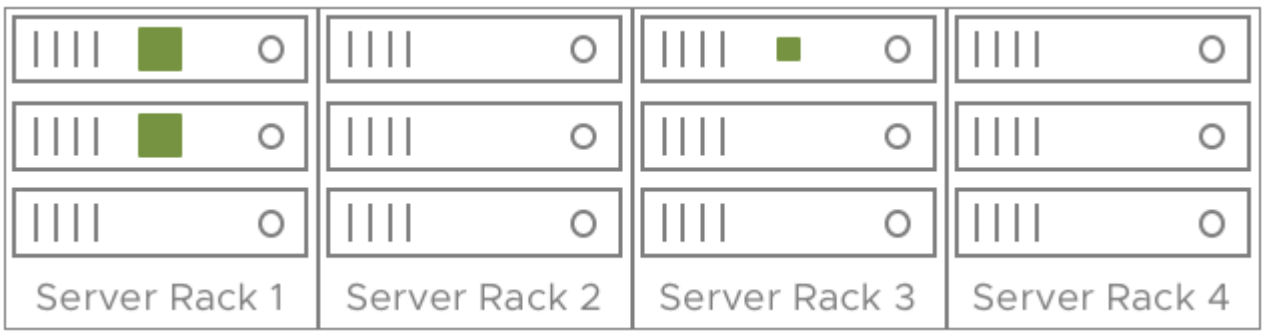
When using RAID 1 and deploying a virtual machine with a *NumberOfFailuresToTolerate* = 1, there are 2n + 1 hosts required (where n = *NumberOfFailuresToTolerate* . This means that to tolerate 1 failure, 3 ESXi hosts are required. To tolerate 2 failures, 5 hosts were required and if the virtual machines are to tolerate 3 failures (maximum), then 7 hosts were required.

RAID 5/6 erasure coding fault tolerance methods added more considerations. See the chart below. For information on When to use RAID 5 see this video.

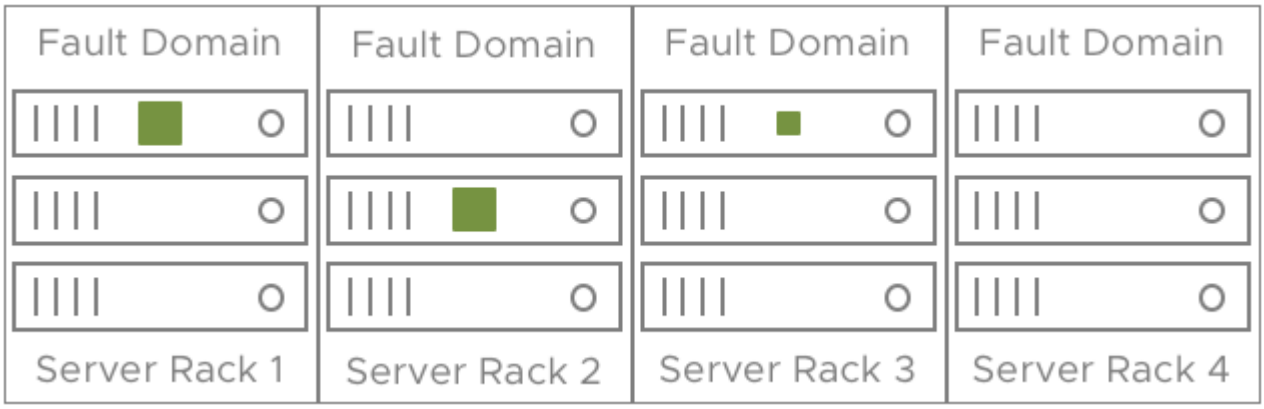
vSAN 8 Express Storage Architecture Introduced a new adaptive RAID 5 policy that starts as small as 3 hosts, and changes stripe size to a 4+1 for more capacity efficiency when using 6 hosts or more, or support a 2+1 policy for smaller clusters.

vSAN OSA RAID

Take the following example where there are 12 hosts in the vSAN cluster, split across four racks. Let's assume that there are three ESXi hosts in each rack. When a virtual machine that tolerates 1 failure is deployed, it is possible for both replicas to be deployed to different hosts in the same rack.



If fault domains are activated, this allows hosts to be grouped together to form a fault domain. This means that no two copies/replicas of the virtual machine's data will be placed in the same fault domain. In this example, mirrored components and the witness are distributed across three racks. The loss of an entire server rack (fault domain) would not result in the loss of availability of the object and virtual machine.



To calculate the number of fault domains required to tolerate failures, use the same equation as before; when deploying a virtual machine with a NumberOfFailuresToTolerate = 1 on a cluster with fault domains, 2n + 1 fault domains (containing 1 or more hosts contributing storage) is required. In this way, the above table can be replicated replacing "hosts" with "fault domains" to understand how many fault domains are needed.

This is true for compute resources as well. In such a scenario, 1 fault domain worth of extra CPU/Memory is needed, as a fault domain failure needs to avoid resource starvation.

You will need an extra fault domain to activate the full restoration of policy compliance once a failure occurs. Be sure to factor this into storage, and compute overheads when designing fault domains.

Design decision: When designing very large vSAN clusters, consider using fault domains as a way of avoiding single rack failures impacting all replicas belonging to a virtual machine. Also, consider the additional resources and capacity requirements needed to rebuild components in the event of a failure.

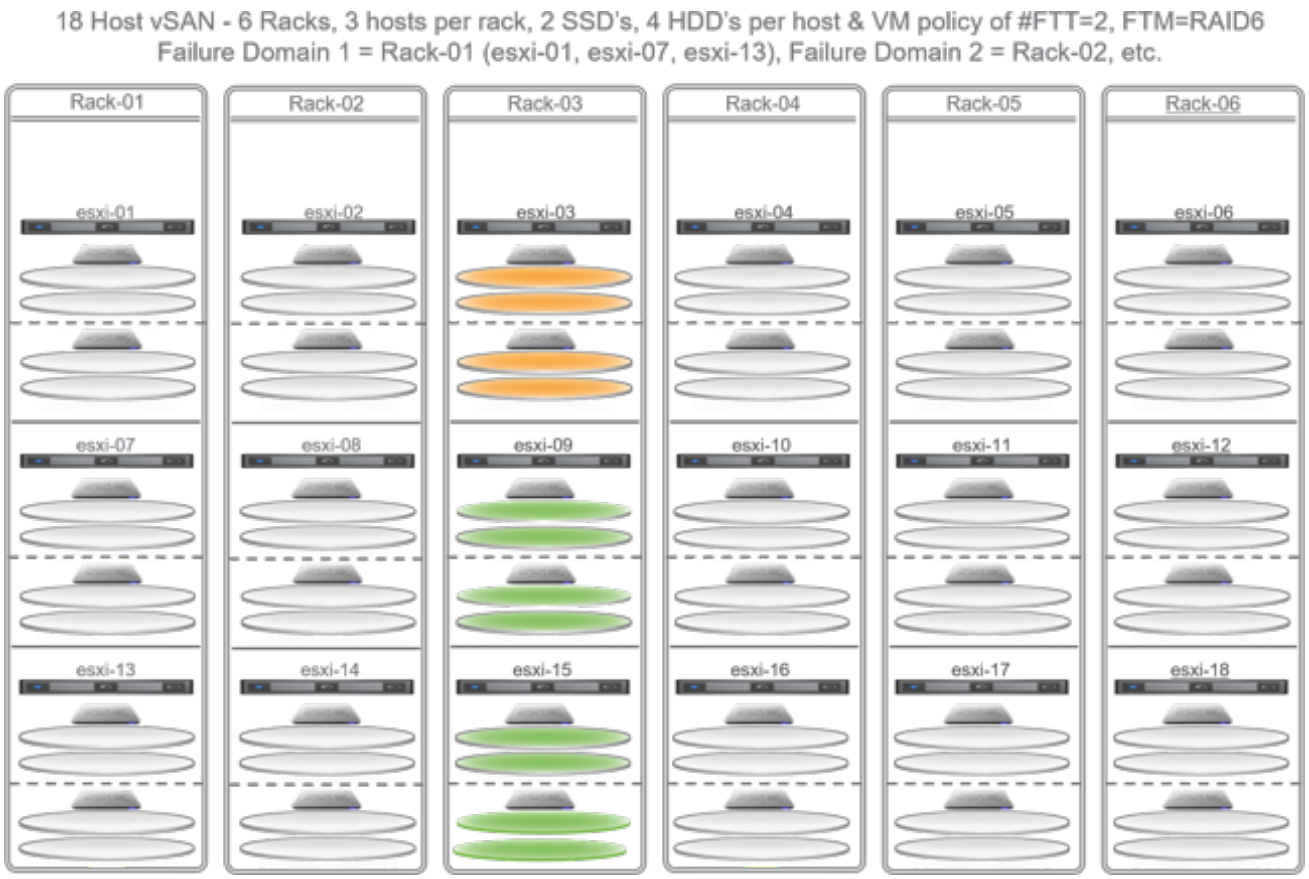
There should be a strategy and an operational run book for how maintenance will be performed for the environment. Design for one of the following:

If Total Fault Domains = Required Fault Domains to satisfy policy:
- Only one host at a time should enter maintenance mode.
- Capacity Used on any one host in the fault domain must be less than Total Free Space across all other hosts in that same fault domain.

Spare capacity may be calculated as follows, where:

D= number of hosts in Fault Domain
T= number of hosts to put into maintenance (or hosts to tolerate the failure of... the 'Tolerance' value)
A= active data per host
(Active data per host x Number of hosts to put into maintenance)
divided by
(number of hosts in the fault domain – hosts to be put in maintenance mode)

(A*T/D-T)



Example:

Using the image above as an example, assuming a policy of FTT=2 with RAID6, to be able to place host esxi-03 into maintenance mode, you may choose to evacuate the data on host esxi-03. In this case, the only hosts available to ingest that data are within the same FD. Therefore, to understand how much capacity to have available:

Assume A=3.7TB consumed capacity per host to be rebuilt/relocated.

(3.7 TB * 1 host) / (3 hosts in FD – 1 host)
3.7/(3-1)
3.7/2
1.85 TB spare capacity required per host in the FD on each of the remaining hosts after taking 1 down.

If Total Fault Domains > Required Fault Domains to satisfy policy:
- Apply method above
---OR---
- Capacity Used on any one host must be less than the Total Free Space across the excess fault domains.
- The best practice is one host in maintenance mode at a time. However, in cases where an entire fault domain must be serviced, hosts from more than one fault domain should not be in maintenance mode simultaneously.

Calculating Capacity Tolerance – Across Fault Domains
If the number of configured Fault Domains exceeds the required Fault Domains as indicated by policy, and there is insufficient capacity within the fault domain to ingest evacuated data, it is possible to burden the additional fault domains with the extra capacity. Therefore capacity calculations must include the number of available or extra fault domains, and determine the amount of spare capacity the hosts in those fault domains must have to ingest the data from the hosts in maintenance mode.

Spare capacity may be calculated as follows, where:

F= number of total Fault Domains
R= number of required fault domains to satisfy the policy
D= number of hosts in Fault Domain
T= number of hosts to put into maintenance (or hosts to tolerate the failure of... your 'Tolerance' value)

A= active data per host
(Active data per host x Number of hosts to put into maintenance)
divided by
(total fault domains – fault domains required) x (number of hosts in each fault domain)

(A*T) / ((F-R)*D))

Example:

Using the image above as an example, again using a policy of FTT=2 and RAID6, to be able to take 2 hosts down for maintenance from FD3, yet still have enough capacity to rebuild impacted data in the event of a host failure within the fault domain, capacity may be calculated as follows:

Assume A=3.7TB consumed capacity per host to be rebuilt/relocated.

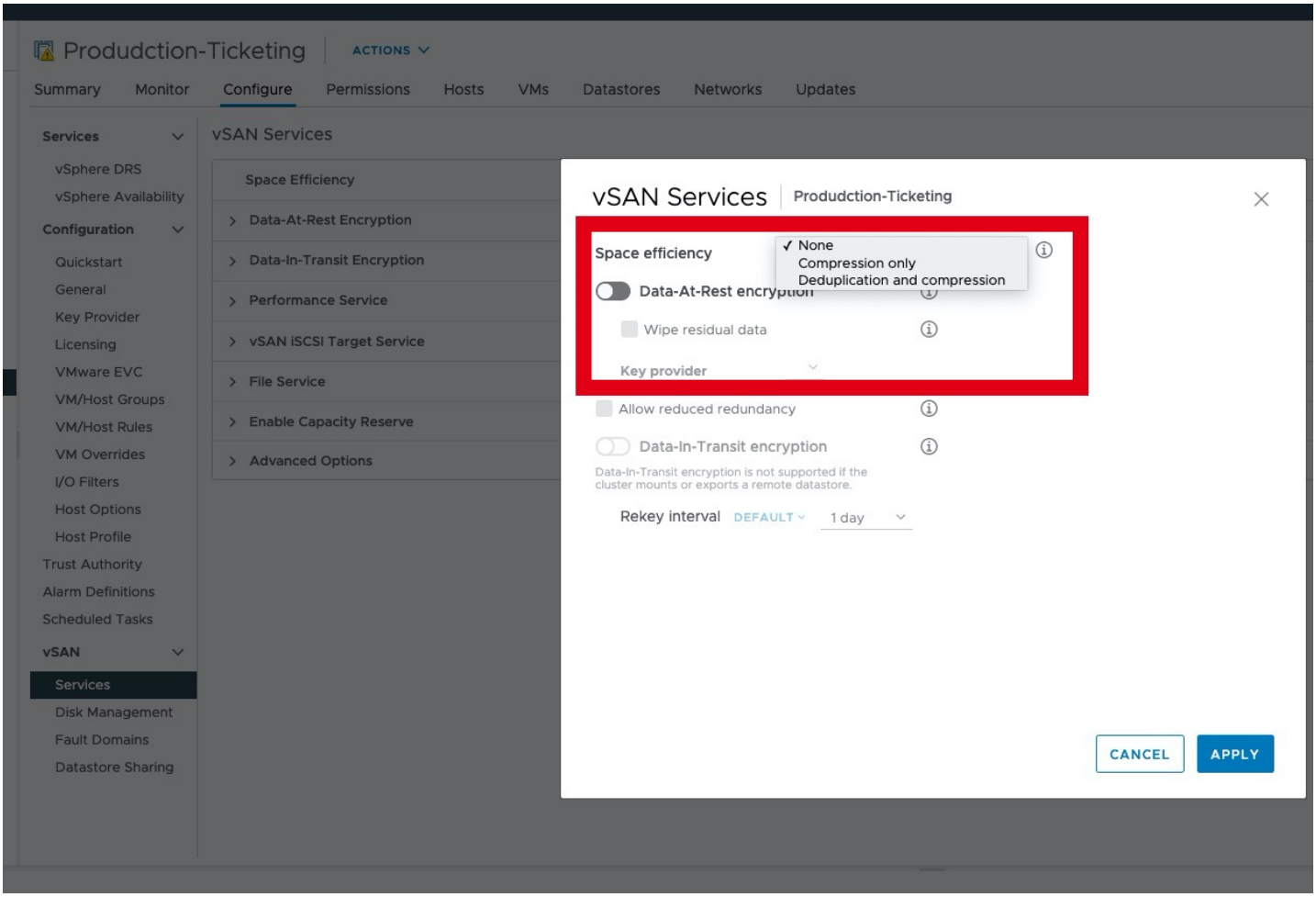
(3.7 TB * 3 hosts) / ((8FD total – 6FD required)*3 hosts per FD)
(3.7*3)/((8-6)*3)
11.1/(2*3)
11.1/6
1.85 TB spare capacity required on each of the 6 hosts in the remaining 2 FD to evacuate FD #3

Additional Content:

- Read "Considerations using vSAN fault domains"

Deduplication and Compression Considerations

In addition to the default data layout, 2 additional options exist for all-flash vSAN clusters to use a more space efficient data layout.



Deduplication and Compression (vSAN OSA) - For VSAN OSA this feature leverages both deduplication as well as compression to maximize space efficiencies.

When this feature is activated, objects will not be deterministically assigned to a capacity device in a disk group but will stripe across all disks in the disk group. This will reduce the need for component re-balancing, and should reduce the need to increase the *NumberOfDiskStripesPerObject* policy to balance random read performance

This option also requires an all-flash cluster to be activated.The *ObjectSpaceReservation* policy when used with deduplication and compression will result in reduced capacity availability as space must be reserved in case the blocks become non-unique. This policy will not change the performance characteristics (the blocks will still be hashed and compressed) but it will impact usable storage.

vSAN 8 (ESA) Compression Only - VSAN 8 Express Storage Architecture by default enables compression only as a cluster service. You can optionally disable compression for new writes on virtual machines using SPBM policies. This new compression system is up to 4x more efficient than the old compression system allowing up to 8x compression.

Compression-Only (OSA) - Introduced in vSAN 7 Update 1. Ideal when looking for a middle ground with space efficiency. For workloads that may yield limited results from deduplication, but still generate compressibility this is an option that reduces the additional compute and IO overhead associated with deduplication.

For more information on recommended use cases see the VMware vSAN Space Efficiency Technologies.

Cluster Size Consideration

The flexibility of cluster design and sizing is one of the key benefits with vSAN. This naturally leads to questions on the performance capabilities of vSAN relative to cluster size. Will an application run faster if the cluster consists of more hosts versus fewer hosts? The simple answer is "no" and has been covered in vSAN Cluster Design - Large Clusters Versus Small Clusters.

The question of performance and cluster size stems not from comparing the total capability of a 4-host cluster versus 64 hosts, but rather, a data center that has dozens or hundreds of hosts, and have questions about the optimal cluster size for performance.

A vSphere cluster defines a boundary of shared resources. With a traditional vSphere cluster, CPU and memory are two types of resources managed. Network-aware DRS is omitted here for clarity. Adding additional hosts would indeed add additional CPU and memory resources available to the VMs running in the cluster, but this never meant that a single 16vCPU SQL server would have more capabilities as hosts are added: It just enlarged the boundary of available physical resources. vSAN powered clusters introduce storage as a cluster resource. As the host count of the cluster increases, so does the availability of storage resources.

Data Placement in vSAN

To understand elements of performance (and availability), let's review how vSAN places data.

With traditional shared storage, the data living in a file system will often be spread across many (or all) storage devices in an array or arrays. This technique sometimes referred to as "wide striping" was a way to achieve improved performance through an aggregate of devices, and allowed the array manufacturer to globally protect all of the data using some form of RAID in a one-size-fits-all manner. Wide striping was desperately needed with spinning disks, but still common with all-flash arrays and other architectures.

vSAN is different: Using an approach for data placement and redundancy most closely resembling an object-based storage system. It is the arbiter of data placement, and which hosts have access to the data. An object, such as a VMDK may have all the object data living on one host. In order to achieve resilience, this object must either be mirrored to some other location (host) in the vSAN cluster or if using RAID-5/6 erasure coding, will encode the data with parity data across multiple hosts (3 or 5 hosts for RAID 5 with vSAN ESA, 4 hosts for RAID-5 with vSAN OSA , 6 hosts for RAID-6). Thanks to Storage Policy-Based Management (SPBM), this can be prescribed on a per-object basis.

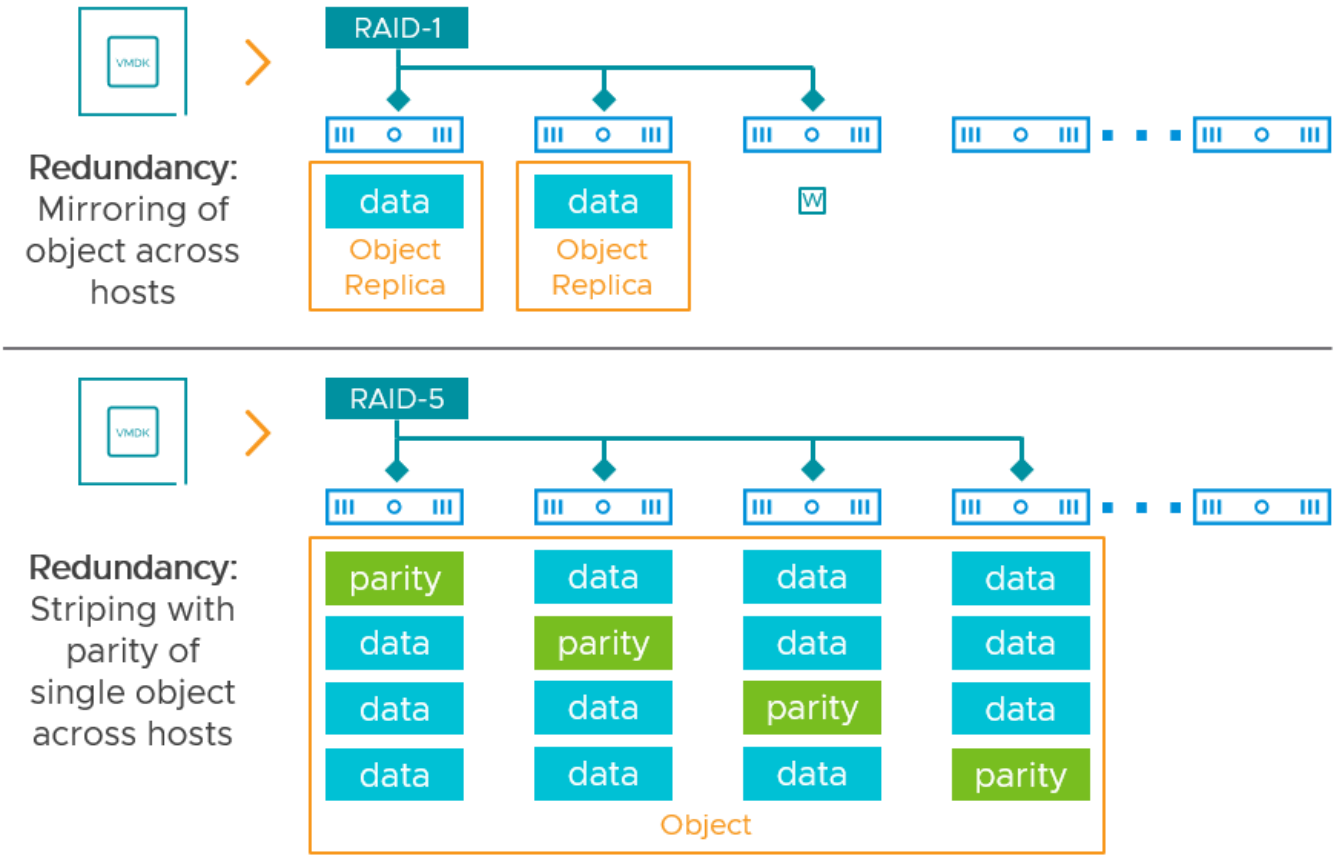
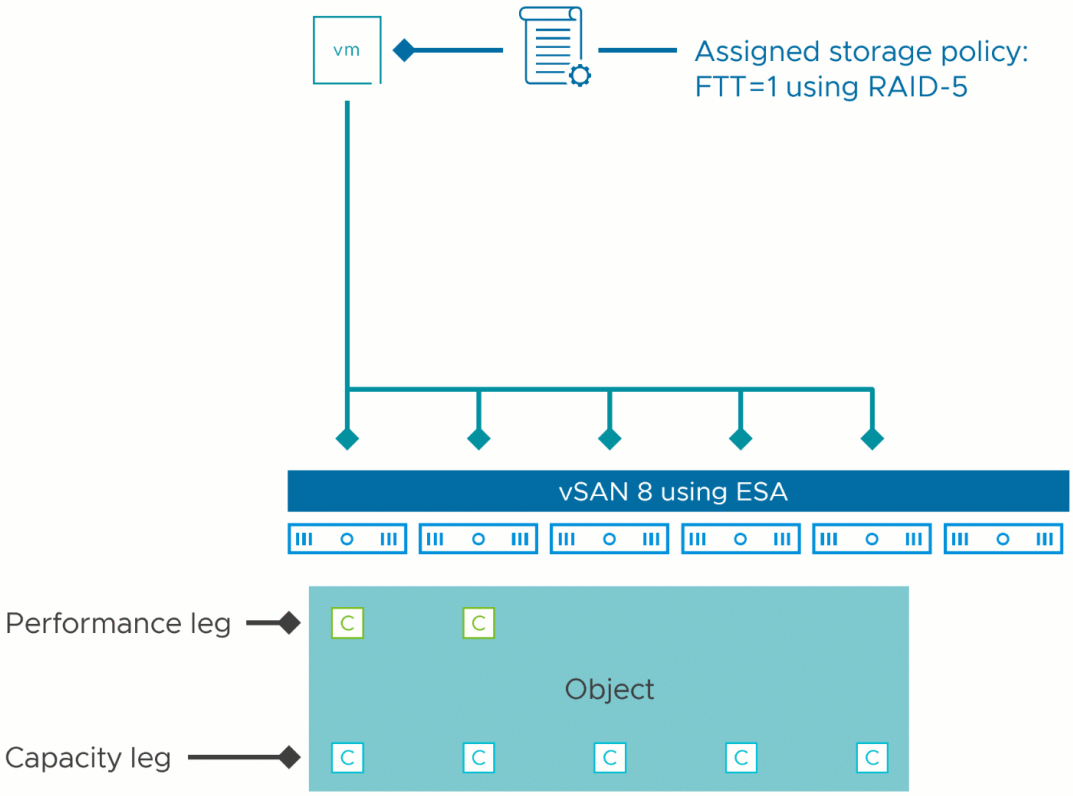


Figure 1. Examples of vSAN's approach to data placement with resilience

New with vSAN 8 ESA, RAID 5 now automatically adjusts the stripe width from 2+1 to 4+1 (or back) when the cluster expands to 5 hosts.



Therefore, whether the cluster is 4 hosts or 64 hosts, the number of hosts it will use to store the VM to its prescribed level of resilience will be the same. Some actions may spread the data out a bit more, but generally, vSAN strives to achieve the desired availability of an object prescribed by the storage policy, with as few hosts as possible.

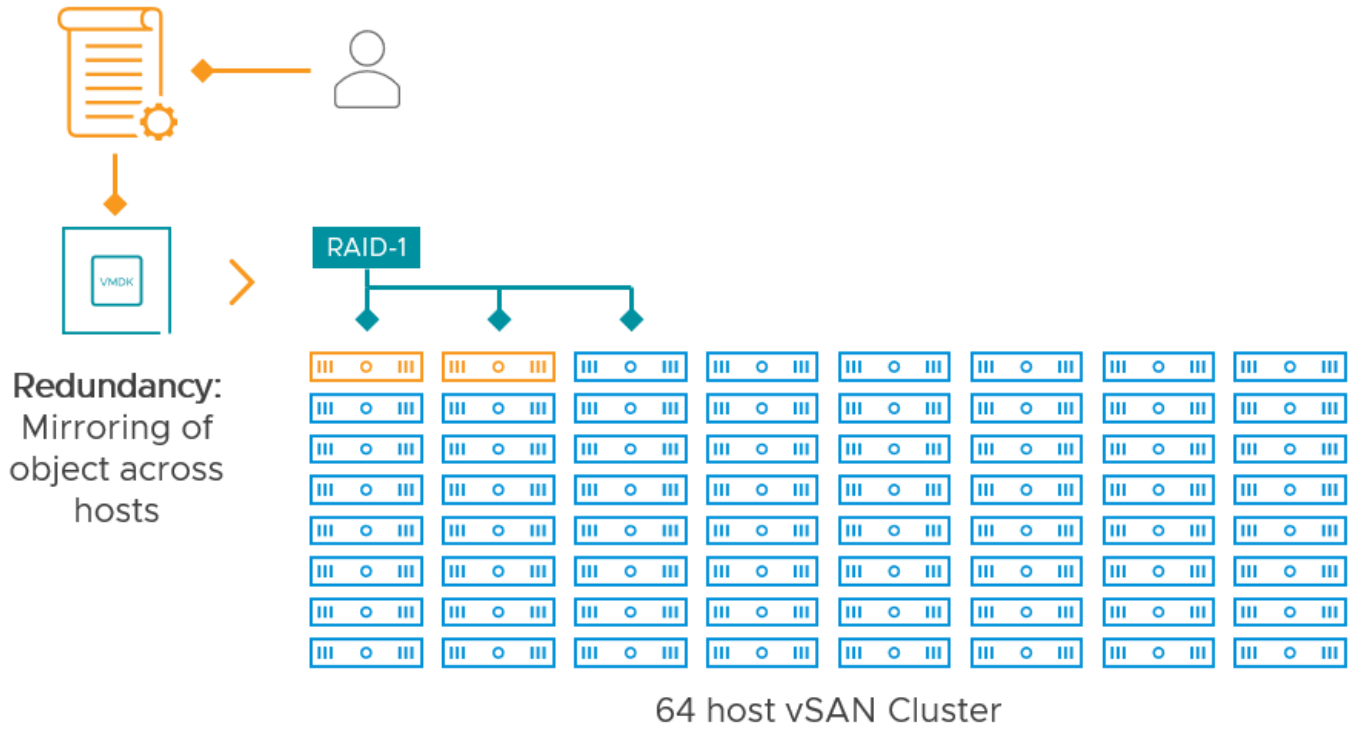


Figure 2. A VM with an FTT=1 policy using mirroring, in a 64 host cluster

The benefit to the approach used by vSAN is superior resilience under failure conditions and simplified scalability. When a level of failure to tolerate (FTT) is assigned to an object, availability (and performance) is only referring to the hosts where the specific object resides. Other failures can occur in the cluster and have no impact to that specific object. Wide-striping, occasionally found in other approaches, can increase the fragility of an environment because it introduces more contributing resources that the data depends on for availability. Wide-striping can make scaling more difficult.

When does cluster size matter as it relates to performance?

The size of a traditional vSphere cluster can impact VM performance when physical resources are oversubscribed. Oversubscription most commonly occurs when there are too many taxing workloads for the given hardware capabilities of the hosts. When there are noticeable levels of contention of physical resources (CPU and memory) created by the VMs, then performance will degrade. DRS aims to redistribute the workloads more evenly across the hosts in the cluster to balance out the use of CPU and memory resources and free up contention.

With a vSAN cluster, storage is also a resource of the cluster - a concept that is different than with traditional three-tier architecture. If the I/O demand by the VMs is too taxing for the given hardware capabilities of the storage devices, then there may be some contention. In this situation, adding hosts to a cluster could improve the storage performance by reclaiming lost performance, but only if contention was inhibiting performance in the first place. Note that the placement and balancing of this data across hosts in the cluster is purely the responsibility of vSAN, and is based on capacity, not performance. DRS does not play a part in this decision making.

With a vSAN cluster, performance is primarily dictated by the physical capabilities of the devices in the storage stack (cache/buffer, capacity tier, storage HBAs, etc.) as well as the devices that provide communication between the hosts: NICs, and switchgear.

Testing of a discrete application or capabilities of a host can be performed with a small cluster or large cluster. It will not make a difference in the result as seen by the application, or the host. If synthetic testing is being performed with tools like HCIBench, the total performance by the cluster (defined by IOPS or throughput) will increase as more hosts are added. This is due to an increase in worker VMs as more hosts are added to the cluster. The stress test results are reporting back the sum across all hosts in the cluster. The individual performance capabilities of each host remain the same.

What about network switches?

A common assumption is that there must be some correlation between the needs of network switching and cluster size. Given the same number of hosts connecting to the switchgear, there is very little correlation between cluster size and demands of the switchgear. The sizing requirements for NICs or the switchgear they connect to is most closely related to the host capacity and type/performance of gear the hosts consist of, the demands of the application, storage policies and data services used, and the expectations around performance.

Figure 3. vSAN cluster size versus demand on network

The performance of the hosts can be impacted if the switchgear is incapable of delivering the performance driven by the hosts. This is a sign the underlying switchgear may not have a backplane, processing power, or buffers capable of sustaining advertised speeds across all switch ports. But this behavior will occur regardless of the cluster sizing strategy used. It is possible that as you add additional hosts you will hit networking buffer limitations, or internal queue limits on the switching ASICs. If a cluster is expected to grow significantly in performance requirements, consider deploying adequately powerful switches at initial deployment. Also, many common top of rack switches are limited to 48 access ports, so larger clusters may require additional throughput between switches in order to prevent these links from becoming bottlenecks. Consider for large clusters that will span multiple switches making sure adequate bandwidth is allocated and strongly consider using leaf/spine CLOS designs. 100Gbps switches that support breakout connections to 4 x 25Gbps connectors are increasingly popular for large clusters. This will allow for even the largest clusters to maintain traffic on a single pair of leaf switches.

RDMA Switch Support

vSAN 7 Update 2 introduces (RDMA Converged over Ethernet version 2 (RCoEv2) support. Please talk to your switch vendor and confirm that RCoEv2 requirements can be met.

Operational considerations of large clusters

Another consideration can be the time it takes to patch a cluster. While many customers take advantage of vSAN's ability to non-disruptively update the cluster, some customers may chose to operate in a manner that still mandates fixed patch windows. For this legacy posture, a number of techniques, improvements and can be used to limit the length of time to patch a cluster:

1. Upgrade vSAN - Newer vSAN releases resync data faster and smarter and with more care to not impact production workloads.

2. Consider use of "QuickBoot" for patching ESXi hosts. vSphere Quick Boot is an innovation in conjunction with major server vendors that restarts the VMware ESXi™ hypervisor without rebooting the physical host, skipping time-consuming hardware initialization. For workloads that can accept a interruption consider the "Suspend to Memory" option introduced for vLCM patching with Quickboot in 7 Update 2. Do note, that while this does significantly speed host patching, it will stun virtual machines. For Quickboot capability see KB52477. For more information about vLCM and vSAN see this video.

Where cluster size considerations do come into play is operation and management. In the document, vSAN Cluster Design - Large Clusters Versus Small Clusters, a complete breakdown of considerations and tradeoffs is provided between environments that use fewer vSAN clusters with a larger number of hosts, versus a larger number of vSAN clusters with a fewer number of hosts.

Considerations when performance is important to you?

If performance is top of mind, the primary focus should be more toward the discrete components and the configuration that make up the hosts, and of course the networking equipment that connects the hosts. This would include, but not limited to:

- Higher performing buffering tier devices to absorb large bursts of writes.
- Higher performing capacity tier devices to accommodate long sustained levels of writes.
- Proper HBAs if using storage devices that still need a controller (non-NVMe based devices such as SATA or SAS).
- The use of multiple disk groups in each host to improve parallelization and increase overall buffer capacity per host.
- Appropriate host networking (host NICs) to meet performance goals.
- Appropriate network switchgear to support the demands of the hosts connected, meeting desired performance goals.
- VM configuration tailored toward performance (e.g. multiple VMDKs and virtual SCSI controllers, etc.).

Conclusion

vSAN's approach to data placement means that it does not wide-stripe data across all hosts in a cluster. Hosts in a vSAN cluster that are not holding any contents of a VM in question, will have neither a positive or negative impact on the performance of the VM. Given little to no resource contention, using the same hardware, a cluster consisting of a smaller number of hosts will yield about the same level of performance to VMs compared to a cluster consisting of a larger number of hosts. If you want optimal performance for your VMs focus on the hardware used in the hosts and switching.

Determining if a Workload is Suitable for vSAN

In general, most workloads are suitable for a properly sized vSAN configuration, with few exceptions

Overview

In general, most workloads are suitable for a properly sized vSAN configuration, with few exceptions.

For hybrid configurations, thought should be given as to how applications will interact with cache. Many applications are cache-friendly most of the time, and thus will experience great performance using vSAN.

But not all applications are cache-friendly all of the time. An example could be a full database scan, a large database load, a large content repository, backups and restores, and similar workload profiles.

The performance of these workloads in a hybrid configuration will be highly dependent on the magnetic disk subsystem behind cache: how many disks are available, how fast they are, and how many other application workloads they are supporting.

By comparison, all-flash configurations deliver predictably high levels of performance through low latency all of the time, independently of the workload profile. In these configurations, cache is used to extend the life of flash used as capacity, as well as being a performance enhancer.

The vSAN Ready Node documentation can provide examples of standardized configurations that include the numbers of virtual machines supported and the estimated number of 4K IOPS delivered.

For those that want to go deeper, VMware has a number of tools to assist with determining whether or not vSAN will meet performance requirements, such as the vSAN TCO and Sizing Calculator

For profiling workloads, the Live Optics Virtual Assessment can be deployed to measure workload demand.

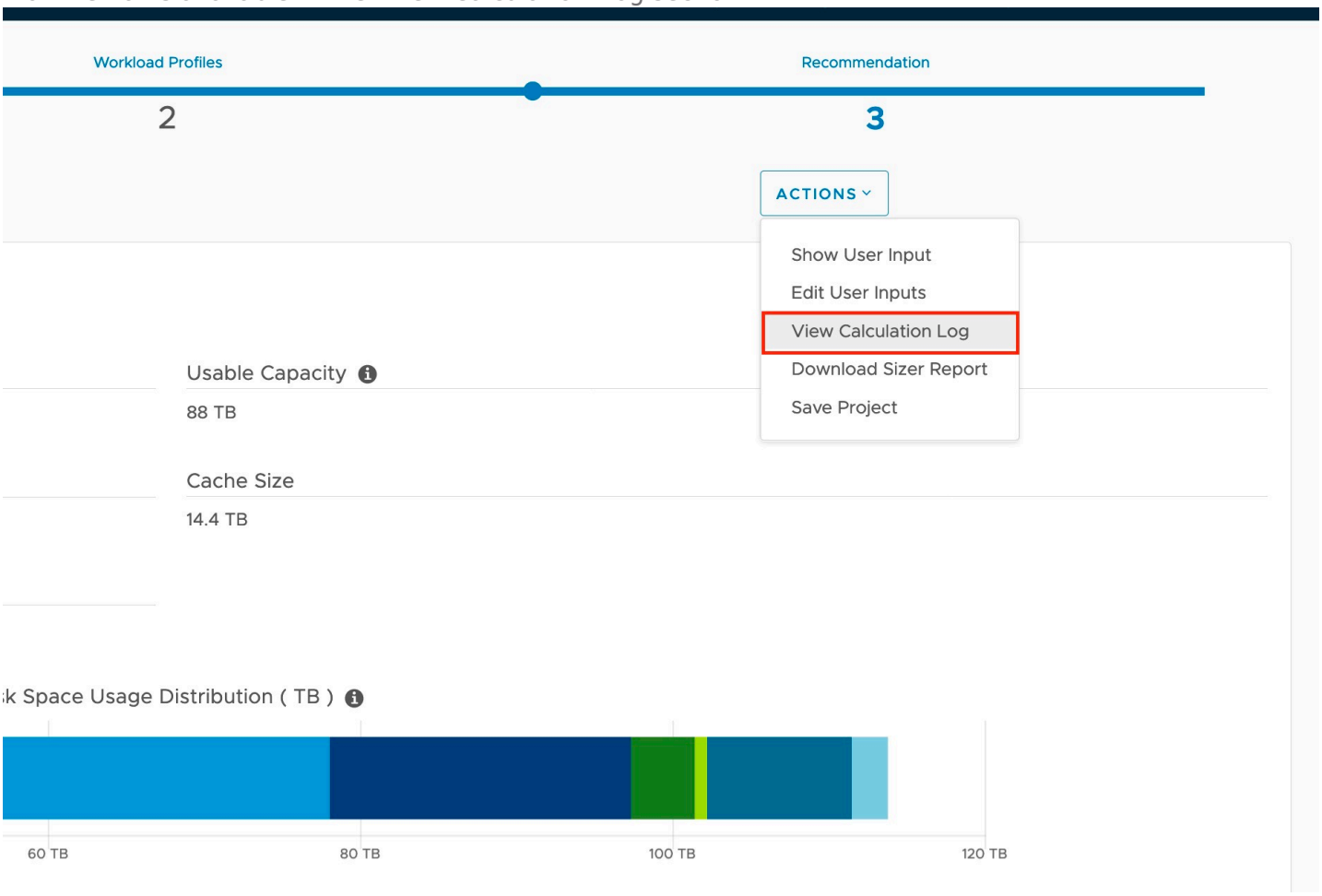
Sizing Examples

Sizing Examples

Previously this section of the design and sizing guide included written out explanations of how to hand calculate the overheads when designing a vSAN cluster. This section has been replaced by the vSAN ReadyNode Sizer that can be found at "https://vsansizer.vmware.com/".

For a video walkthrough of the sizing tools, the following short presentation provides an overview.

If you wish to understand the specific overheads in relation to why it sized a cluster a specific way the math remains available in the "View Calculation" log section.



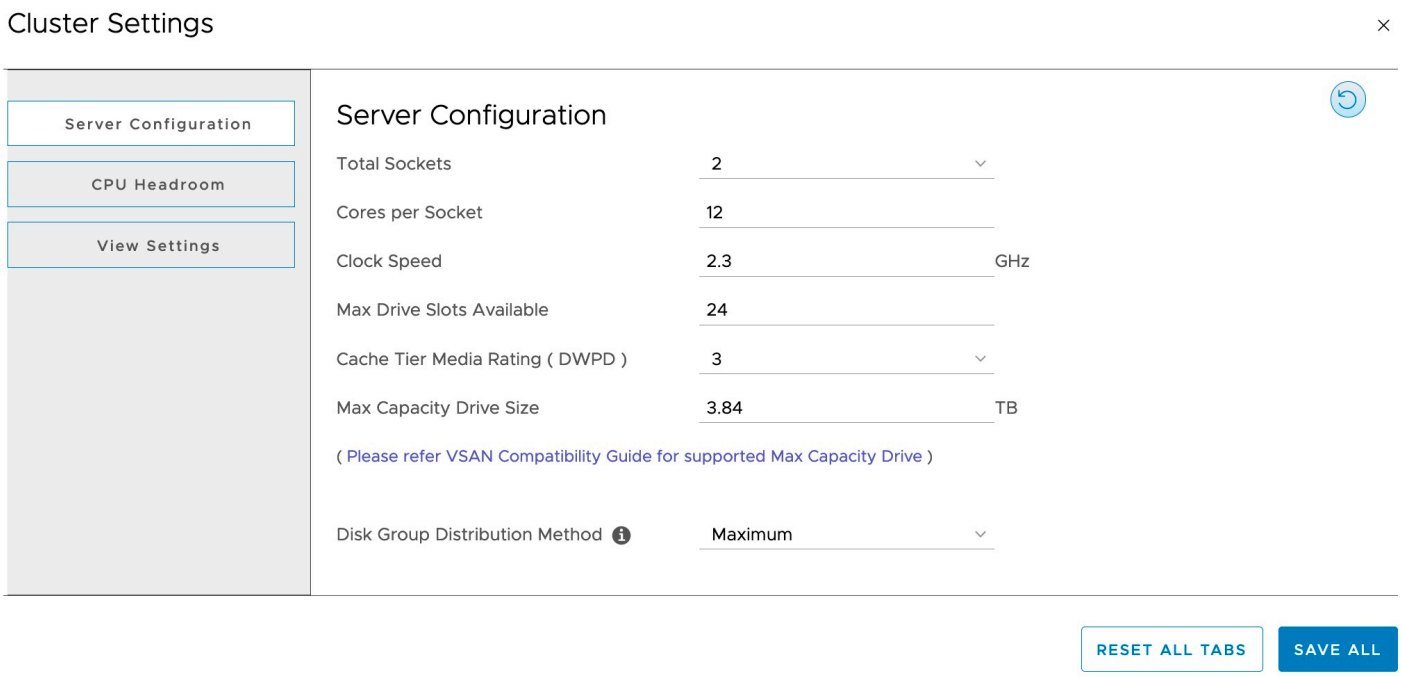
You can review the capacity overheads, as well as calculations for handling host failures, and to see if a workload is CPU or capacity bound.

log Help | CPU Bound | IO Bound | Capacity Bound | HFT | Memory Bound | Disk Layout | Final Recommendation

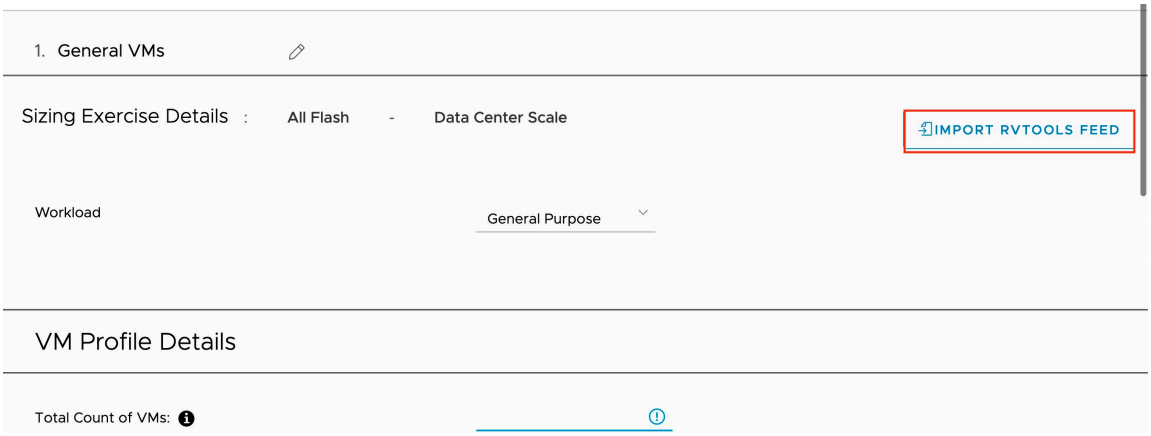
Disk Layout

```
Resource Statistics - CPU % = 221.056 < Total Cores Consumed by vSAN And App > * 100 / (20.4 < Total Cores per Host > * 12.0 < Number of hosts - final number of h
Resource Statistics - IOPS % = 200000.0 < Total IOPs per cluster > * 100 / (25812.0 < Max possible IOPS per Diskgroup > * 2 <No of Diskgroup > * 12.0 < Number of
Resource Statistics - Capacity % = 9483.3< Raw capacity per host with Overheads >* 100 / 15360.0< User Available Space  > = SizerConfigValue [value=61.74, units=%
Resource Statistics - Memory % = 183.9267 < Required Memory (GB) > * 100 / 256.0< Available memory (GB) > = SizerConfigValue [value=71.85, units=%]
=====    Calculate Resource Statistics - End     =====


=====    Calculate DiskSpaceUsage Statistics - Start     =====
cluster Size : 12.0
Disk Format overhead per cluster : 0.065 <Disk format overhead> * 12.0 <cluster size> : 0.78
Physical reservation overhead per cluster : 0.1931 <Physical reservation overhead> * 12.0 <cluster size> : 2.3171999999999997
Remaining free space per cluster : 5.8767 <Remaining free space overhead> * 12.0 <cluster size> : 70.5204
Slack  space overhead per cluster : 1.6093 <Slack space overhead> * 12.0 <cluster size> : 19.3116
Aggregated Raw Storage per cluster : 6502.222222222223 <AggregatedrawStorage > * 12.0 <cluster size> * 0.001 <MemoryGB_TO_TBConversion>  : 78.02666666666667
Aggregated Raw Storage with Overheads per cluster : 9.4833 <Aggregated Raw Storage with Overheads> * 12.0 <cluster size> : 113.7996
Aggregatedrawcapcity per Cluster  : 15.36<Physical space per host>  * 12.0 <cluster size>  : 184.32
Checksum overhead per cluster : 0.7725 <Checksum overhead> * 12.0 <cluster size> : 9.27
Checksum Overhead : 9.27 TB
Dedup overhead per cluster : 0.3412 <dedup overhead> * 12.0 <cluster size> : 4.0944

Dedup overhead : 4.0944 TB
```

There are a number of advanced settings. The cluster Settings allows you to specify the CPU and memory configuration of the servers that will be used.



RVtools a popular third party inventory capture tool that can be found at https://www.robware.net/rvtools/ and can be used to import workloads into the VM profile section.



Common Sizing Mistakes

Although most vSAN design and sizing exercises are straightforward, careful planning at the outset can avoid problems later.

Based on observed experiences to date, the most frequent design issues are:

- Failing to use VCG-listed components, drivers and firmware, resulting in unpredictable behavior. Flash devices and IO controllers are particularly sensitive.
- Not properly sizing cache for capacity growth (e.g. thin volumes progressively getting fatter), resulting in declining performance over time.
- Using 1Gb networking for performance-intensive environments, with predictably poor results.
- Not understanding 3-node limitations associated with maintenance mode and protection after a failure.
- Using very large, slow disks for capacity, resulting in poor performance if an application is not cache-friendly.
- Failure to have sufficient extra capacity for operations like entering maintenance mode, changing policy definitions, etc.

Additional Links

VMware Ready Nodes: https://vsanreadynode.vmware.com/RN/RN
vSAN Configuration Guide: https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan

vSphere Community Page: https://communities.vmware.com/community/vmtn/vsan

Key Blogs:

- http://cormachogan.com/vsan/
- http://www.yellow-bricks.com/virtual-san/
- http://www.virtuallyghetto.com/category/vsan
- http://blogs.vmware.com/vsphere/storage
- http://www.thenicholson.com/vsan
- http://www.vmware.com/products/virtual-san/resources.html - vSAN Product Page
- https://www.vmware.com/support/virtual-san - vSAN Support Center
- vspeakingpodcast.com - Virtually Speaking Podcast
- vSAN documents, demos, podcasts, web based learning etc.

vSAN Sizing Tool

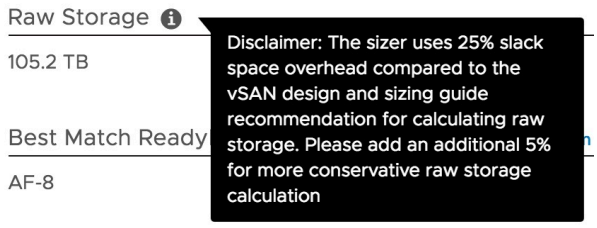
The new vSAN Sizer can be found at: https://vsansizer.vmware.com

This sizing tool is a tool for HCI sizing with vSAN. It is not limited to storage, but also factors in CPU and memory sizing also. This tool incorporates into sizing overheads for Swap, Deduplication and Compression metadata, disk formatting, base 10 to 2 conversion and CPU for VSAN.

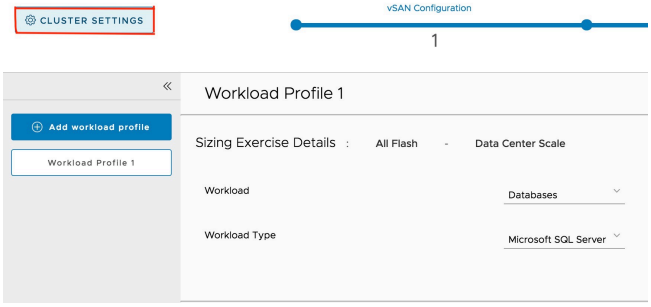
Assumptions

The Tool currently assumes a minimum of 3 nodes and 2 disk groups per server.

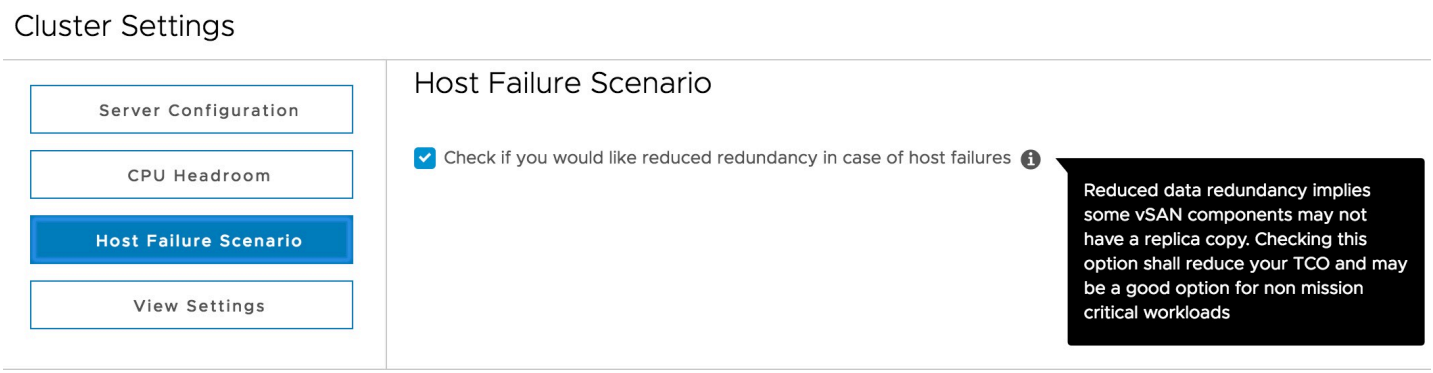
Currently the Slack calculates at a 25%. To follow the 30% recommendation for slack add an additional 5% capacity as well as any additional capacity needed for growth beyond what was put into the workload profiles.



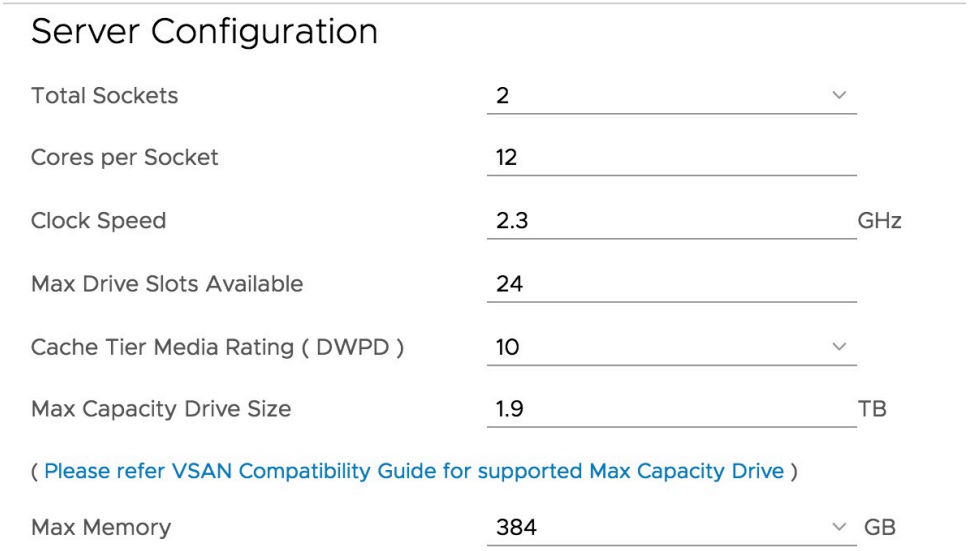
Cluster Settings contains additional assumptions regarding compute sizing, headroom, server configuration, and failure tolerance.



If you want to maintain slack space and have host overhead for a rebuild, unchecking the " *Check if you would like reduced redundancy in case of host failures*" checkbox.

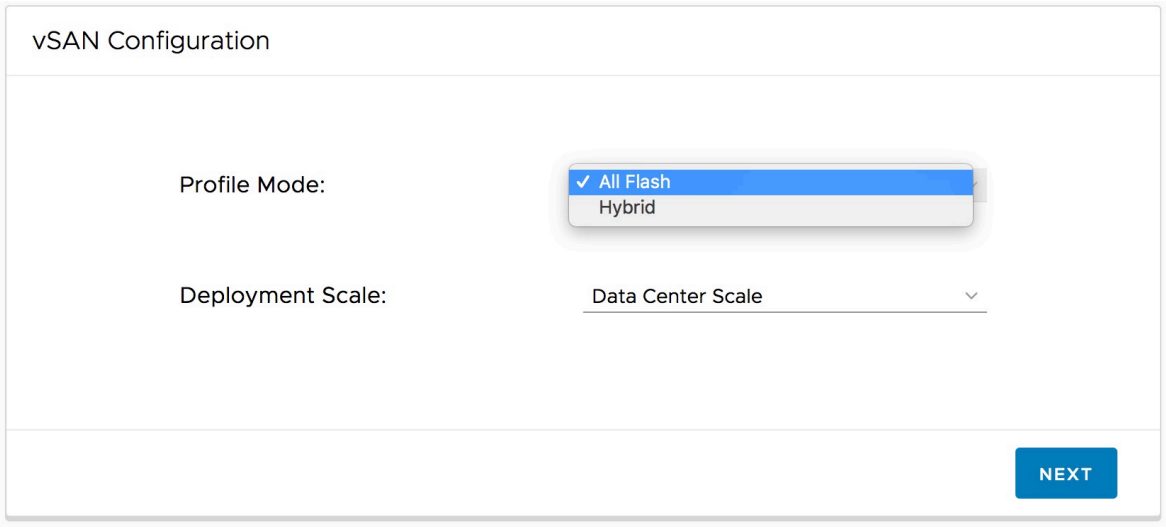


You can adjust the server configuration from the default. This will allow you to change what kind of drives can be chosen, what the number of drive bays are, the type of CPU used, and the maximum supported memory configuration.



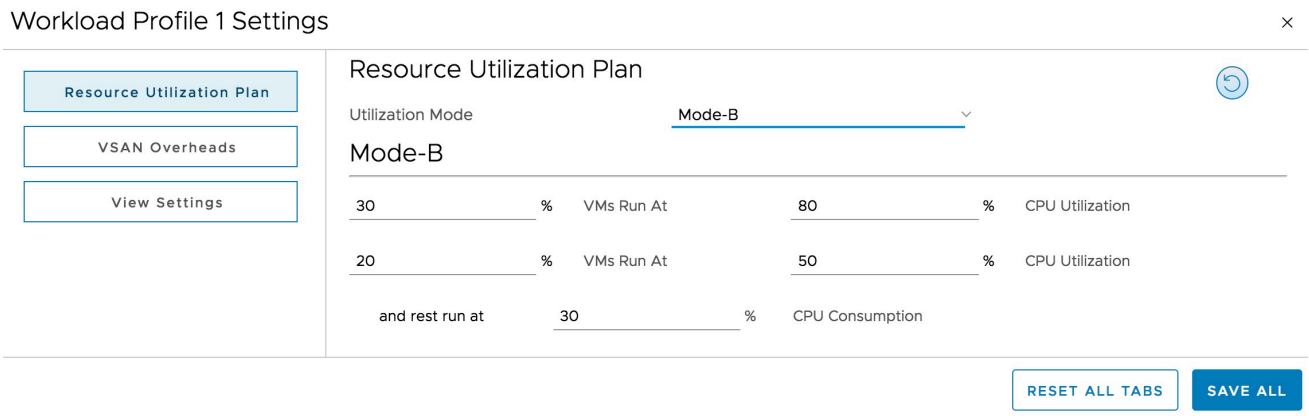
Guidance for the vSAN Sizing Tool

Start by choosing All Flash or Hybrid.

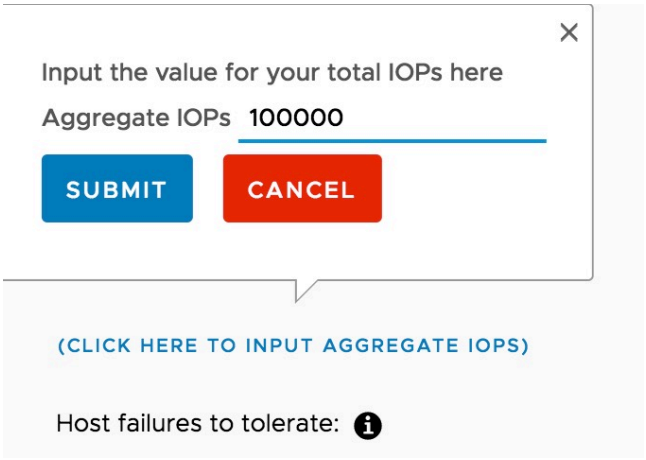


Workload Profiles

It should be noted that you can add multiple workload profiles, as well as edit the base assumptions within a workload profile. *Advanced Settings* within workload sizing allow for customization of the compute demand vs. vCPU allocation for virtual machines. if you have a homogenous workload on CPU demands, or more detailed information on CPU usage you may need to adjust the resource utilization pan for the workload group.



For storage performance demands, you can submit a total IOPS for a workload group of VM's and it will divide that across the pool



If VDI is chosen as a workload profile, note that it defaults to a 200GB Capacity used. Make sure to pay attention to adjust this, as linked clones and instant clones can use significantly less capacity.

VMware vSAN GSS Support

- https://my.vmware.com/web/vmware/login
- http://kb.vmware.com/kb/2006985 - How to file a Support Request
- http://kb.vmware.com/kb/1021806 - Location of VMware Product log files
- http://kb.vmware.com/kb/2032076 - Location of ESXi 5.x log file
- http://kb.vmware.com/kb/2072796 - Collecting vSAN support logs