vSAN Design Guide

Design recommendations for vSAN in VMware Cloud Foundation 9.1

May 5, 2026

Table of Contents

Introduction

VMware vSAN™ is a hyperconverged, software-defined storage system integrated directly into the hypervisor. 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.

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, drives, network configuration, and number of capacity devices, along with detailed configuration questions to help correctly and successfully deploy vSAN.

There are two ways to build a vSAN cluster:

- Certified vSAN ReadyNodes from any of the leading server OEMs.
- Emulated ReadyNodes, or ReadyNodes with changes made following this knowledge base article.

Recommendation: Strongly consider Emulated ReadyNodes for best pricing, but do verify all components are certified and match the Broadcom BCG

New to the Guide is content for vSAN 9.1. Design Guidance for vSAN 8 and original storage architecture can be found in this archived copy of the design guide. VMware strongly recommends new vSAN clusters be deployed using Express Storage Architecture.

Scope of Topics

The information provided in this document will assume the use of vSAN 9.1 and/or VMware Cloud Foundation (VCF) 9.1. VCF deployments may have additional requirements and support limitations that fall outside of the scope of this document.

ReadyNode Profiles

There are three broad categories of ReadyNodes to cover HCI, storage cluster, and disaster/cyber recovery use cases. Performance numbers now include a minimum expected performance for each category. These IOPS per node numbers are assuming a 8KB block size, and a 8K, 70/30 read/write ratio. The latest ReadyNode profile requirements can be found here.

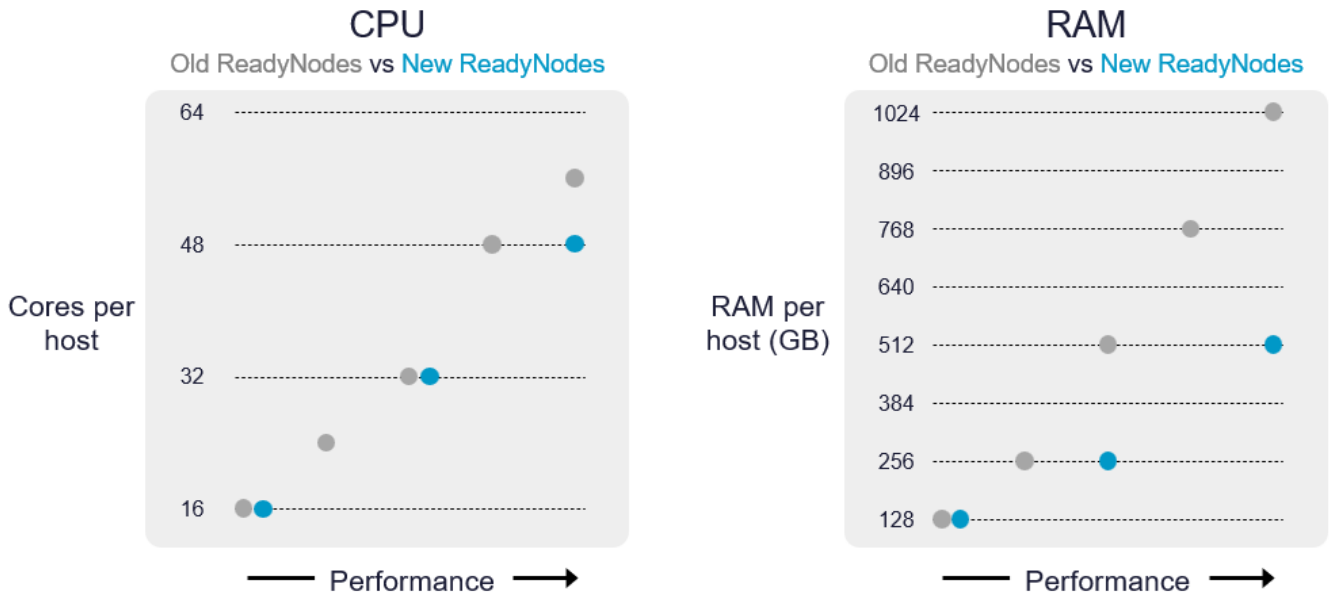
updates, and reduced resource requirements

The previous 4 categories were simplified down into a “Small/Medium/Large” sizing, along with reduction in the CPU and memory requirements. In addition a review of telemetry and phone home was performed to right size nodes based on real world overheads for computers. This led to substantial reductions.

- Up to a 67% decrease in RAM for ReadyNodes certified for vSAN storage clusters.
- Up to 33% decrease in CPU cores for ReadyNodes certified for vSAN storage clusters.
- Up to a 50% decrease in RAM for ReadyNodes certified for vSAN HCI clusters.

For more information about the reduced CPU/Memory requirements, see this blog post.

ReadyNodes Certified for vSAN HCI Clusters



vSAN ESA HCI ReadyNode™

Description	vSAN-HCI-SM	vSAN-HCI-MED	vSAN-HCI-LRG
Node Capacity (TB, min)	3.2	20	50
CPU (cores/node, min)	16	32	48
Memory (GB/node, min)	128	256	512
Max Storage Devices for base configuration	12	18	24

Networking (GbE, min)	10	25	100
Performance for base configuration (IOPS/node, up to)	50 K	100 K	150 K

vSAN ESA Storage Cluster ReadyNode™

Description	vSAN-SC-SM	vSAN-SC-MED	vSAN-SC-LRG
CPU (cores/node, min)	16	32	48
Memory (GB/node, min)	128	192	256
Nodes per Cluster (min)	4	6	6
Max Storage Devices for base configuration	12	18	24
Networking (E/W GbE) (min)	25	25	100

Networking (N/S GbE) (min)	10	25	25
Performance for base configuration (IOPS/node, up to)	50 K	75 K	125 K

Recommendation: Always consider splitting front and back end networking for storage clusters into dedicated network interfaces. North/South connectivity that connects to the compute cluster can be older/slower legacy 10Gbps networking..

vSAN ESA Cyber Recovery ReadyNode™ Hardware Guidance

Description	CyberRecovery-SM	CyberRecovery-MED	CyberRecovery-LRG
CPU (cores/node, min)	16	24	32
Memory (GB/node, min)	128	192	256
Nodes per Cluster (min)	4	6	6
Max Storage Devices for base configuration	8	18	24
Networking (GbE, min)	10	25	25

Replication Throughput (node, up to)	**150 MB/s**	**225 MB/s**	**300 MB/s**

Note: Minimum number of devices across profiles must be 4

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 (Being deprecated summer 2026) or the new VCF Private Cloud Sizer (Currently in Beta to replace it).

Adhere to the Broadcom Compatibility Guide (BCG)

There are a wide range of options for selecting a host model, storage controller as well as types of flash devices.. It is extremely important that you follow the Broadcom Compatibility Guide (BCG) 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. For VSAN ESA you no longer need to consult storage controllers, as vSAN ESA is not supported behind RAID controllers, Tri-mode controllers or HBAs. For a list of vSAN ESA compatible storage devices see this vSAN BCG link.

The ESA is much easier to design for, as it can 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 BCG makes very specific recommendations on hardware models for NVMe storage devices. 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 RAID/HBA Controllers (OSA Only) and NVMe drives the exact version supported is specified. Ensure that the hardware components have these levels of firmware, and that any 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.

In previous versions, vSphere Lifecycle Manager validates device firmware and drivers against the HCL when a third-party Hardware Support Manager (HSM) is present. In 9.1, vSphere Lifecycle Manager reports the current running driver and firmware of a device and validates against the HCL for vSAN clusters. Some devices may not report their firmware without an appropriate HSM. This provides a first-level validation of devices when used in a vSAN cluster

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.

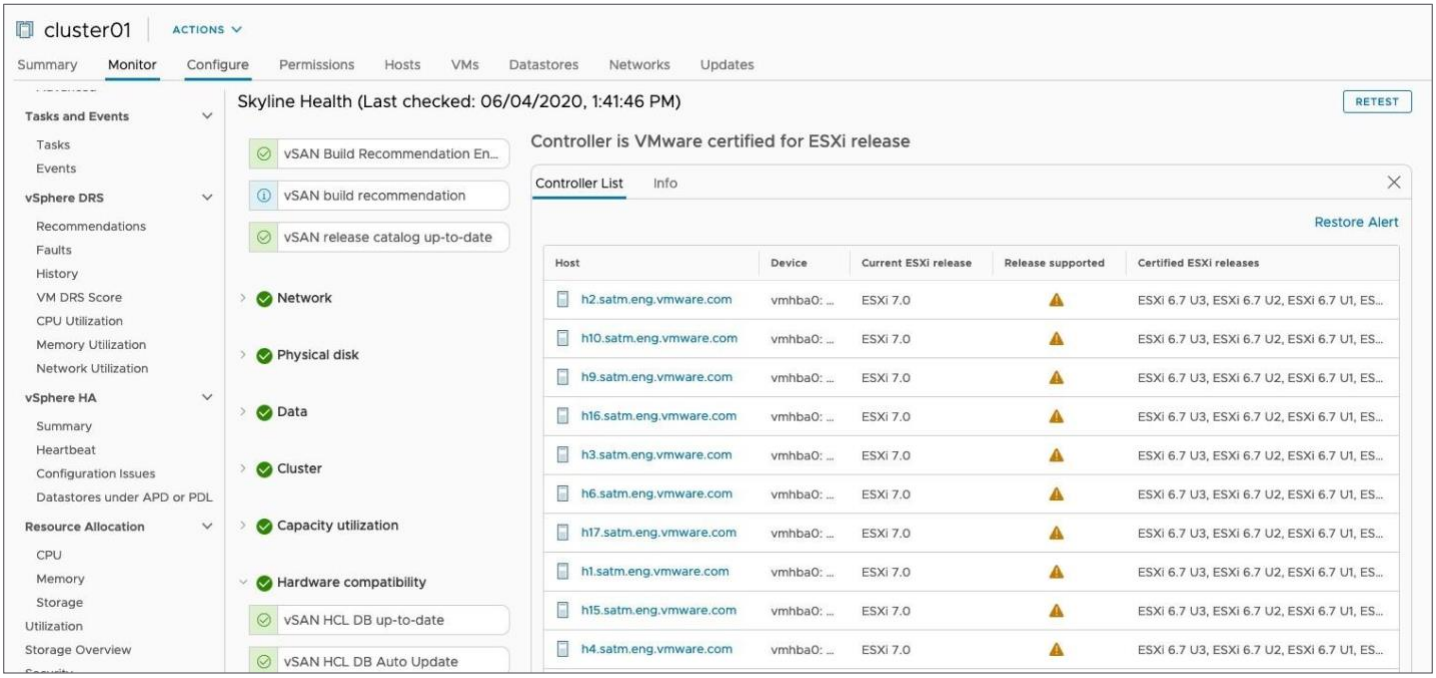
Network Interface Cards (NIC) and RDMA Support

While it is strongly recommended to choose NICs from the vSAN BCG, any vSphere supported NIC is supported for TCP usage. For RDMA usage, only certified NICs with certified firmware and drivers are supported. For supported NICs please see the vSAN networking BCG and confirm that the NIC is explicitly supported for vSAN RDMA usage.

Consider increasing ring buffer allocations to NIC drivers if you are seeing high pNIC error rates.

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

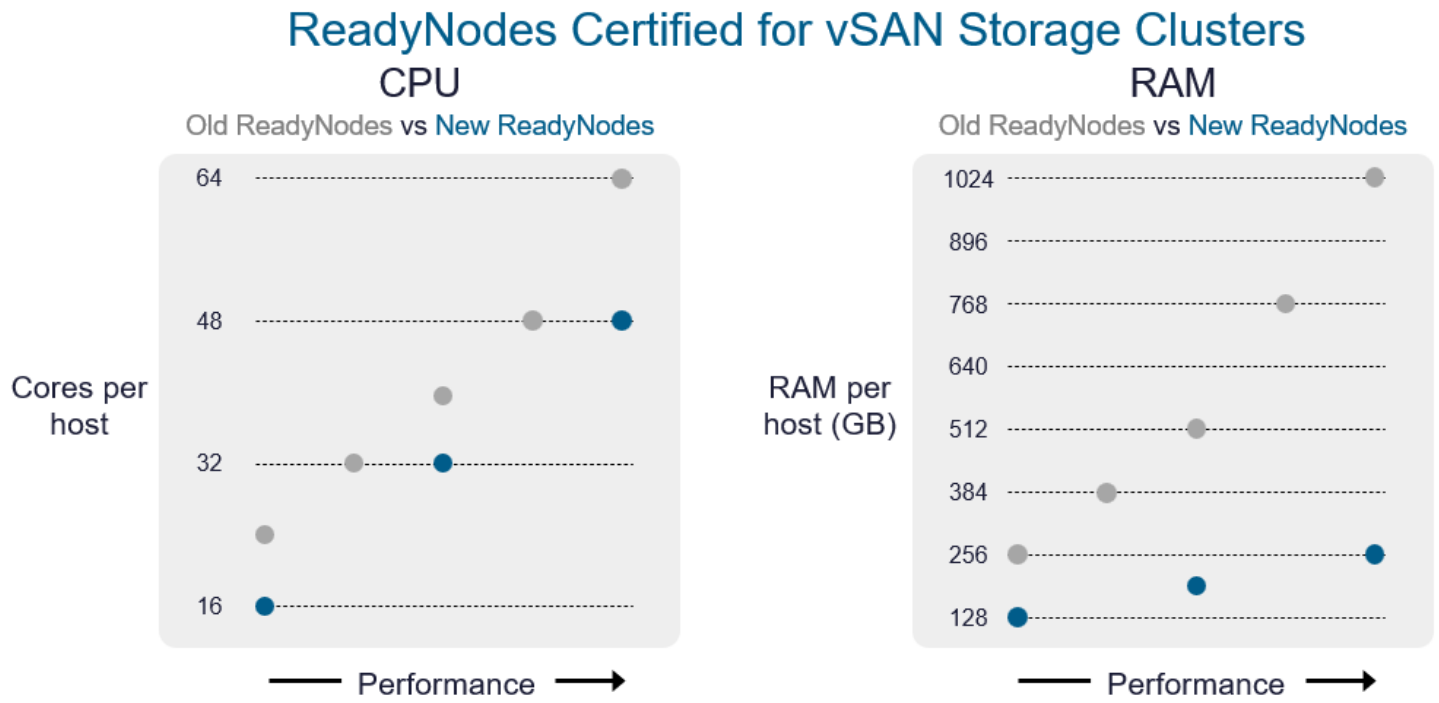
Recommendation: Verify all software, driver and firmware versions are supported by checking the VCG and using the vSAN Health Service. The screen shot below shows an example of a controller driver that is not on the VCG.



For more information on vSAN over RDMA, see the post: "vSAN Networking – Is RDMA Right for You?"

VMware vSAN datastore sharing and vSAN storage clusters (formerly HCI Mesh and vSAN Max)

Datastore sharing 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. Dedicated vSAN storage clusters can be deployed to explicitly provide shared storage to other vSAN and compute clusters. For more information see the updated Design and Operations Guidance for vSAN storage clusters. vSAN Storage cluster storage nodes require 25Gbps for the inter-storage node communication. Older, legacy 10Gbps may be used to connect the compute clusters to the vSAN storage clusters. Separate VMkernel ports can be devoted to back end vs. front end connectivity. Work has recently been done to decrease the compute requirements of storage clusters. More information can be found here.



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 1.92TB SSDs become more difficult to find, adding a comparable in performance 3.84TB 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 NVMe devices that using different Performance Classes, or endurance levels (Read Intensive vs. Mixed Use) within or across hosts that comprise a vSAN cluster will generally only provide performance levels as fast as the slowest devices used. If one is transitioning existing hosts to newer, faster devices such as Gen 5 PCI-Express NVMe drives, 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.

If you are adding drives with different capacity sizes, try to do so evenly across all hosts in the cluster, as to not create a abnormally large or small hosts in the cluster.

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."

Recommendation: 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 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. Do consider that significant capacity and performance demand increases in a vSAN host, may require additional CPU or memory.

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 pools by adding capacity devices, or replacing existing devices if additional drive bays are not available.

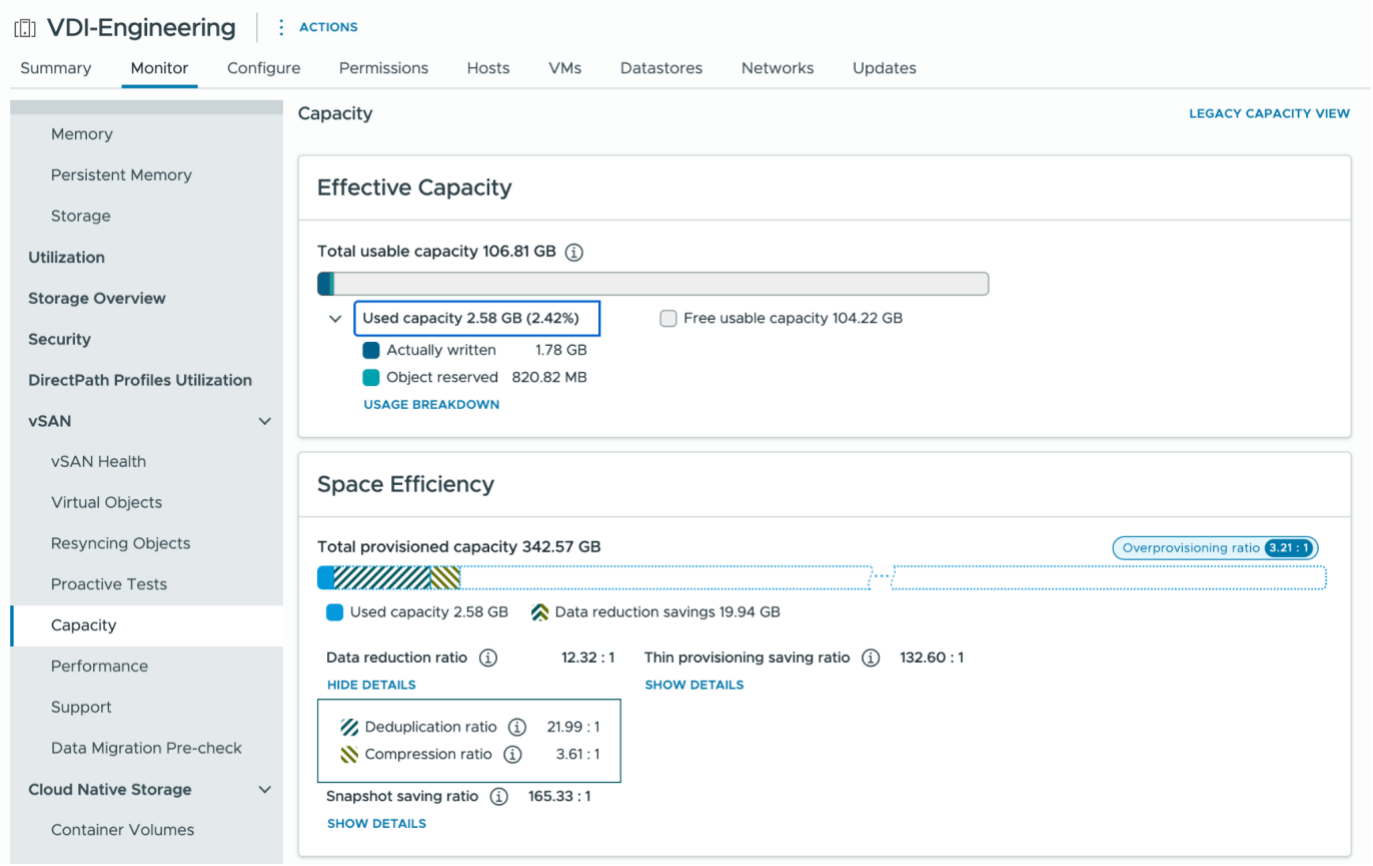
This consideration is covered in depth throughout this guide. In particular, one should consider choosing hosts for a design that have additional disk slots for more capacity, as well as providing an easy way to install additional devices into these slots.

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 document has a detailed list of considerations when reviewing Scale Up vs. Scale Out considerations.

Designing for operational, and host rebuild overheads

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. After that release an optional operational reserve, and host rebuild reserve feature were added as optional toggles that could be used to manage free space as an incremental step in simplifying capacity reservations. New to 9.1 free space should be managed by Auto-RAID. Now capacity views will reflect true usable space accounting for host rebuilds, oeprational needs, and raid overhead. Once enabled the capacity monitoring for vSAN will reflect true usable capacity remaining.



Actual overheads will vary, and are dependent on the configuration of the cluster. These overheads are built into the VCF sizer, and will be accounted for.

For more information and examples on Auto-RAID in vSAN, see the post: "Auto-RAID in vSAN."

Maintenance Considerations

Later versions of vSAN 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 was 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. For more information about vSAN availability technologies see the tech note.

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.

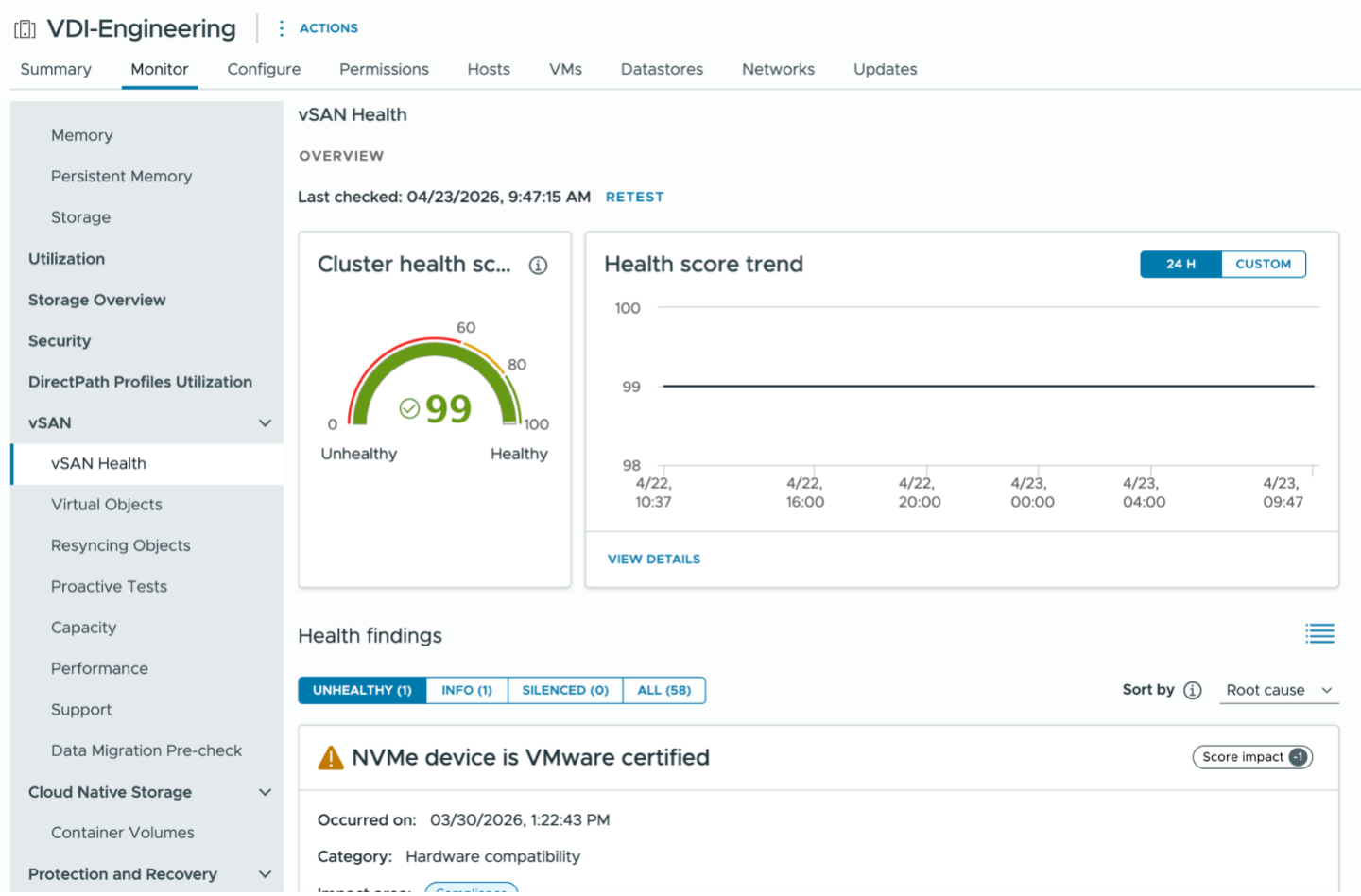
One additional consideration is the size of the capacity layer. Since virtual machines deployed on vSAN are policy-driven, and one of those policy settings (NumberOfFailuresToTolerate) will make a mirror copy of the virtual machine data, one needs to consider how much capacity is required to tolerate 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 Health

vSAN includes the vSAN Health service. This complimentary feature available to vSAN that 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 Broadcom Knowledge Base article for guidance on correcting the issue. Online health checks 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 and Review Interoperability

Verify that the vSphere components in your environment meet the software requirements for using vSAN. To use the full set of vSAN 9.1 and later capabilities, the ESXi hosts and vCenter Server must be on vSphere 9.1. 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.

vSAN shares version numbers with ESXi and vCenter server, and if both are updated will be considered the same version of vSAN. vSAN 9.1 is effectively vSphere 9.1 + vCenter Server 9.1.

If you want to check vCenter server backwards compatibility with a given ESXi and vSAN version, use the Interop matrix tool.
vSAN functionality may require a non-disruptive update to the vSAN file system version. To see which versions of vSAN require which file system versions see this KB.

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 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 connectivity 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 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 cluster expansion. It is recommended starting with vSAN 9 for new clusters to always deploy vSAN Express Storage Architecture.

Hybrid Considerations

vSAN Original Storage Architecture (OSA) - Hybrid Deprecation

The hybrid configuration in vSAN Original Storage Architecture feature will be discontinued in a future VCF release. This was announced with vSAN 9.0. Support for QLC (initially in 9.1 for CyberVault use cases), global deduplication, improved compression, raid 5-6 have effectively closed the gap making magnetic storage unattractive on a cost basis.

Summary of Design Overview Considerations

- For new clusters, vSAN ESA should be preferred.
- When using the Express Storage Architecture in vSAN 9 , please make sure proper design principles 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
- Where possible, 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 activates the cluster to automatically remediate in the event of a failure, or when using 2 Node clusters using protection levels inside of the host.
- 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. There are vSAN configuration limits that impact design and sizing. Refer to "Configuration Maximums" in the Broadcom Config Max tools. When deploying high VM counts, be mindful of component per disk limitation (discussed in vSAN limitations section of this document).

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..

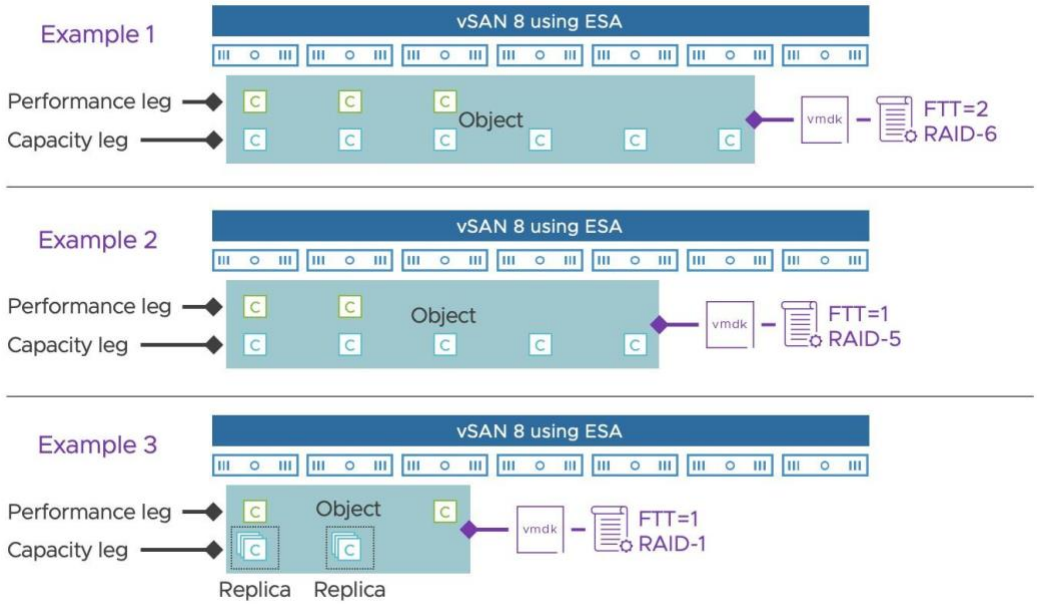
Design decision: Ensure there are enough hosts (and fault domains) in the cluster to accommodate a desired NumberOfFailuresToTolerate requirement.

Maximum VMDK Size and Component Counts

The maximum VMDK size on a vSAN datastore is 62TB. This is the same maximum as NFS and VMFS.

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. Historically If the policy contains a stripe width, the object will be striped across multiple devices in the capacity layer, but this feature no longer has significant impact with vSAN ESA. 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 image below.

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. In addition, the amount of objects per host is tied to the number of physical disks. Each ESA disk is limited to 3000 DATA components and 3000 metadata components, so 4 disks would provide for 24,000 (12,000 data components, and 12,000 metadata components). There is a health check for object counts.



Within an object one will see a concatenation of two data structures.

- **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

The following highlights key take aways from this section.

- 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 accommodate performance and component counts.
- 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.

Network Connectivity and Bandwidth

vSAN Express Storage Architecture (ESA) supports 10Gbps for vSAN-HCI-SM (Formerly ESA-0) profiles and vSAN-HCI-MED and above profiles will 25Gbps or higher speed requirements. ESA ReadyNodes configured for vSAN ESA will be configured with 25100/200Gbps NICs. The 10Gbps requirement for the vSAN-HCI-SM profiles should be considered for brownfield deployments, but it is strongly advised to still purchase 100Gbps NICs for future proofing as the QSFP interface can be broken out to 10 or 25Gbps if needed, while future proofing the hosts. Inversely 100Gbps switches should be strongly considered even if hosts only have 25Gbps interfaces, as breakout cables can be used to provide SFP28 to the hosts from the 100Gbps switch ports. vSAN ESA prior to vSAN 9 may require custom configurations to use 10Gbps connectivity. See this knowledge base article.

vSAN storage clusters can split networking to the remote compute clusters, and the back end storage node to storage node traffic. See vSAN storage cluster 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 switches. 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. 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 must 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"
- Read "vSAN Networking – Network Topologies"
- Read "vSAN Networking – Network Oversubscription"
- Read "vSAN Networking – Optimal Placement of Hosts in Racks"
- Read "vSAN Networking – Teaming for Performance"
- Read "vSAN Networking – Teaming for Redundancy"
- Read "vSAN Networking – Is RDMA Right for You?"

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 standard switch 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.

In certain configurations, multiple VMkernel ports can be used for vSAN storage clusters. Starting in vSAN 9.0 you may split back-end vSAN traffic from front-end storage traffic requested by the vSphere hosts mounting the datastore.

For more information, see the post: "vSAN Networking – Teaming for Redundancy."

MTU and Jumbo Frames Considerations

vSAN supports jumbo frames. VMware testing finds that using jumbo frames can reduce CPU utilization and improve throughput. Do be aware, 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.

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

Multicast Considerations

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. The tcptransmission dashboard may be used to validate NIOC is reducing out of order packets.

NIOC only impacts outbound (Transmission, or TX performance). For inbound performance congestion, consider switch specific performance policies, enterprise congestion notification (ECN), Data Center TCP (DCTCP), and simply larger interface speeds.

Recommendation: Use NIOC, but do not use this as a replacement for faster networking links.

Summary of Network Design Considerations

The following highlights key take aways from this section.

- 25Gbps or faster (50/100Gbps) are strongly sugested for vSAN Express Storage Architecture. 10 Gbps is supported for vSAN-HCI-SM ReadyNode profiles to support legacy/brownfield environments, but 25Gbps NICs should be used for these configurations. For edge 2 node configurations, direct connection of 25/100Gbps is cost effective and easy.
- 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 supported for availability/redundancy, but is not required.
- Jumbo frames can provide benefits in a vSAN environment.
- Always use a 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
- Configure DCTP, ECN, DCBX and other network fabric prioritization to reduce network congestion.
- Consider ultra deep buffer leaf switches, for larger clusters, or clusters with bursty workloads. Consider larger interface speeds for longer running high throughput workloads.
- Avoid oversubscription of leaf to spine switches, if clusters will span multiple leaf switches. CLOS networking avoids congestion, especially with spines with shallow buffers.

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

This guide has been updated to only provide guidance for VSAN Express Storage Architecture. For more information comparing vSAN Express Storage Architecture (ESA) to OSA please see this blog.

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 storage profiles.Extreme write performance, and sustained writes may still necessitate using mixed use, but most use cases can use read intensive TLC NVMe drives.

QLC drive support (Cyber Recovery)

As part of 9.1 a new class of CyberRecovery ReadyNodes(™) featuring support for QLC flash has been added. QLC flash has unique capabilities, in that it offers ultra dense, lower cost NVMe devices. This comes at the expense typically of sustained write performance and endurance. These drives are being supported for this use case where Cyber Recovery use cases mandate longer data retention, useful for being able to recover before workloads were compromised. To use these drives,

Scale up Capacity and memory, Ensure Adequate Networking

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

As you scale up storage and memory within a host you should strongly consider the bandwidth requirements required to support maintenance activities, host rebuilds, and performance access demands. Beyond the vSAN considerations of “bigger” hosts, large quantities of memory in a host will require more networking bandwidth for vMotion activities. As we scale hosts to 1TB and more of memory, we should strongly consider 100Gbps and faster networking.

Design decision: Consider how long you want a rolling update of the cluster to take and external requirements (change management windows, estimated time for vMotion to evacuate a host, urgency of DRS rebalancing, vSAN repair from drive failure) when sizing the network of a host.

NVMe flash device form factor considerations

vSAN Express Storage Architecture explicitly requires ReadyNodes configured with ESA compatible NVMe devices. Devices can be added by a number of form factors ranging from older U2 and U3 to newer EDSFF E3 family of form factors. E3 form factor is offering as many as 50-100% more drive slots per rack unit. While previously storage dense configurations would require consideration of 2 Rack Unit (RU) servers to reach 24 drives bays, it is now possible to get as many as 16 E3 drives in 1RU server form factors. The newer E3 is often also required to fully leverage Generation 5 PCI Express storage performance capabilities. While U2/U3 drives will be around for some time, do consider newer designs newer form factors.

vSAN NVMe Device Considerations

vSAN ESA requires NVMe devices. are a number of considerations when deciding what NVMe devices to consider. The considerations fall into three categories; cost, performance & capacity.

- NVMe offers low latency, higher performance, and lower CPU overhead for IO operations.
- SAS and SATA devices are not supported with vSAN ESA.

- NVMe namespaces are not supported at this time.
- Consider at least 4 drives, when high numbers of components per host will be necessary. (See below for more information)

Selecting a NVMe drive quantity

While vSAN will support as few as one storage device per host in an HCI deployment, and two devices in a storage cluster deployment, you may find specifying four (PCIe Gen 5) to six (PCIe Gen 4) NVMe storage devices a good starting point for many ReadyNode configurations. This helps ensure that vSAN has an adequate number of storage devices to meet your performance objectives. 4 hosts, also gets close to the maximum per host component count. 4 drives provides 24,000 of 27,000 as there are 6,000 per host added..

Design consideration: vSAN ESA has simplified device support selection by providing a curated list of NVMe devices. Review the performance and endurance capabilities of devices chosen.

How Much Capacity do I Need?

When determining the amount of capacity required for a vSAN design, historically the 'NumberOfFailuresToTolerate' policy setting previously played an important role in this consideration. RAID 5 and 6 and RAID 1 formerly would be selectively used in different ways with this setting (allowing for multiple RAID 1 mirrors, or the selection between RAID 5 and RAID 6).

Going forward,vSAN Auto-RAID will automatically adjust from RAID 5 to RAID 6 (and apply by default site mirroring for stretched clusters, or host mirroring for 2-node clusters. The overheads, for Auto-RAID when using RAID 5, and RAID 6 will be the same (150% of consumed capacity). This simplifies capacity measurement/projection as the only distinction between RAID 5 and RAID 6 is RAID 6 will be used once you have 6 hosts.

Design decision: Auto-RAID will simplify capacity overhead, as the RAID 5 and RAID 6 have the same capacity overhead consideration.

N+1 or N+2 Design Considerations

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: 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 Sizer will take into account file system, metadata and checksum overheads.

Snapshot Capacity Sizing Considerations

vSAN ESA offers the ability to retain snapshots for data recovery, and data copy management reasons. You may also replicate and orchestrate disaster recovery of these snapshots remotely using VMware Live Recovery (VLR).

Design consideration: If doing longer retention be aware that change rates for virtual machines will impact capacity requirements.

Choosing a Storage I/O Controller

vSAN does not support NVMe drives being placed behind RAID controllers or Tri-Mode controllers. Beyond being unsupported, this has been shown to cause performance problems, especially in configurations that limit NVMe drives to a single PCI-Express lane. While some ReadyNodes have used PCI-Express switches or expansion devices, these are commonly not needed on the most modern of CPU architectures today.

NVMe Hotplug Support

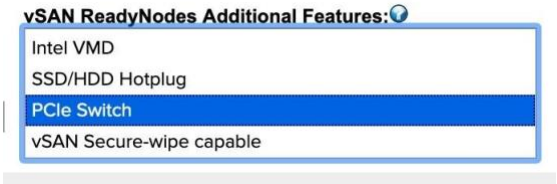
Please consult with the ReadyNode vendor if it will support or not support NVMe hotplug.

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 for OSA usage, 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. Additionally be aware of backplanes used for Tri-Mode controllers often limit NVMe drives to a single PCI-E lane making this unsupported configuration have extreme performance issues.

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.



Storage Controller Queue Depth

Storage Controllers are not in the I/O path of vSAN ESA, and so queue depth is no longer a concern.

Disk drives as a storage failure domain

vSAN OSA required disk groups that would sometimes cause multiple drives to be inaccessible off of a single disk failure. Those limitations do not exist in vSAN ESA, and a single drive failing will not impact other drives within the disk pool, or host.

Small Disk Drive Capacity Considerations

The smallest drive supported for production use for vSAN ESA is 1.6TB.

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. The Maximum vSAN ESA capacity component size is 765GB and the maximum performance/metadata leg component size is 255GB.

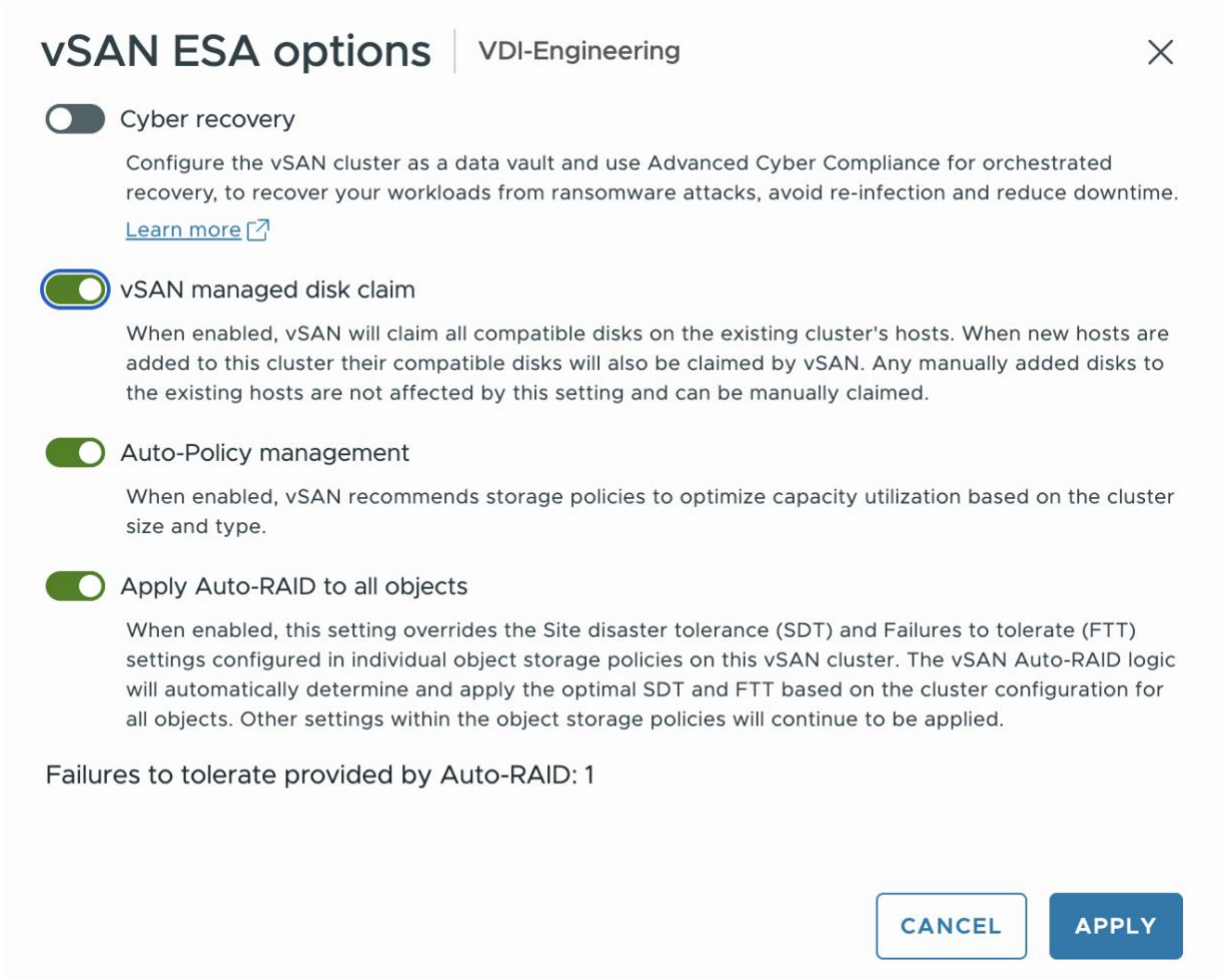
When creating very large VMDKs, the object will be split (striped) into multiple components. This may quickly consume the component count of the hosts, and this is especially true when NumberOfFailuresToTolerate is considered

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 RAID 1, 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 host. 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.

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.

Summary of Storage Design Considerations

The following highlights key take aways from this section.

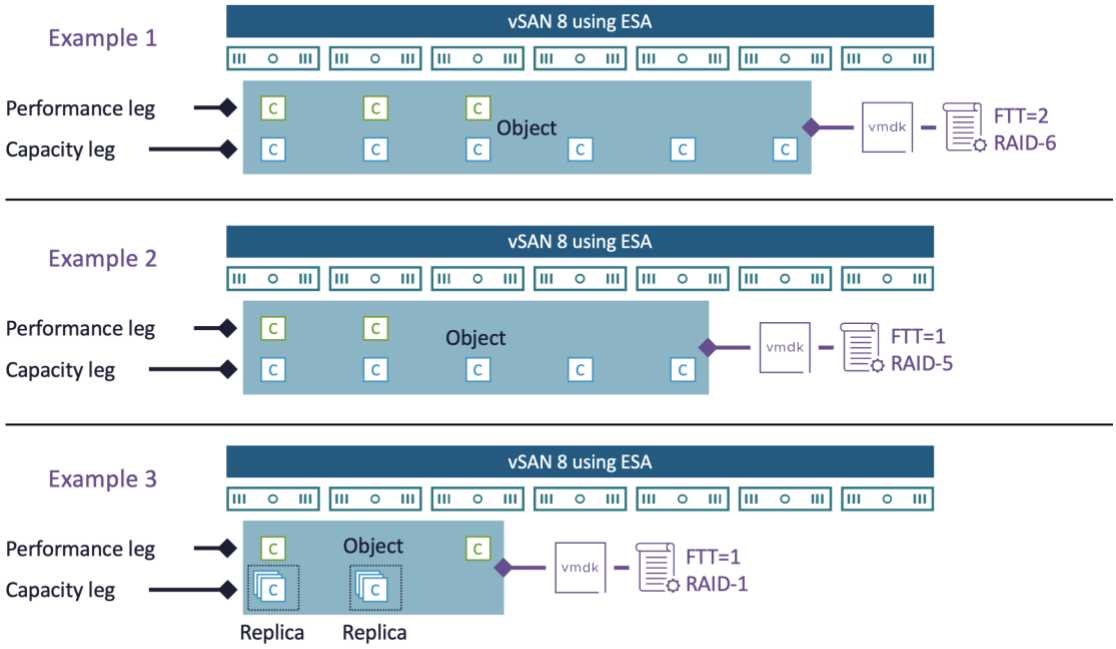
- Determine the endurance required for the flash cache, and the flash capacity requirement for all-flash solution designs.
- Consider hosts with E3.S form factor drive slots. They support Gen5 PCI-E, denser configurations, and are more future proof.
- Design with one additional host with enough capacity 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 use the sizer to account for file system and other metadata overhead.
- Consider object imitations when deploying over 200 virtual machines. You may need more physical devices to increase component count. .
- Consider a design that will facilitate easy replacement of failed components.
- Consider capacity requirements for snapshot or replicated snapshots as part of vSAN data protection and VMware Live Recovery workflows.
- Although not required, it is best to create vSAN disk pools using the same drive configuration (number of drives, endurance and performance requirements) 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 consists 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 composed 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 composed of at least one component. If Auto-RAID is used, then RAID 5 will be used up to 5 hosts, and RAID 6 will be used at 6 hosts for the capacity leg. The performance leg will always use mirroring (with a triple mirror for RAID 6).

NumberOfDiskStripesPerObject is a legacy setting that can be safely ignored in vSAN ESA. It will be removed from the storage policy option, if Auto-RAID is configured for the policy, and will be ignored on legacy policies if Auto-RAID is enabled for the cluster..

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

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. In vSAN ESA components are distributed in such a way that vSAN can still guarantee failures to tolerate without the use of witnesses. The exceptions to this are for 2 Node or stretched cluster configurations, where the witness appliance will act as a witness for quorum proposes and hold a witness component. For vSAN OSA RAID 1 a witness was still 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.

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

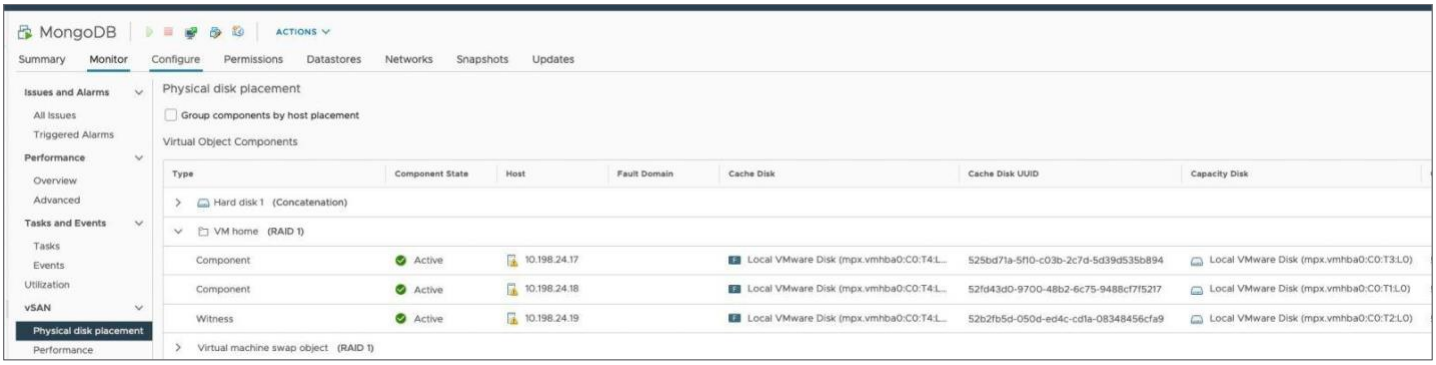
vSAN OSA uses its own SparseSE snapshotting system. This requires IO to merge on snapshot deletion and while improved there can be performance impacts from long-term running with many snapshots.

vSAN ESA uses a native file system snapshot system, that removes performance impacts, stuns and instantly deletes snapshots. For more information see this blog or this video demo.

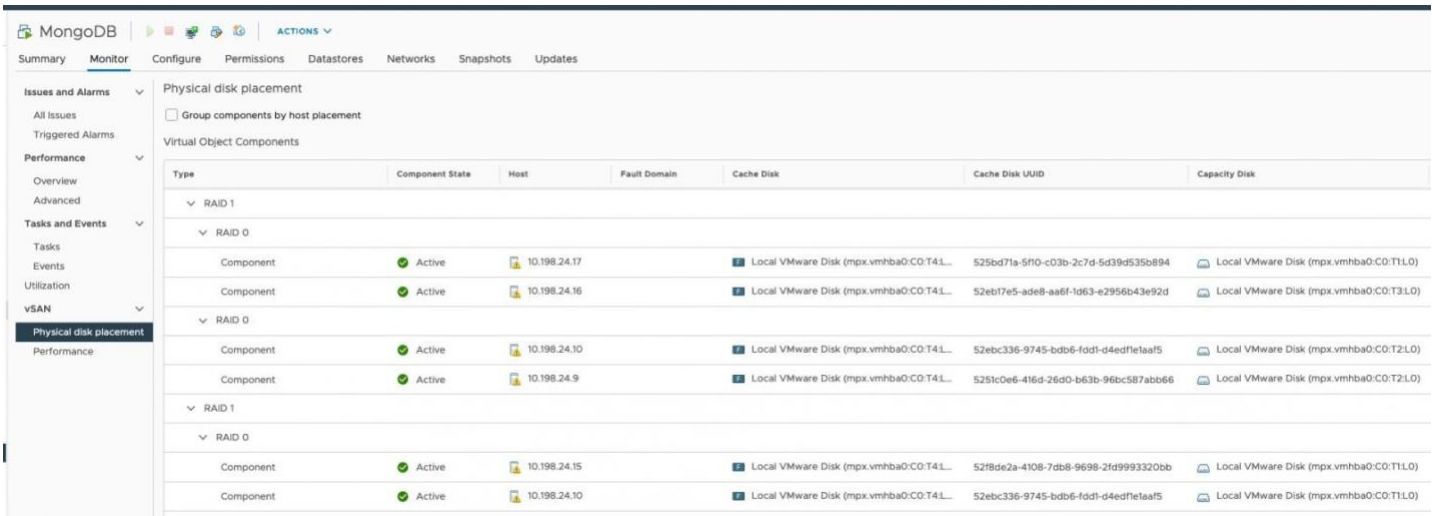
Design consideration: The virtual machine memory snapshot size needs to be considered when sizing the vSAN datastore, if there is a desire to use virtual machine snapshots and capture the virtual machine's memory in the snapshot.

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 setting 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. New in 9.1 Auto-RAID drastically simplifies policy decisions.

vSAN ESA 9.1 - Auto-RAID

vSAN 8 U1 introduced Auto-Policy a way in the ESA to try to generate and maintain for each cluster an optimized storage policy. This was an incremental step, that 9.1 Auto-RAID is the next generation in policy, capacity and availability simplicity. At setup, or after setup a simple slider

takes over configuration of Site disaster tolerance (SDT) and Failures to tolerate (FTT) settings In addition policies that are no longer relevant will no longer be exposed in the UI.

For Auto-RAID enabled clusters, other storage policy rules that are no longer relevant to vSAN ESA will not be displayed within the policy, as they are no longer applicable. This includes:

- Force provisioning (Now automatically handled)
- Number of disk stripes per object (Irrelevant)
- Flash read cache reservation (Irrelevant)
- Disable checksum Iirrelevant)
- Compression (Now an always-on cluster service in vSAN for VCF 9.1)

Resilience Settings of Auto-RAID

The logic Auto-RAID uses for resilience settings is noticeably different from past versions of vSAN ESA. When resilience is possible, it will always default to space-efficient erasure coding for everything except site resilience for stretched clusters, and host resilience for 2-Node topologies. In those cases, the Site Disaster Tolerance will be set to a mirror.

Standard single site clusters

- 6 or more hosts in a cluster. Auto-RAID will use FTT=2 using RAID-6. 1.5x object capacity overhead
- 3-5 hosts in a cluster. Auto-RAID will use FTT=1 using RAID-5. 1.5x object capacity overhead
- Fewer than 3 hosts in a cluster. Auto-RAID will use FTT=0. 1.0x object capacity overhead.

Stretched clusters

- 6 or more hosts per site/fault domain. Auto-RAID will use a site disaster tolerance of a RAID-1 mirror, plus FTT=2 using RAID-6. 3.0x object capacity overhead
- 3-5 hosts per site/fault domain. Auto-RAID will use a site disaster tolerance of a RAID-1 mirror, plus FTT=1 using RAID-5. 3.0x object capacity overhead
- Fewer than 3 hosts per site/fault domain. Auto-RAID will use a site disaster tolerance of a RAID-1 mirror, plus FTT=0. 2.0x object capacity overhead

2-Node clusters

- 6 or more storage devices per host. Auto-RAID will use a site disaster tolerance of a RAID-1 mirror, plus FTT=0. 2.0x object capacity overhead. (secondary levels of resilience are not currently available for 2-Node clusters using Auto-RAID in 9.1)
- 3-5 storage devices per host. Auto-RAID will use a site disaster tolerance of a RAID-1 mirror, plus FTT=0. 2.0x object capacity overhead. (secondary levels of resilience are not currently available for 2-Node clusters using Auto-RAID in 9.1)
- Fewer than 3 devices per host. Auto-RAID will use a site disaster tolerance of a RAID-1 mirror, plus FTT=0. 2.0x object capacity overhead.

This means that the standard overhead for standard clusters will be 1.5x, stretched clusters will be 3x, and 2-Node clusters will be 2x. This capacity overhead is prior to savings from compression and deduplication. This setting is required to be used, to use the new in 9.1 vSAN capacity views that show "usable" capacity remaining on the cluster.

One noteworthy item is that when Auto-RAID assigns FTT=1 using RAID-5, it will always use the 2+1 scheme. The optional 4+1 RAID-5 erasure code in previous versions of vSAN ESA is not used.

For more information about this policy management, see this blog.

Number of Disk Stripes Per Object/Stripe Width

Note: Use of vSAN Auto-RAID on ESA policies, removes this option from appearing.

Given that the new data structure and Log Structured Object Manager allow vSAN to deliver near device-level performance of NVMe devices, increasing the stripe width value does little more than create more data components, and complicate placement decisions for vSAN. Increasing the stripe width is not recommended as a mitigation step when troubleshooting vSAN performance in the ESA.

Why is it still there? If it is no longer relevant for the ESA, then why does the policy rule still exist? Storage policies and the rules that make up a storage policy are a construct of the vCenter Server. A given vCenter Server may be responsible for many vSAN clusters, some of which may be running the OSA, while others use the ESA.

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

Force Provisioning (legacy: non-Auto-RAID clusters).

Note: Use of vSAN Auto-RAID on ESA policies, removes this option from appearing.

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.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.

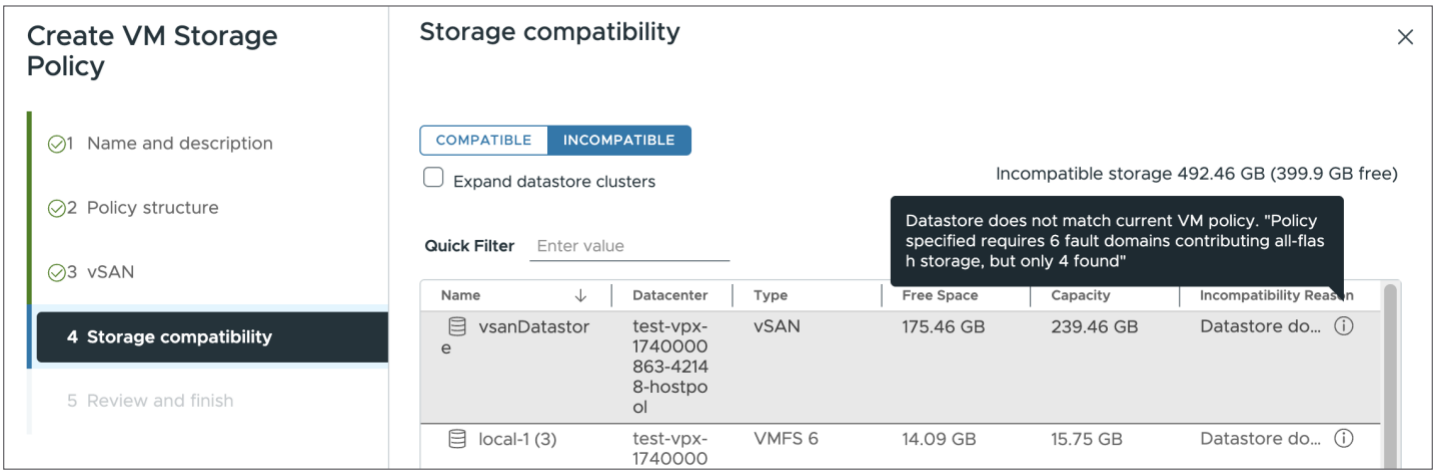
Object Space Reservation (thin provisioning)

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.

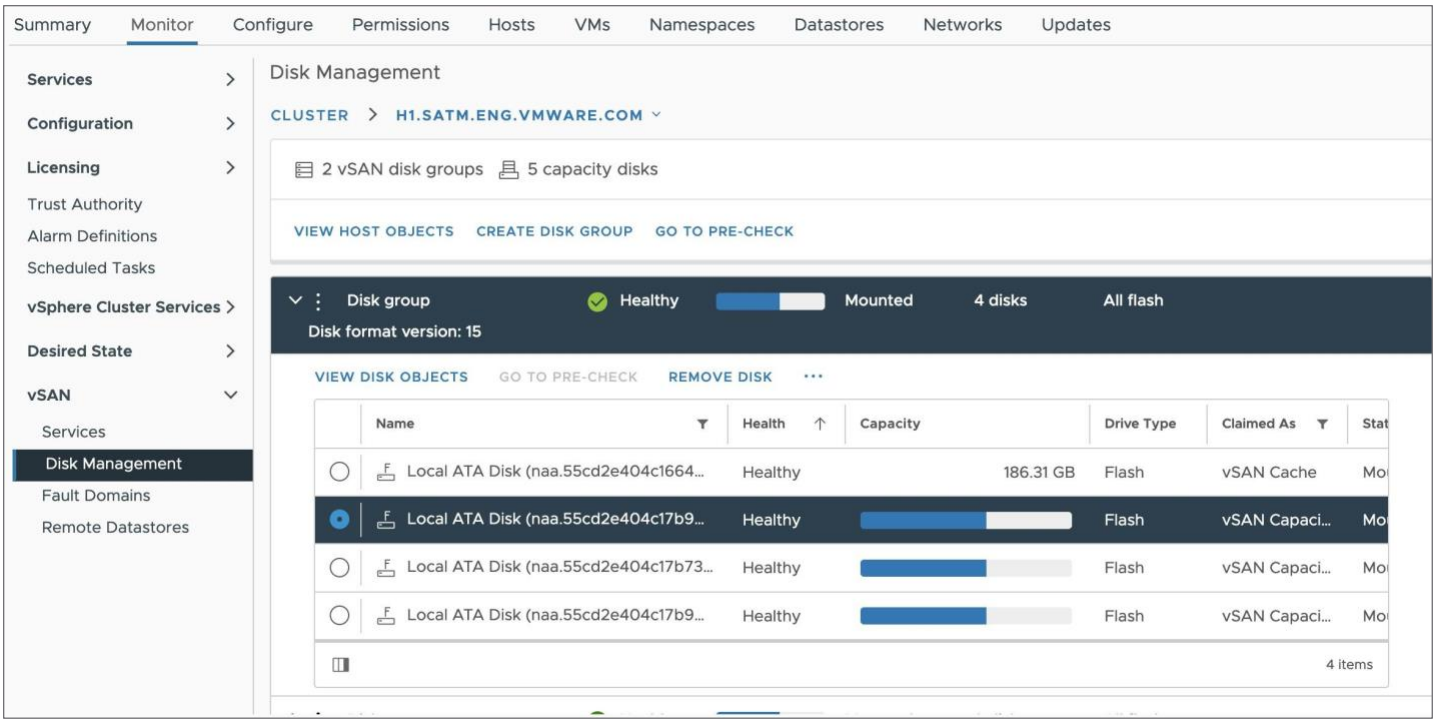
By default, virtual machine storage objects deployed on vSAN are thinly provisioned. This capability, ObjectSpaceReservation (OSR), specifies the percentage of the logical size of the storage object that should be reserved (thick provisioned) when the virtual machine is being provisioned. The rest of 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 hosts in the cluster to satisfy a raid 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 vSAN policies 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.

IOPs 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.

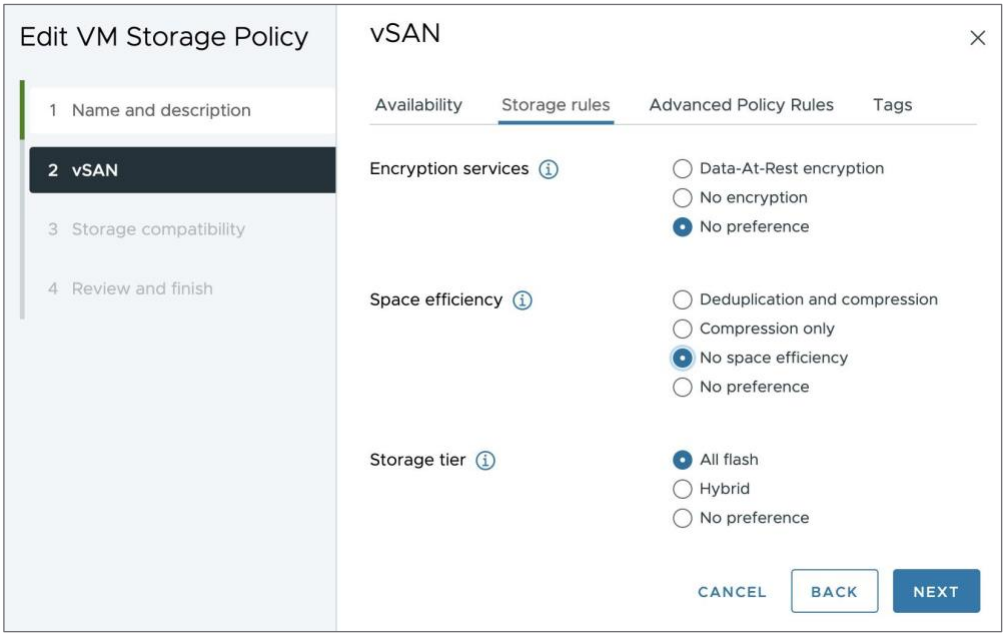
This capability IopLimitForObject limits the amount of performance available to an object. This is normalized to a 32KB block size. A virtual machine reading or writing at 16KB would be treated the same as one performing 32 KB-sized operations. A 64KB read or write however would be treated as two separate operations, leading to half of the configured IOP limit being the number of operations performed.

Deactivate Object Checksum

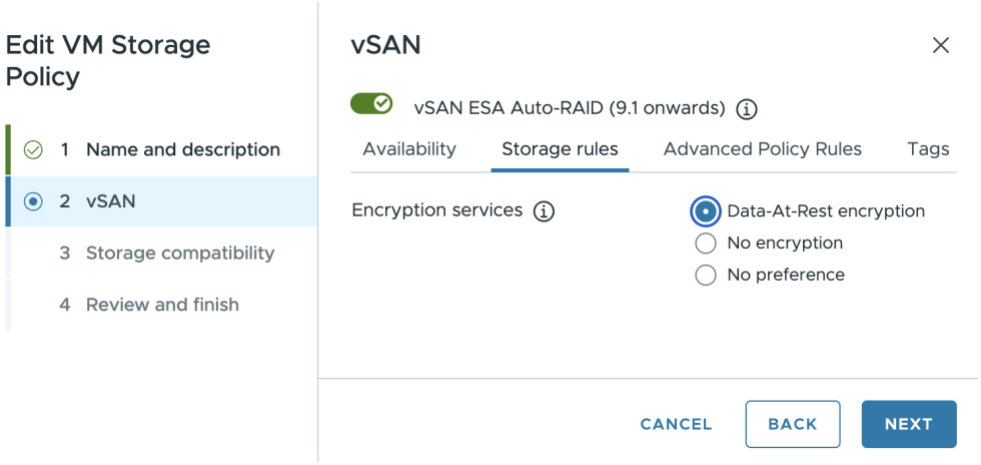
This is a legacy policy that has no impact on vSAN ESA. It will be ignored by the vSAN ESA cluster, and exists for migration compatibility.

vSAN ESA Compression (vSAN 8- 9.0 only)

VSAN 8 Express Storage Architecture by default enables compression only on all virtual machines created. **Starting in vSAN 9.1 compression cannot be disabled, and this policy will be ignored.**



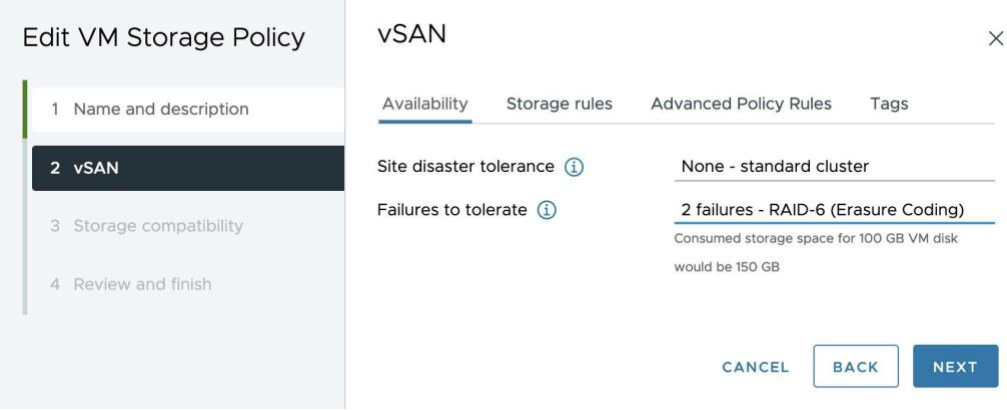
When Auto-RAID is used (9.1 onwards), the storage rules will be simplified, and only include checking if a cluster is encrypted or not.



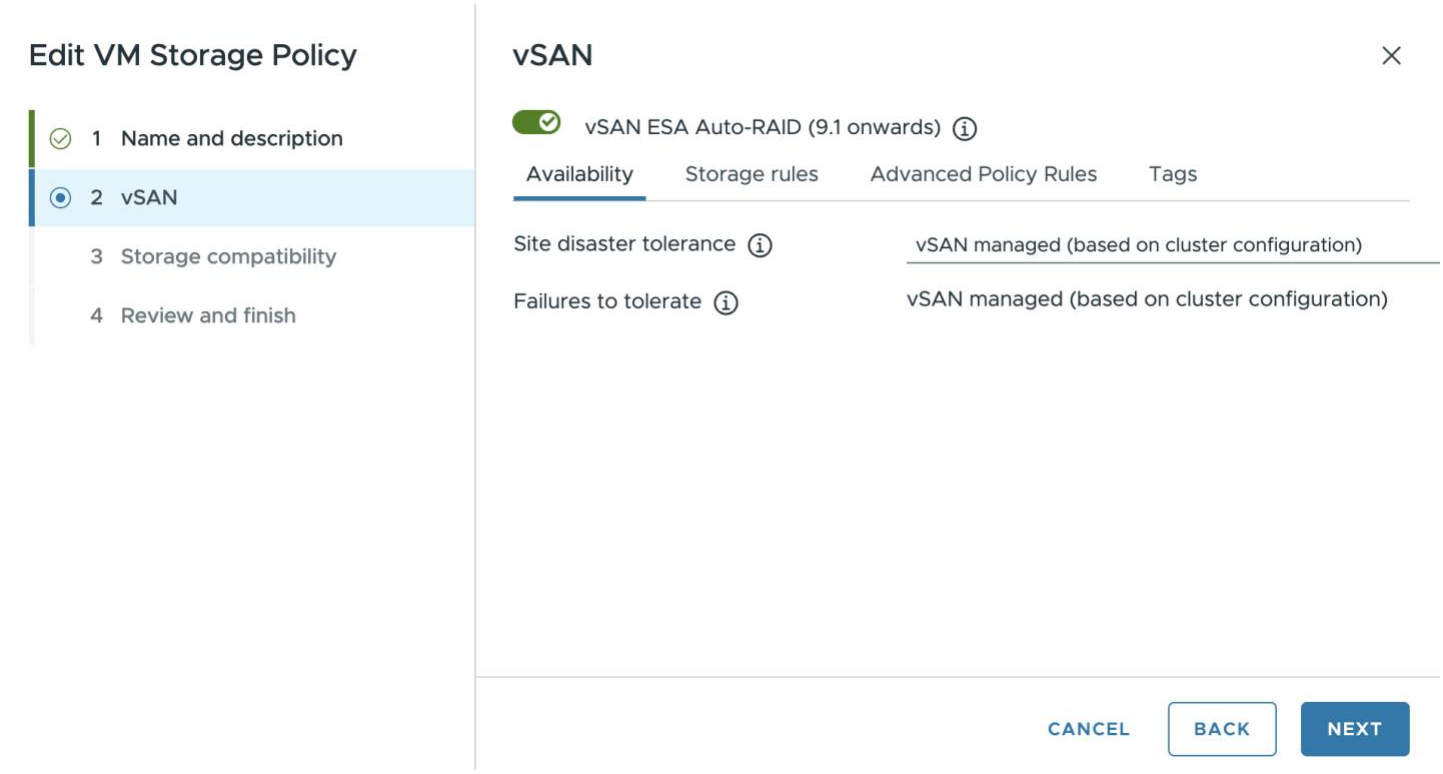
Failure Tolerance Method (vSAN)

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

Further simplification in 9.1 has these settings controlled by vSAN Auto-RAID.



This option simplifies to a vSAN managed policy when vSAN ESA 9.1 Auto-RAID is in use.



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

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.

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. (This applies to OSA only)

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

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.

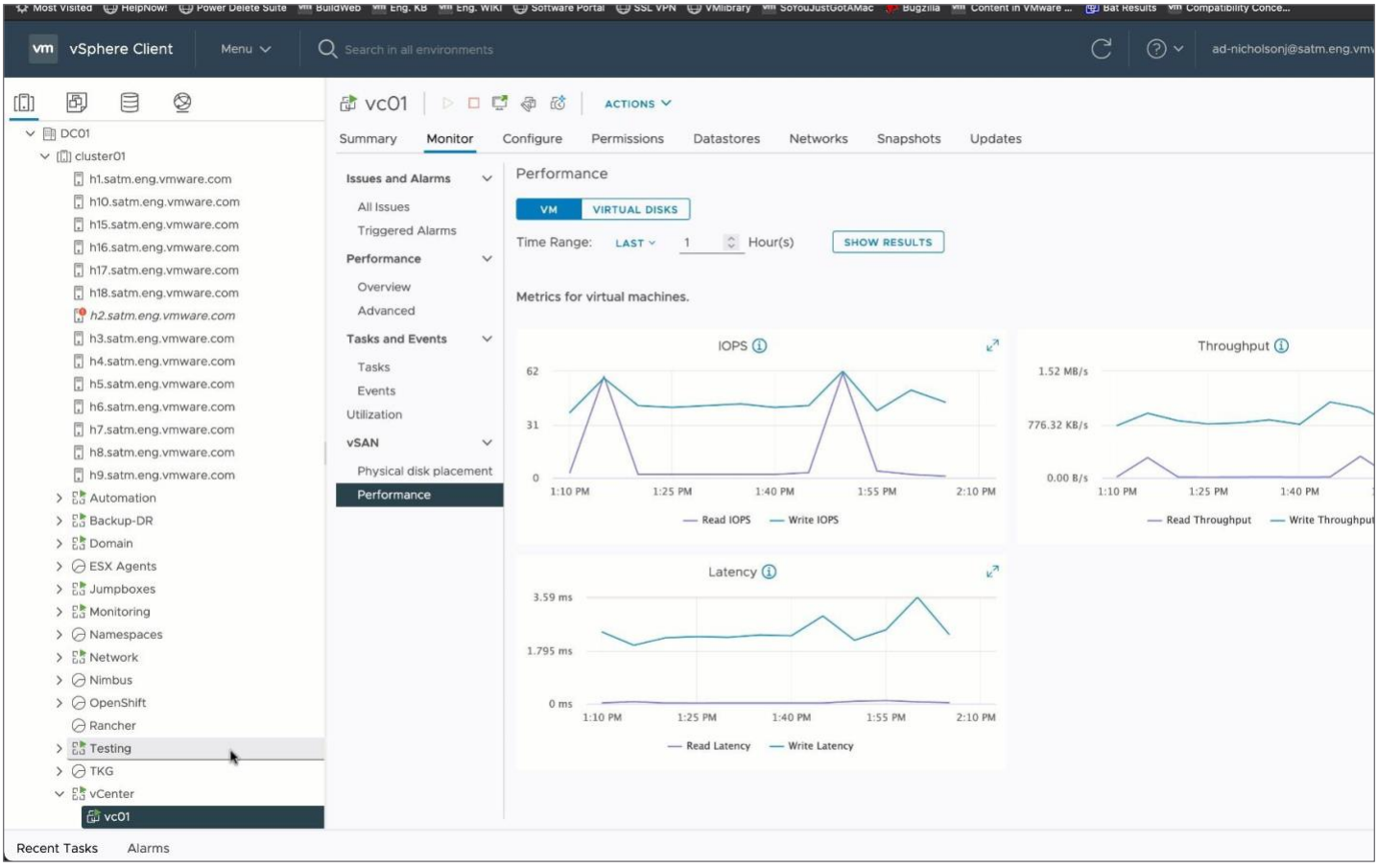
Shortly, a number of capacity sizing examples will be looked at in detail, and will take the considerations discussed here into account.

Changing a VM Storage Policy Dynamically

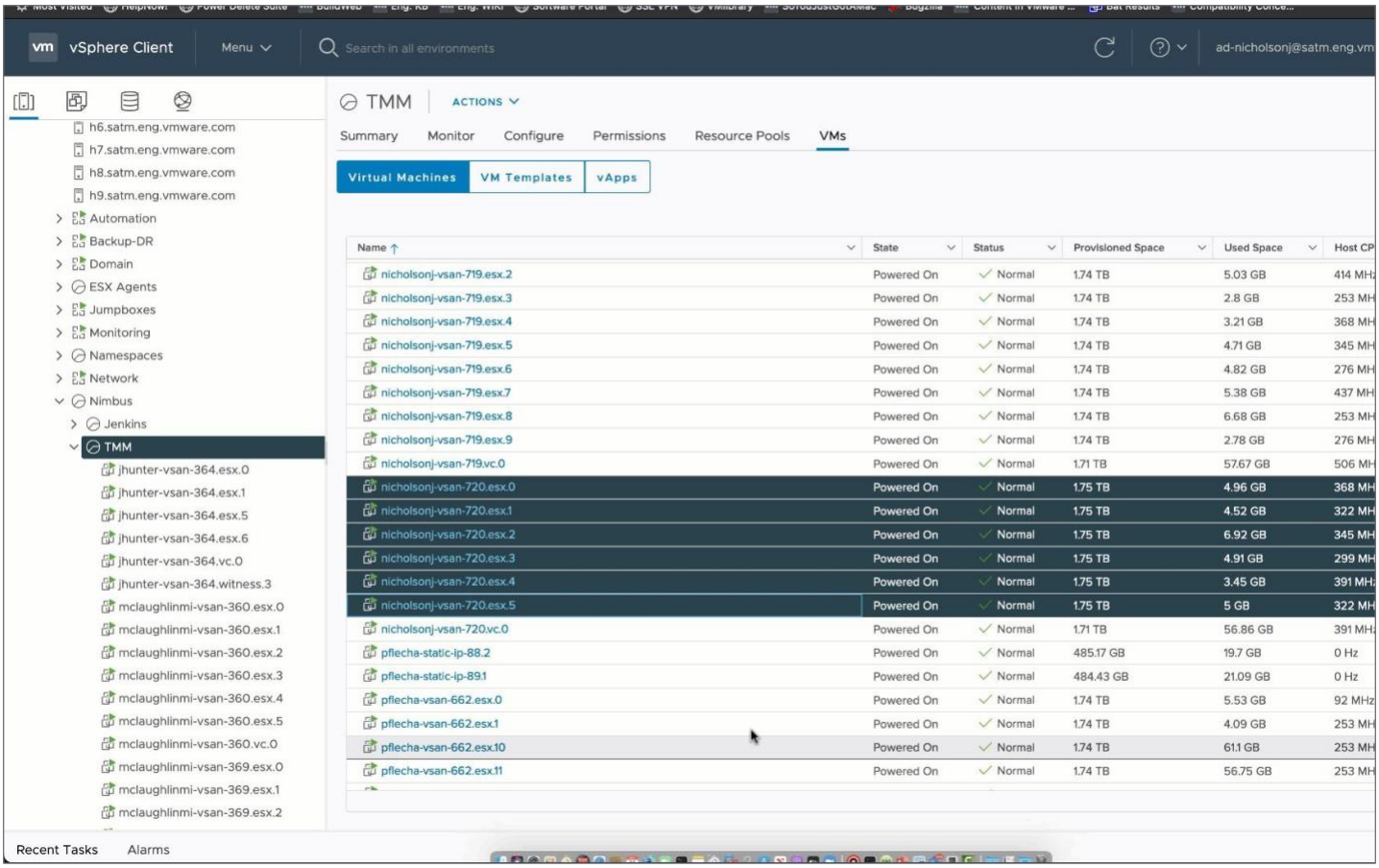
Dynamically changing the policy associated with a virtual machine in a non-disruptive manner has been a core feature of vSAN from the earliest version.

The following examples show how this process works:

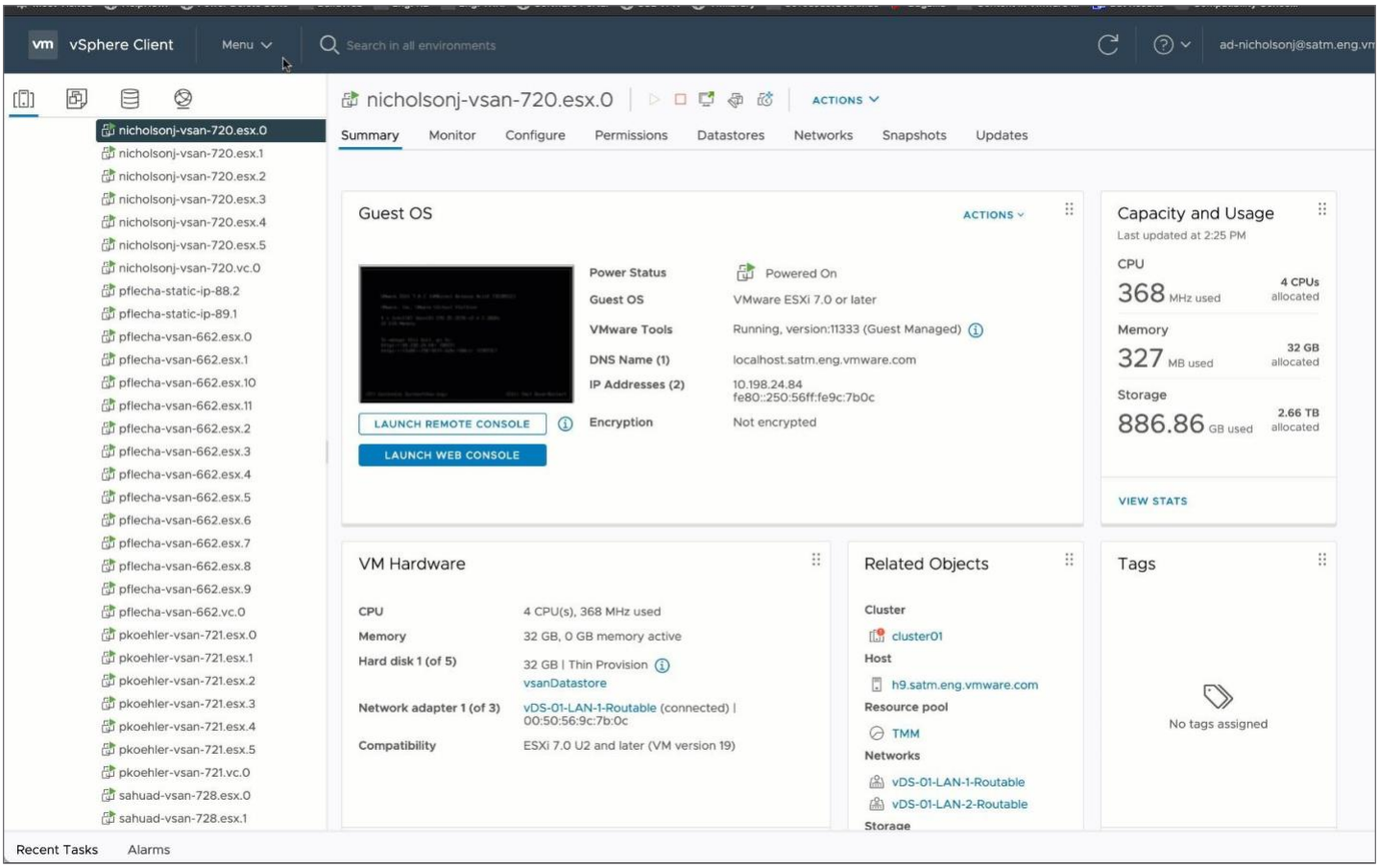
1. Changing the policy associated with a specific virtual machine.



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.

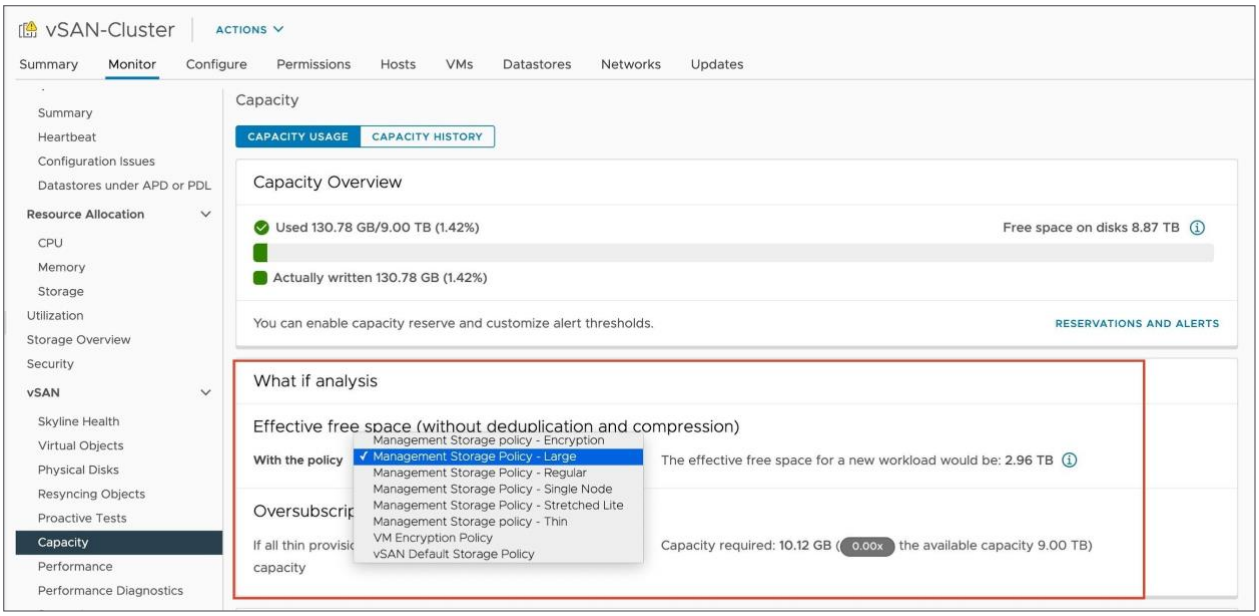
Capacity considerations of policy changes.

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.

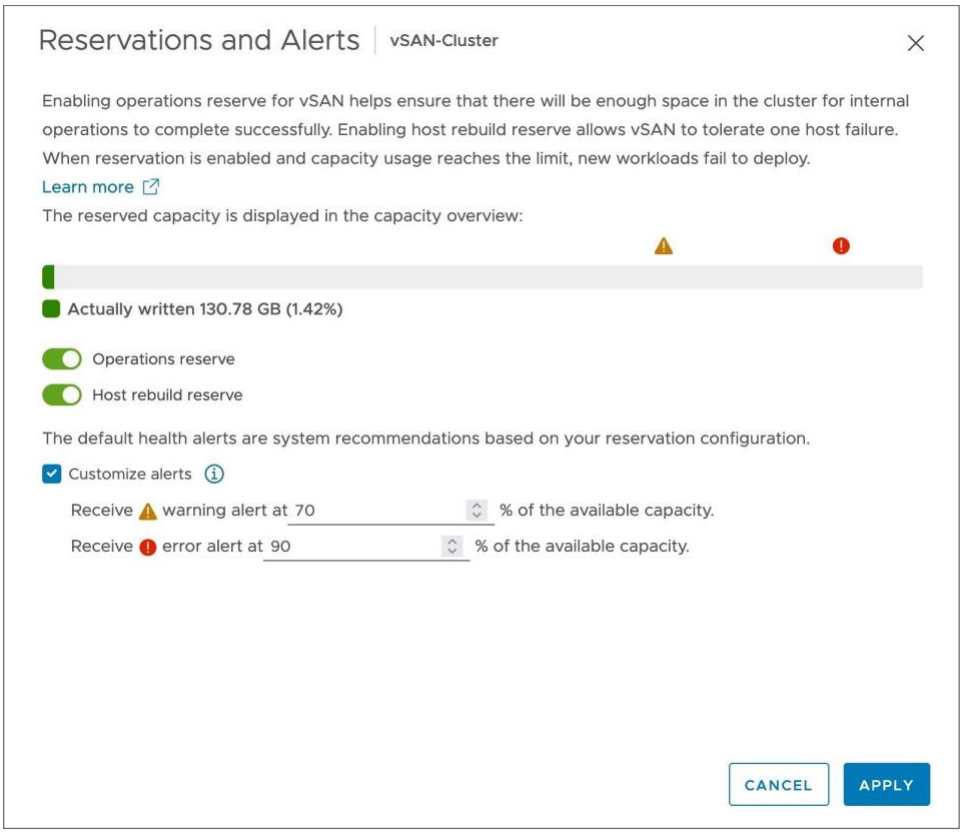
When administrators make a change to a VM Storage Policy and then apply this to a virtual machine to make the change, vSAN will attempt to find a new placement for a replica with the new configuration. If vSAN fails to find a new placement, the reconfiguration will fail. In some cases existing parts of the current configuration can be reused and the configuration just needs to be updated or extended. For example, if an object currently uses RAID 1, and the user asks for

RAID 5, if there are additional hosts to use, vSAN will build out a RAID 5 tree and delete the RAID 1 tree.

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. Now in 9.1 when Auto-RAID is enabled this capacity is already accounted for in the new simplified usable capacity views.

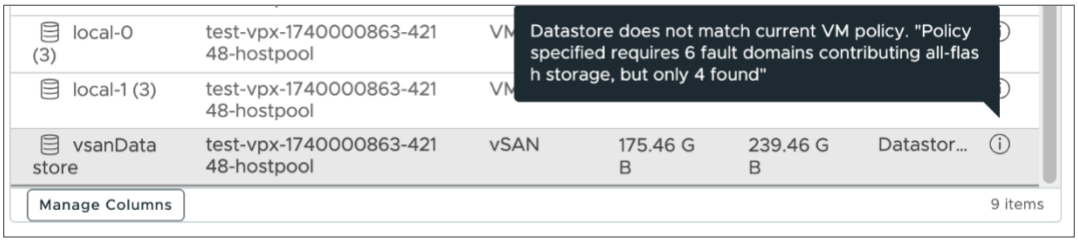


Provisioning a Policy that Cannot be Implemented

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.

Summary of Storage Policy Design Considerations

The following highlights key take aways from this section.

- Any policies settings should be considered in the context of the number of components that might result from said policy.
- StripeWidth will likely not improve performance for vSAN ESA, and is no longer visible when Auto-RAID is in use.
- FlashReadCacheReservation is a legacy policy for hybrid OSA vSAN.
- NumberOfFailuresToTolerate using Auto-RAID no longer 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. Auto-RAID will automatically select raid 6 once 6 hosts are available.
- ForceProvisioning will allow non-compliant VMs to be deployed, but once additional resources/capacity become available, these VMs will consume them to become compliant. This setting is no longer needed, and Auto-RAID will fall back to RAID 0, for bootstrapping, and it will allow provisioning to continue to occur to the surviving side of a stretched cluster operating with a degraded site. This setting is no longer an option when Auto-RAID is in use on a cluster or policy.
- 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 overcommitment that the administrator will need to monitor.

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

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

Older hardware reuse

For VMware VSAN ESA, Intel Icelake or newer are currently generally supported. Cascade Lake and older, can run vSAN OSA, or optionally be connected to a vSAN ESA storage cluster. New to vSAN 9.1, OSA clusters can mount ESA clusters, and vice versa, making it possible to connect newer ESA backed storage clusters to older compute clusters running vSAN OSA. .

For more information on how to use older servers best with vSphere and VSAN see the following post: VCF 9.0 Server Certification: Preserving Your Hardware Investment and giving the best ROI

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)

A minimum of 128GB is required for vSAN ESA AF-0 hosts. The vSAN sizer includes memory overhead requirements.

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. The 9.1 Zero Touch Provisioning is supported.

Despised State Configuration Support

vSAN 9.1 includes integration for desired state configuration. vSphere Configuration Profiles ensures that configuration and remediation changes do not negatively impact vSAN. vSAN maintenance mode policies and object accessibility policies are honored when remediating vSAN clusters. Advanced vSAN configuration can be applied cluster-wide.

Core Dump

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 deprecated 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 Devices

vSAN as of vSphere 7 Update 3 Supports the use of TPM 2.0 devices to cache encryption keys used by vSAN Encryption with 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 Clusters and Hosts

Compute-Only Hosts

While it is supported to add a compute only node to a vSAN cluster, it operationally is generally not advised. It is best to extend clusters in a like for like fashion to scale out capacity, storage performance, object counts as the cluster grows.

Compute only clusters

Compute only clusters are commonly used in situations where:

- Re-use of older hardware that may not be suitable for vSAN, or VSAN ESA.
- Extreme isolation of database applications for licensing, or other concerns.
- Service provider enviroments with dedicated clusters per tenants
- Segmentation of control between storage management and compute management

Compute clusters, can often reuse existing more limited networking, especially if the vSAN storage cluster connected to them will utilize separated networking for it's back end storage network. There is no requirement that the compute clusters conform to the ReadyNode certification list, or CPU generation list.

For extended guidance of designing and operations of storage clusters see Design and Operational Guidance for vSAN Storage Clusters

Recommendation: 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: licensing for vSAN is now based on raw capacity of the storage devices in the cluster. No special "client licenses" are required, for compute only clusters.

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 utilize RAID 5 or RAID 6 (6 hosts).
2. Consider the number of capacity devices left on the remaining hosts to handle rebuilds, free space, and component count maximums.
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

Blade and Compostable 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 often simply not enough disk slots in the hosts.vSAN storage clusters, and

datastore sharing from HCI clusters, offers a way to leverage existing blade investments. For more information see the updated Design and Operational Guidance for vSAN Storage Clusters.

External Storage Enclosure Considerations

As of vSAN ESA 9.1, external storage enclosures are not supported. While work has been tested with storage partners, this is not currently a supported configuration.

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

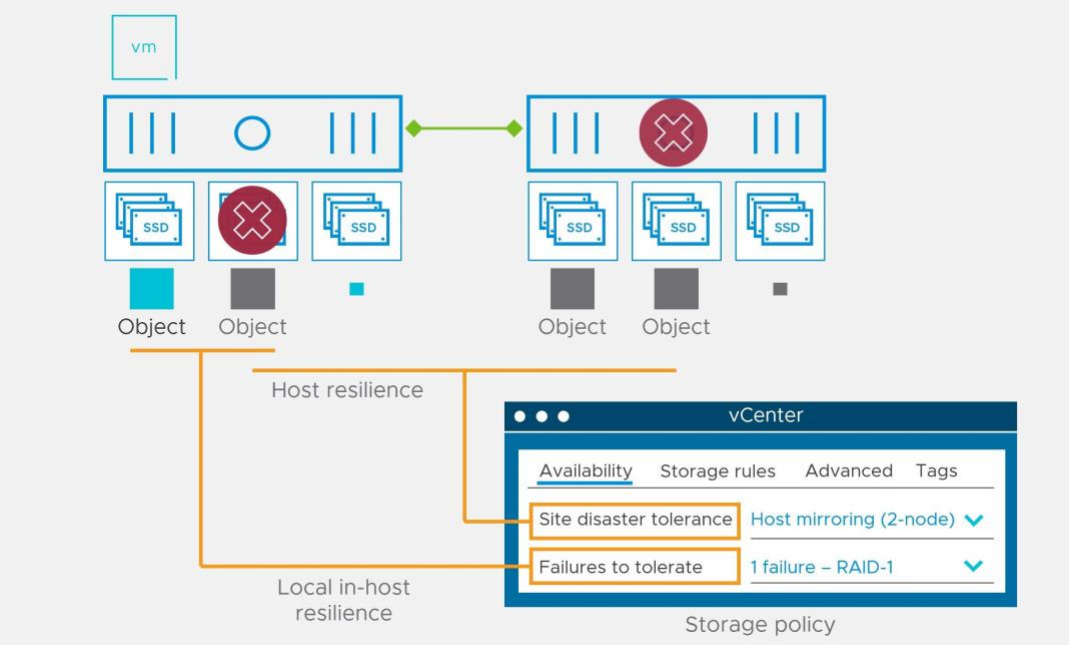
This section of the guide looks at cluster specific design considerations.

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 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 drives RAID 5 (2+1) can be used. For additional information on this capability see the vSAN 2-Node Cluster Guide.



3 - Node Considerations

vSAN with 3 hosts will use RAID-1) 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. For small edge locations, also consider vSAN 2 Node deployments were you can use a RAID 5 inside the host (2+1) and combine it with a mirroring of that data between the two hosts.

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.

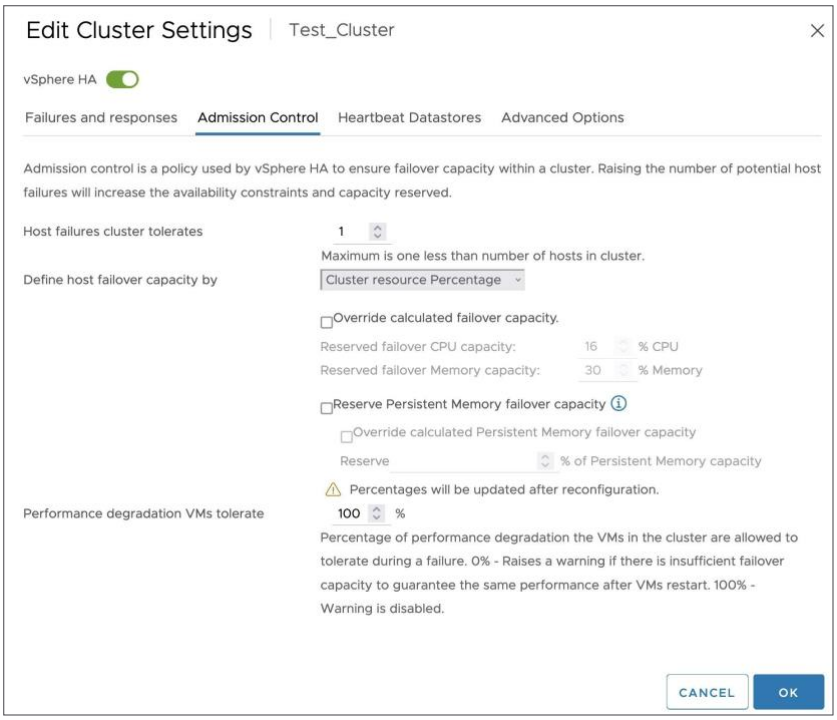
There are a number of requirements for vSAN to operate with vSphere HA.

- 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.

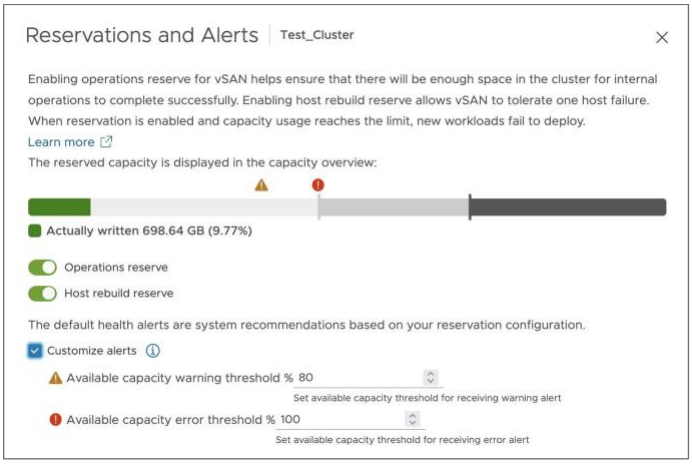
One major sizing consideration with vSAN is interoperability with vSphere HA. Current users of vSphere HA are aware that the NumberOfFailuresToTolerate setting will reserve a set amount of CPU & memory resources on all hosts in the cluster so that in the event of a host failure, there are enough free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

HA Admission 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 Admission Control does not reserve storage capacity, a vSAN cluster using Auto-RAID will automatically reserve capacity for host failure in it's calculations of free space shown at the Cluster → monitor → vSAN → Capacity view. While the legacy operational reserve, and Host rebuild reserve can still be found under the cluster configuration option (along with the alarms) this is fundamentally legacy and redundant with the new effective capacity views when using Auto-RAID.



VMCP Consideration

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.

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

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.

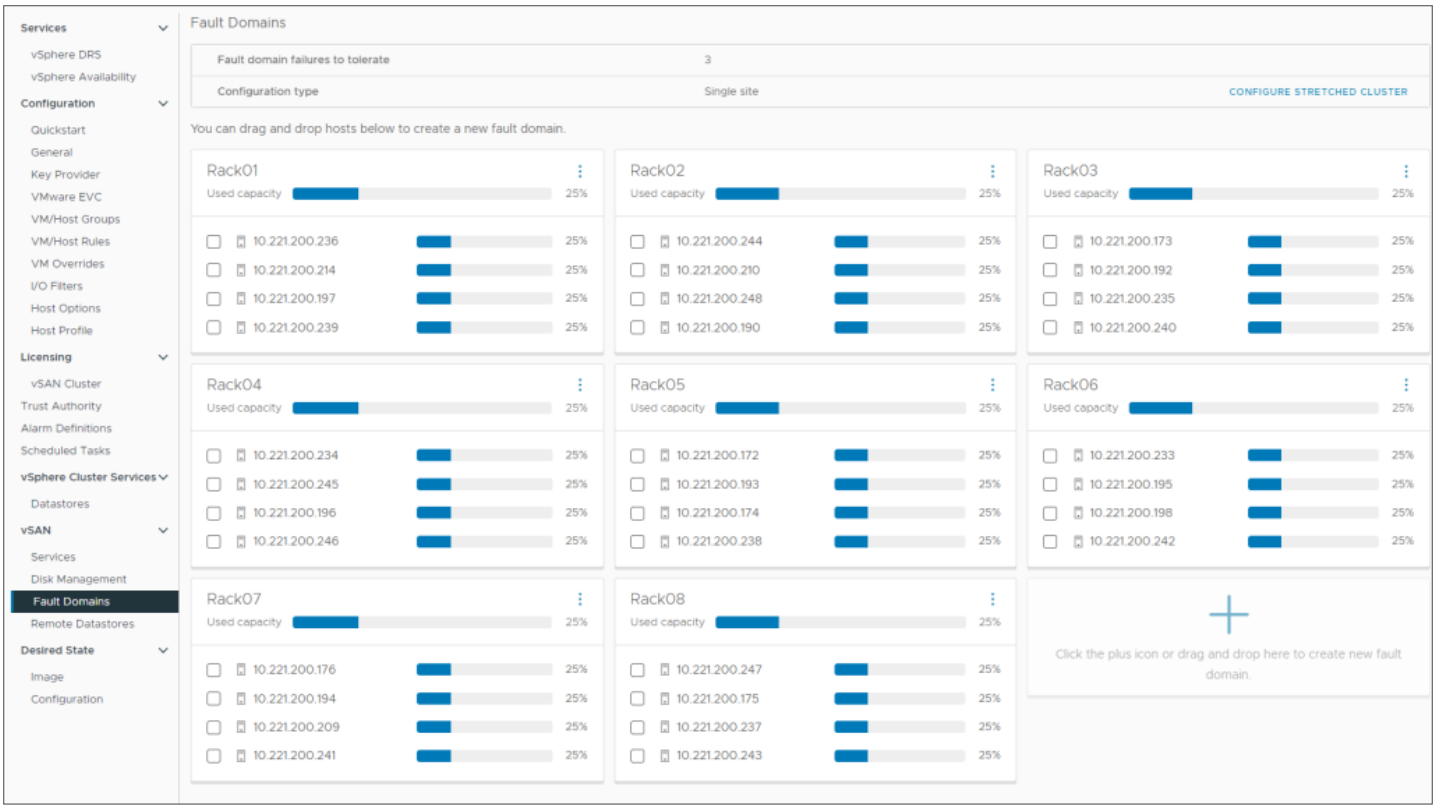
Note: “Shutdown” may also be used anytime “power off” is mentioned if it is likely that a VM will retain access to some of its storage but not all during host isolation. (This is unlikely in the case of a vSAN datastore.) However, note that in such a situation some of its virtual disks may be updated while others are not, which may cause inconsistencies when the VM is restarted. Further, a shutdown can take longer than power off.

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

Fault Domains with vSAN ESA

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.

Using the Fault Domains feature in a cluster running vSAN’s Express Storage Architecture allows you to provide unique resilience capabilities that align with your physical topology



Recommended Minimums for Fault Domains

Overall cluster size. Since the fault domains feature within vSAN is designed to provide rack or room-level resilience in the event of a failure, it is really intended for larger clusters. Given the recommendations of having at least 3 hosts within each fault domain, and one more fault domain than is absolutely required by the desired storage policy, a realistic minimum host count would look like the following:

FTT=1 using RAID-5 (using its 2+1 scheme plus one extra fault domain) is 12 hosts: 4 fault domains with 3 hosts in each.

FTT=1 using RAID-5 (using its more space efficient 4+1 scheme plus one extra fault domain) is 18 hosts: 6 fault domains with 3 hosts in each **note: this option is no longer used by Auto-RAID.**

FTT=2 using RAID-6 (using its 4+2 scheme plus one extra fault domain) is 21 hosts: 7 fault domains with 3 hosts in each.

There are a few features that are unavailable or unsupported when using vSAN's Fault Domain feature.

- Reserved Capacity toggles. As noted in the post, "Understanding Reserved Capacity Concepts in vSAN" and "Design and Operation Considerations when using vSAN Fault Domains" when a cluster (ESA or OSA) is configured with the fault domains feature, the "Operations Reserve" and "Host Rebuild Reserve" toggles are not available. It is recommended to follow free capacity recommendations in vSAN prior to those capabilities coming out, which was to maintain about 25% of free capacity.
- vSAN ESA Auto-Policy Management feature. The new Auto-Policy Management Capabilities with the ESA in vSAN 8 U1 are currently not supported when using the fault domains feature in vSAN ESA.

Remember that it is not possible to configure a cluster using vSAN's Fault Domains feature with just 2 fault domains. You must have at least 3 fault domains, with the recommended number of fault domains being more depending on which RAID data placement scheme is used.

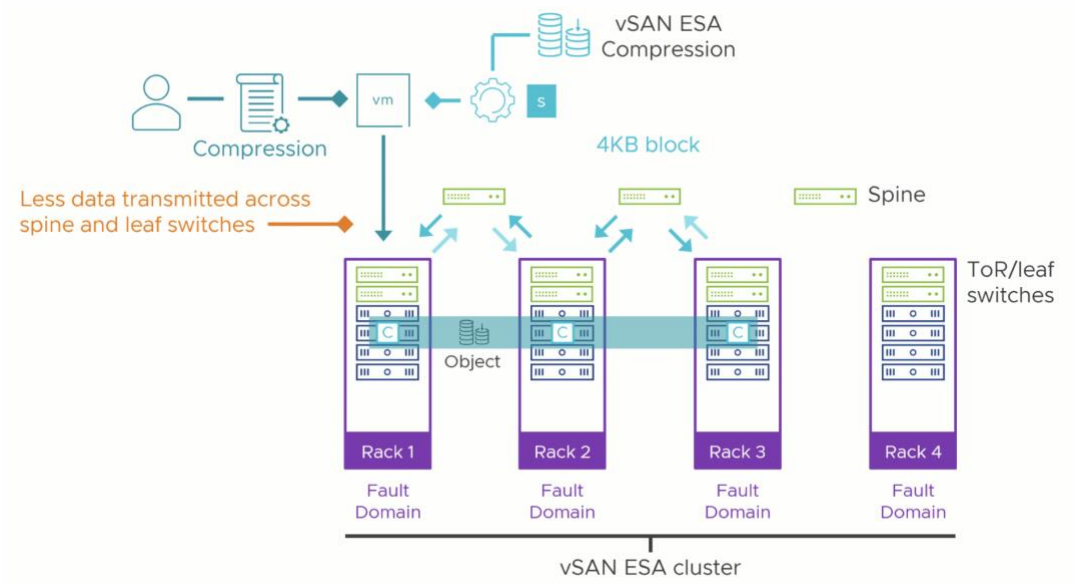
Design decision: It is best to build fault domains with the same number of hosts per fault domain. Hardware configurations should be consistent across the entire cluster. Do not activate this feature without careful review of the considerations it has for impacting host rebuild and rebalance data placement. For a deeper explanation of this feature and vSAN ESA see this blog.

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.

Two improvements to vSAN ESA improve operations for fault domains.

Compression on the network reduces the amount of bandwidth used between hosts. This is especially valuable in cases where spine switches are over subscribed that separate fault domains. Do pay special attention to bandwidth saturation of the network with fault domains, and make sure to not undersize leaf to spine bandwidth

RAID 5 can now be done with 2+1. This lowers the minimum number of fault domains to 3, and for a N+1 design lowers the minimum number of fault domains to 4.



This is true for compute resources as well. In the use of fault domains, 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. It is recommended to not use this feature if the groupings of hosts do not align with any unique physical failure domains as it will increase capacity and compute requirements, and potentially slow rebuilds

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 faultdomain.

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:

Assuming 6 racks with 3 hosts each, assuming a policy of FTT=2 with RAID6, to be able to place a host 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:

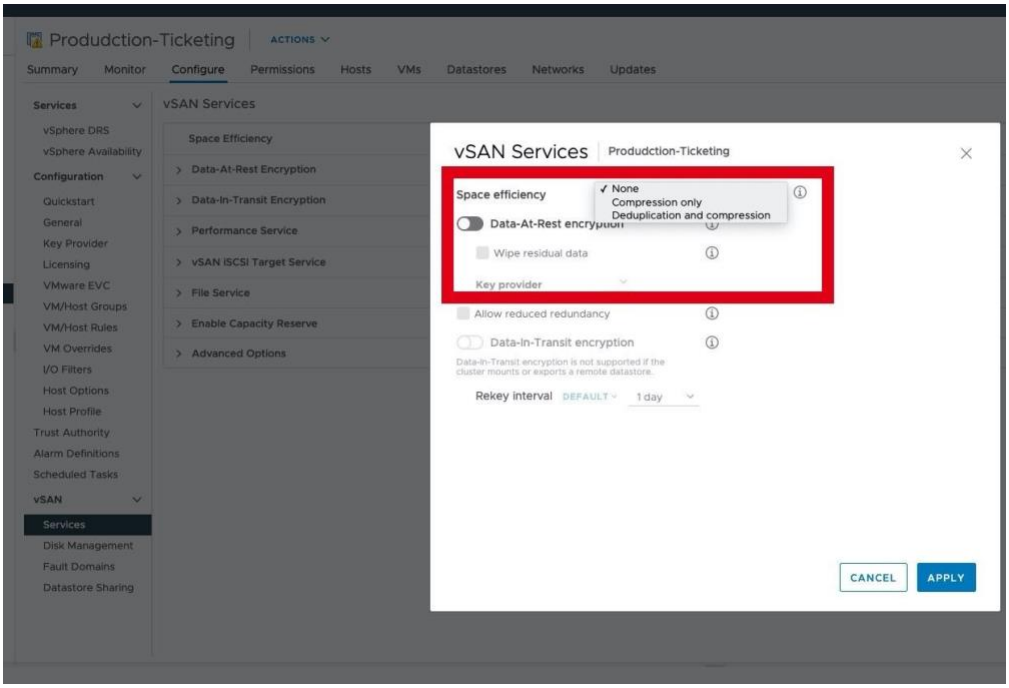
- F= number of total Fault Domains
- R= number of required fault domains to satisfy the policy
- D= number of hosts in Fault Domain
- T= number of hosts to put into maintenance (or hosts to tolerate the failure of... your 'Tolerance' value)
- A= active data per host
- (Active data per host x Number of 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))

Additional Content:

Read "Considerations using vSAN fault domains"

Deduplication and Compression Considerations

vSAN 9.1 Express Storage Architecture by default now enables compression only as a cluster service. Optionally, global deduplication can be enabled as a cluster service.



vSAN 9.1 (ESA) Deduplication and Compression and space efficiencies

A dedicated vSAN Space Efficies technote can be found here.

vSAN 9.1 Compression

vSAN for VCF 9.1 introduces the use of the Zstandard (ZSTD) compression algorithm. This enhanced data compression capability works well across all workload types, which is the reason why it is now an always-on feature of the cluster.

vSAN 9.1 introduces deduplication under general availability when using vSAN ESA. This is a cluster-wide deduplication technology that is designed for maximum levels of space efficiency with minimal impact on guest VM workloads. The intelligent design of ESA's deduplication technology will adaptively throttle the use of spare resources so that there is not a compromise in performance of guest VM workloads. Furthermore, deduplication's integration throughout the storage stack means that the common vSAN data services are fully compatible with deduplication, encryption, snapshots, and storage clusters. As of 9.1 this is not yet supported with stretched cluster configurations, or 2-node clusters.

vSAN in 9.1 no longer uses the multiplier of reduction (or savings) notation, instead space savings uses a capacity ratio.

vSAN 9.1 also makes visible the unique compression and deduplication ratios in the capacity view.

Sample space efficiencies:

10 hosts with approximately 4TB per host would provide just under 40TB raw capacity (more specifically, ~39TB) as seen in the legacy capacity view. This same cluster, as viewed in the new Effective Capacity View would show just over 20TB in effective capacity (more specifically, 2.03TB), or

approximately half of the raw capacity. But the actual “used capacity” on the cluster would be after all of the space efficiency techniques (compression, deduplication, thin provisioning, snapshot savings, etc.).

•An estimate of effective usable capacity could be easily derived with the following equation:

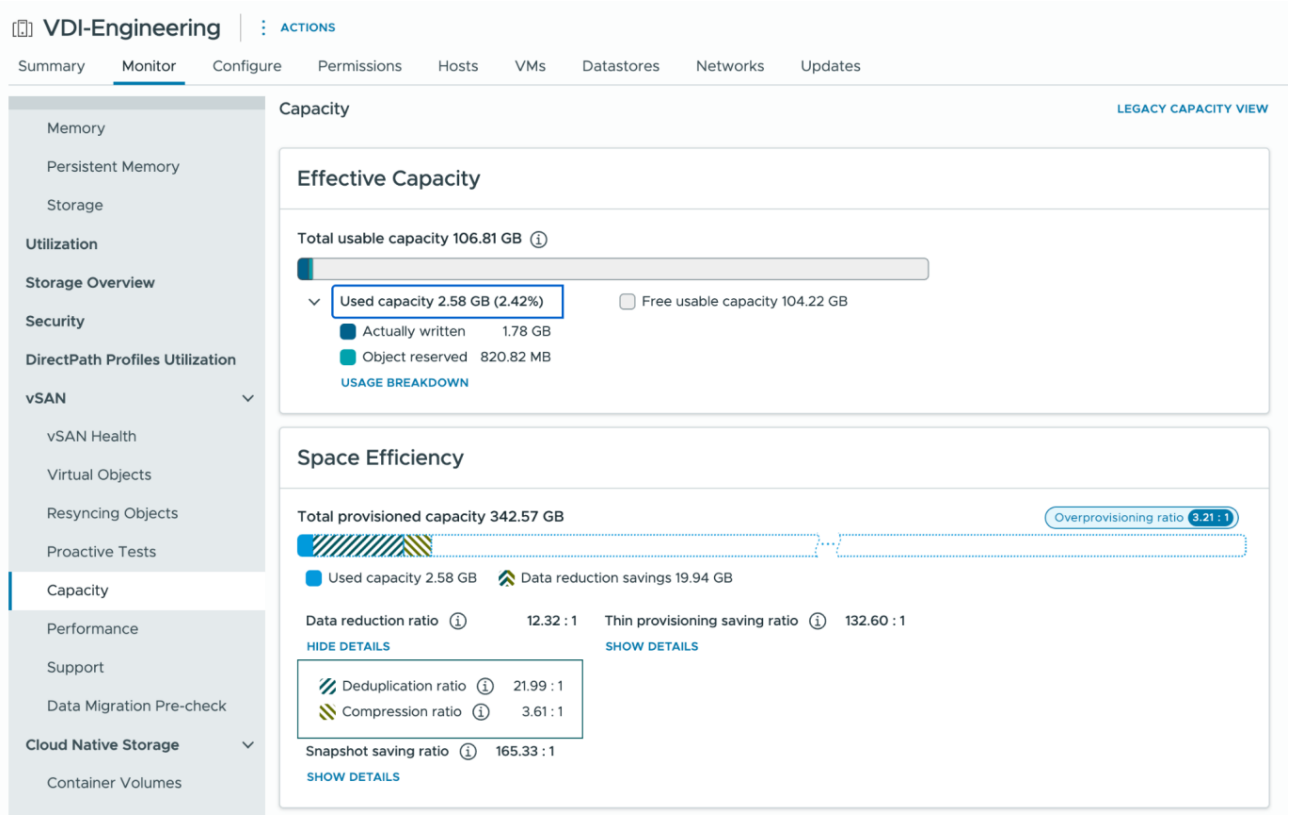
•In this example, if the customer was getting a 2:1 data reduction ratio, they would be able to store up to about 40TB, roughly equal to the aggregate raw capacity of the hosts.

•In this example, if the customer was getting a 4:1 data reduction ratio, they would be able to store up to about 80TB, about 2x the aggregate raw capacity of the hosts.

•In this example, if the customer was getting a 6:1 data reduction ratio, they would be able to store up to about 120TB, about 3x the aggregate raw capacity of the hosts.

In this example, if the customer was getting an 8:1 data reduction ratio, they would be able to store up to about 160TB. About 4x the aggregate raw capacity of the hosts. :

- 2:1 multiplier of reduction means that consumption is just 50% of the original size, and a savings of 50%
- 4:1 multiplier of reduction means that consumption is just 25% of the original size, and a savings of 75%
- 8:1 multiplier of reduction means that consumption is just 12.5% of the original size, and a savings of 87.5%



Deduplication Limitations

Deduplication is now supported with data at rest encryption, but as of 9.1 is not yet supported with stretched clusters.

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.

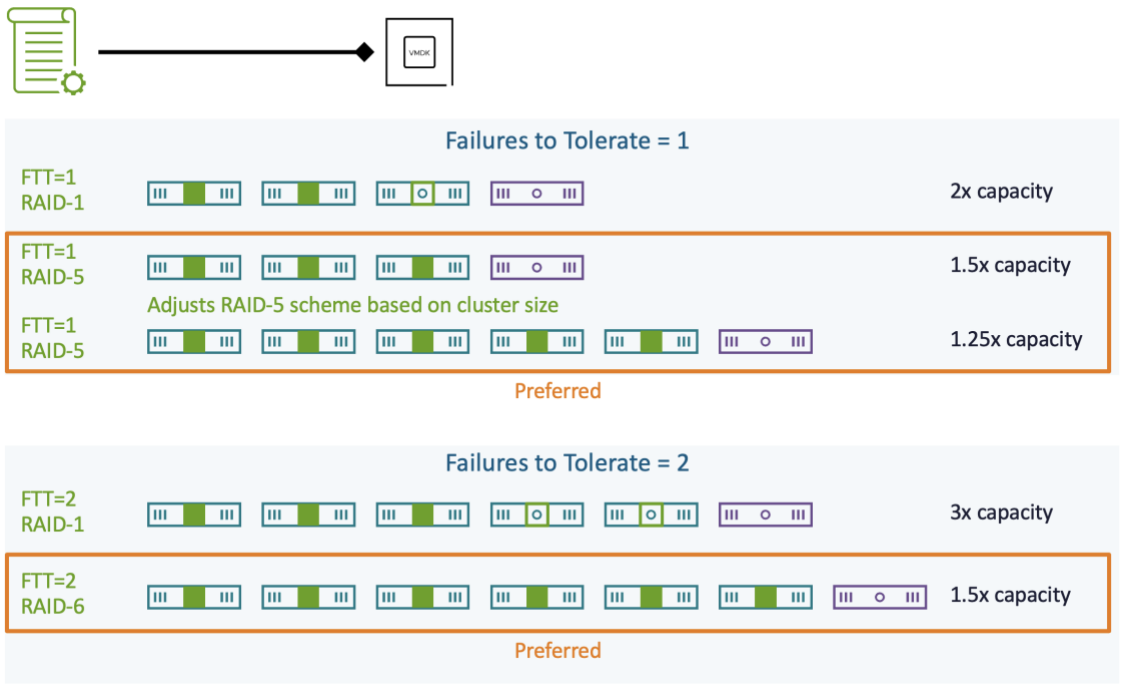
A vSphere cluster defines a boundary of shared resources. With a traditional vSphere cluster, CPU and memory are two types of resources managed. Network-aware DRS is omitted here for clarity. Adding additional hosts would indeed add additional CPU and memory resources available to the VMs running in 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.



New with vSAN 8 ESA, RAID 5 using Auto-RAID will always use the 2+1 strike width, and raid 6 will always use 4+2. Therefore, whether the cluster is 4 hosts or 64 hosts, the capacity overhead 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.

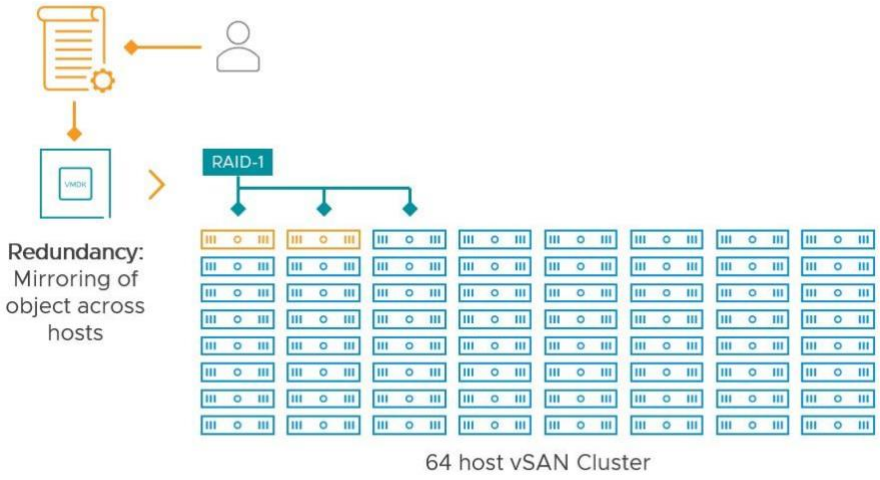


Figure 2. A VM with the legacy RAID-1 policy using mirroring, in a 64 host cluster

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.

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 leaf to spine connectivity is not-oversubscribed, then allocating ultra deep buffer switches (Switches, with multi gigabyte port buffers) to the leaf can help improve performance latency consistency.

If a cluster is expected to grow significantly in performance requirements, consider deploying adequately powerful switches at initial deployment. Also, many common 25Gbps top of rack switches are limited to 48 access ports, while 100Gbps switches are commonly limited to 32 or 64 ports. Larger clusters may require additional throughput between switches to prevent these links from becoming bottlenecks, or adequately allow enough bandwidth to the spine switch. 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, while retaining compatibility with older 25Gbps hosts. 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 for nondisruptive updates of the cluster, some customers may choose 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:

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

Summary

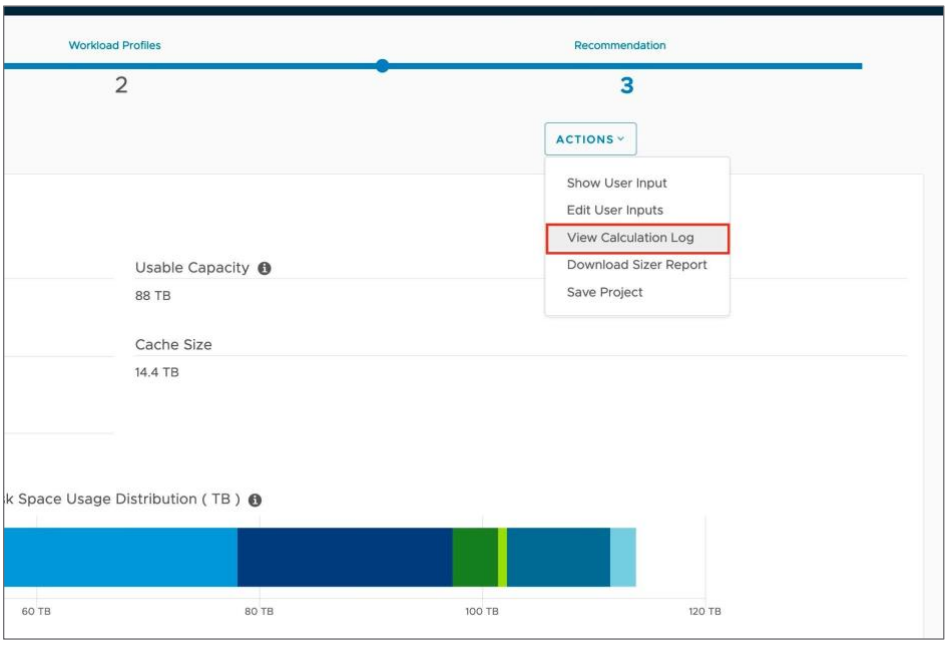
vSAN's approach to data placement means that it does not wide-stripe data across all hosts in a cluster. Hosts in a vSAN cluster that are not holding any contents of a VM in question, will have neither a positive or negative impact on the performance of the VM. Given little to no resource contention, 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.

Sizing Examples

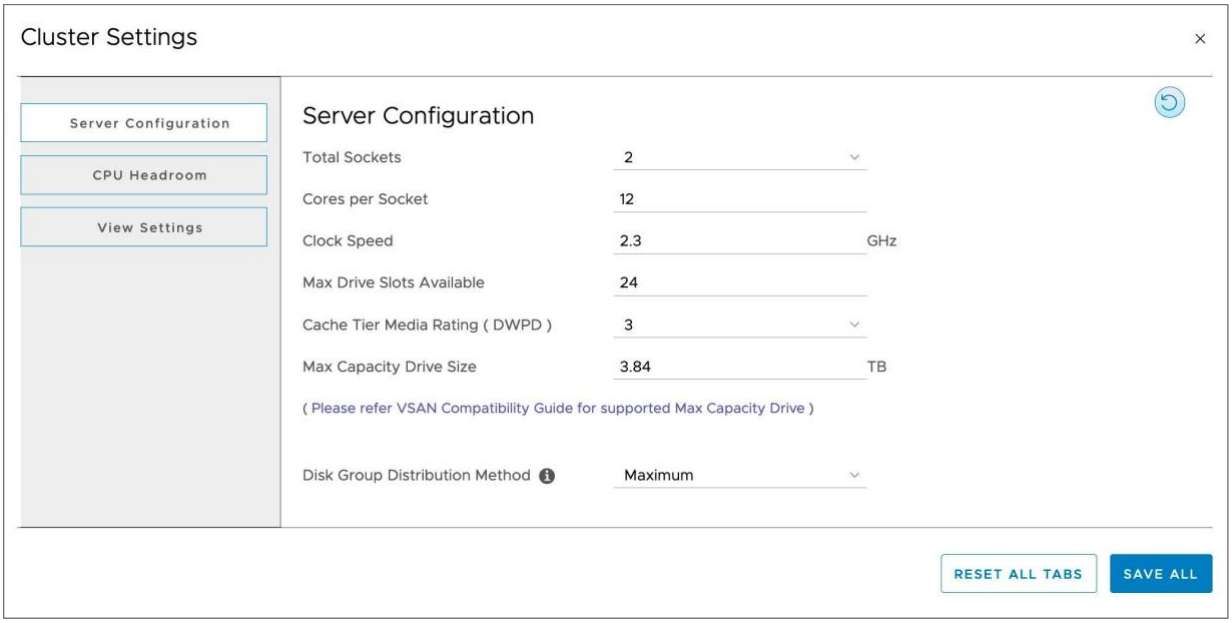
Previously this section of the design and sizing guide included written out explanations of how to hand calculate the overheads when designing a vSAN cluster. This section has been replaced by the vSAN ReadyNode Sizer that can be found at "https://vsansizer.vmware.com/".

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

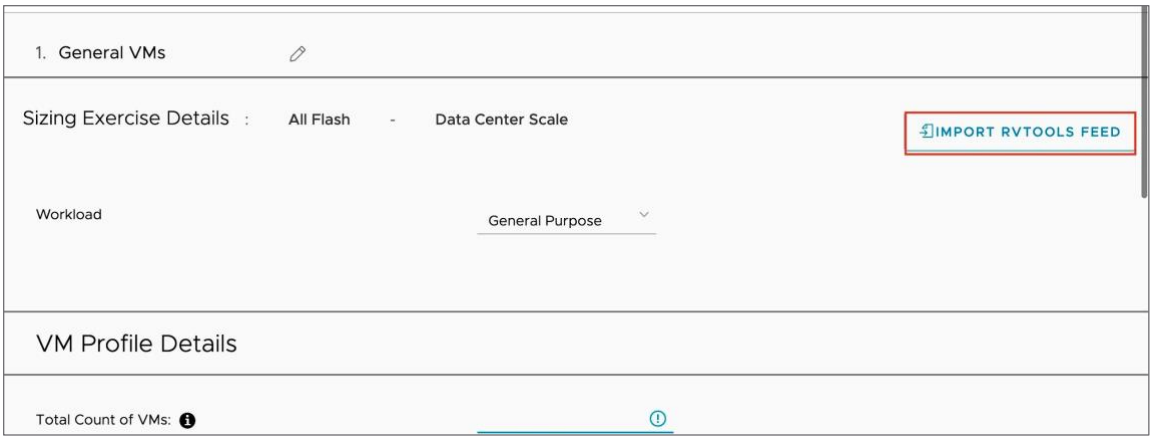
If you wish to understand the specific overheads in relation to why it sized a cluster a specific way the math remains available in the "View Calculation" log section.



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 BCG-listed components, drivers and firmware, resulting in unpredictable behavior. Flash devices and IO controllers are particularly sensitive.
- Not properly sizing for capacity growth (e.g. thin volumes progressively getting fatter), resulting in declining performance over time.
- Under sizing the network. Avoid using 10Gbps with capacity or RAM dense hosts. vSAN ESA can easily saturate a 25Gbps link if the workload demands it.
- Not understanding 3-node limitations associated with maintenance mode and protection after a failure.
- Failure to have sufficient extra capacity for operations like entering maintenance mode, changing policy definitions, etc.

Additional Resources

The following are a collection of useful links that relate to bandwidth sizing for vSAN stretched clusters.

Performance Recommendations for vSAN ESA. This is a collection of recommendations to help achieve the highest levels of performance in a vSAN ESA cluster. Many of these same recommendations apply to vSAN storage clusters.

vSAN Proof of Concept (PoC) Performance Testing. This is a collection of recommendations that will guide users to test the performance of a vSAN cluster. While it is currently written for the OSA, many of the testing methods used are also applicable to the ESA.

Design and Sizing for vSAN ESA clusters. This post offers some nice guidance on using the vSAN Sizer for the ESA that summarizes some key points that can be found in the VMware vSAN Design Guide.

vSAN Network Design Guide. This network design guide applies to environments running vSAN 8 and later.

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

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

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

About the Author

John Nicholson is a Product Marketing Engineer in the VCF division at Broadcom.

vmware® by Broadcom