Design and Operational Guidance for vSAN Storage Clusters

Recommendations for vSAN 8 U3 and VMware Cloud Foundation 5.2

March 5, 2025

vmware by Broadcom

Table of Contents

Introduction

VMware vSAN storage clusters (previously known as vSAN Max™) is a new deployment option within vSAN that provides highly flexible and scalable disaggregated storage for vSphere clusters, powered by the vSAN Express Storage Architecture™, or ESA. It gives customers an ability to deploy a highly scalable storage cluster to be used as primary storage for vSphere clusters*, or augment storage for traditional vSAN HCI clusters.

vSAN storage clusters are powered by the ESA, which provides tremendous flexibility in meeting performance, capacity, and resilience requirements for all types of environments. Since the ESA is designed to power traditional vSAN HCI clusters and disaggregated vSAN storage clusters, the flexibility of ESA may invite non-optimal configurations when deployed as a vSAN storage cluster. The following is a collection of recommendations in the design, operation, and optimization of vSAN storage clusters. Which deployment option is best for you? See: "vSAN HCI or vSAN Storage Clusters - Which Deployment Option is Right for You?" for more information.

Scope of Topics

The information provided in this document will assume the use of vSAN 8 U3, and/or VMware Cloud Foundation (VCF) 5.2. VCF deployments may have additional requirements and support limitations that fall outside of the scope of this document. Please see the "Administration Guide for VMware Cloud Foundation 5.2" for official VCF guidance.

This document is arranged in the order of Planning and Sizing, Day-0 Initial Deployment and Configuration, and Day-2 Operations to help step you through the guidance in an orderly manner. "vSAN storage clusters" is the official name of vSAN's disaggregated deployment option in VCF 9 and newer. Some illustrations, screen captures and hyperlinks in this document may still refer to the deployment option as "vSAN Max" – the previous name used in VCF 5.2.

Planning and Sizing

vSAN ReadyNode Host Specifications and Sizing

Use vSAN ReadyNodes certified for vSAN storage clusters.

vSAN storage clusters must be deployed with vSAN ReadyNodes that are certified for this deployment option. These are special vSAN ReadyNodes that share similarities to vSAN ReadyNodes for ESA, but are optimized for the demands of a storage-only cluster. For more information, see the post: "ReadyNode Profiles Certified for vSAN Max." Recent changes were made that lowers the minimum hardware and cluster requirements. See the post: "Greater Flexibility with vSAN Max through Lower Hardware and Cluster Requirements" for more details.

Note that with vSAN, performance of a VM is derived from the host hardware and the network used to interconnect the hosts in the vSAN cluster, not the cluster host count. While increasing the host count of a cluster will increase the aggregate IOPS and bandwidth achieved by the cluster, in most cases it will not improve the discrete performance capabilities observed by the VM. VM performance will be a function of the host hardware and network connectivity. If high performance is a requirement priority, use higher performing hosts and fast networks with vSAN Max storage clusters to achieve the desired result. See the post: "Performance Capabilities in Relation to vSAN Cluster Size" for more information.

Know what can and cannot be changed in a ReadyNode certified for use with vSAN storage clusters.

The vSAN ESA ReadyNode program for hosts configured in an aggregated, vSAN HCI configuration offers a lot of flexibility in customizing ReadyNode configurations, as noted on the document "What you Can (and Cannot) Change in a vSAN ESA ReadyNode."

Ensure any adjustments to a ReadyNode certified for vSAN storage clusters do not compromise its design objective.

vSAN ReadyNodes certified for vSAN storage clusters are designed to meet specific performance and capacity objectives. There are five profiles to address a wide variety of use cases. While the ReadyNode program is flexible, reducing a resource type such as the quantity of storage devices per host, or using insufficient network bandwidth may inhibit the desired performance capabilities associated with that given ReadyNode profile.

Understand how a desired raw cluster capacity can be achieved differently, and how this may affect resource utilization.

vSAN storage clusters can achieve desired capacity goals in many ways. For a given amount of raw capacity serving the same number of VMs, one can use fewer hosts with a higher density of resources per host, or use more hosts with a lower density of resources per host. Each have their advantages and disadvantages. The vSAN ReadyNode sizer will provide what it determines as an ideal configuration based on your design inputs, and in most cases this will be sufficient. One may wish to adjust the specifications to meet your other design preferences.

vSAN storage clusters **using fewer hosts with higher density of resources per host.**

Advantages	Disadvantages
Lower hardware costs	Larger percentage of resource impact upon a failure of a host
Able to meet capacity objectives and stay within recommended maximum host count for a vSAN storage cluster (24).	Potentially more strain on any given network uplink supporting host.
Less rack space* used with fewer network ports used on ToR switches. *Comparison when using server size of the same physical form factor such as 1U, 2U, 4U.	Higher likelihood of running into per host component limits (27,000) for vSAN ESA.
Higher number of hosts in client cluster(s) that can mount the vSAN storage cluster datastore while staying under the total host limit (128).	May not meet recommended cluster minimums for desire topology and resilience levels (ex: Stretched clusters using secondary levels of resilience through RAID-6)
	Lower aggregate performance across the cluster because of a higher concentration of workloads (and their working sets) on any given host.
	Fewer hosts means a lower allowed component count for the cluster. Depending on the makeup of the data, in some cases this may limit capacity. **A vSAN storage cluster with a lower host count will be able to support fewer VMs/data stored on the datastore.**

vSAN storage clusters **using more hosts with lower density of resources per host.**

Advantages	Disadvantages
Lower percentage of resource impact upon a failure of a host.	Higher hardware costs
Potentially less strain on any given network uplink supporting host.	May not be able to meet capacity objectives while staying within recommended maximum host count for a vSAN storage cluster (24).
Potentially faster resynchronizations because of less resource contention and distributing effort across more hosts.	More rack space* used with fewer network ports used on ToR switches. *Comparison when using server size of the same physical form factor such as 1U, 2U, 4U.

Lower likelihood of running into per host component limits (27,000) for vSAN ESA.	Fewer number of hosts in client cluster(s) that can mount the datastore while staying under the total host limit (128).
May be able to meet recommended cluster minimums for desired topology and resilience levels. (ex: Stretched clusters using secondary levels of resilience through RAID-6)	
Higher aggregate performance across the cluster because of a higher concentration of workloads (and their working sets) on any given host.	
More hosts mean a higher allowed component count for the cluster. This may be especially beneficial for data that is generating a lot of components. **A vSAN storage cluster with a higher host count will be able to support more VMs/data stored on the datastore.**	

"Capacity density" simply represents the number of storage devices in a storage node multiplied by the amount of capacity for each device. While "*hosts with higher density of resources per host"* and *"hosts with lower density of resources per host*" are not specifically defined, the comparison above will help provide some general understanding of advantages and disadvantages of the two approaches if one chooses to deviate from the ReadyNode Sizer results.

vSAN HCI and vSAN storage clusters store data in the form of objects. These objects are comprised of shards of data distributed across the cluster, and are known as components. vSAN ESA (powering both vSAN HCI and vSAN storage clusters) has a limit of 27,000 components per host, which translates to approximately 500 VMs per host (the current supported maximum for vSAN ESA). Some vSAN ReadyNode certified for vSAN storage clusters that have high capacities may approach these limits if the data stored is comprised of many objects, and thus many components. **Adding more hosts to a vSAN storage cluster will help alleviate component exhaustion, and effectively offer support for more VMs.** For more information on objects and components, see the post "vSAN Objects and Components Revisited."

While 500 VMs per host is the supported maximum for vSAN ESA powered clusters (vSAN HCI and vSAN storage clusters), **a design exercise should not use a maximum supported configuration, nor the total number of hosts in a vSAN storage cluster as design inputs.** Just as one does with designing for available memory and compute resources, one must also account for ensuring that component limits are not met during host failure scenarios. For example, a 20 host vSAN storage cluster could theoretically support up to 10,000 VMs, one should account for the total host count if there were a single or a double host failure. In this example, using the host count of "18" (or fewer) instead of "20" paired with using a maximum VM count per vSAN storage cluster host of 400 instead of 500 may be the better approach. With the same 20 host cluster, this would mean the cluster could support up to 7,200 VMs and still retain enough free components in the event of a double host failure.

Understand the difference in endurance for storage devices in vSAN storage clusters ReadyNodes

If vSAN storage clusters ReadyNodes provide options with storage devices that advertise multiple endurance ratings, selection of devices with the most appropriate endurance rating for your environment should be a part of your design. While vSAN 8 U2 and later will include a Skyline Health finding that will monitor endurance of these storage devices, your workloads and environment may be best suited for one endurance rating over the other. The vSAN ReadyNode sizer can help account for this design decision. See the post: "Expanded Hardware Compatibility for vSAN Express Storage Architecture" for more information.

Ensure all vSAN ReadyNodes used for vSAN storage clusters include a Trusted Platform Module (TPM) device.

This will ensure that the keys issued to the hosts in a vSAN storage cluster using Data-at-Rest Encryption are cryptographically stored on a TPM. This will guarantee that the host will have the required keys during host restarts even if the key provider is unavailable. If you are not planning to use vSAN Encryption services, including TPMs in the hosts at the time of purchase is an

affordable and prudent step to future configuration options. For more information, see the "Key Persistence" topic in the vSAN Encryption Services document.

Use the vSAN ReadyNode Sizer™ to meet capacity requirements.
The vSAN ReadyNode Sizer can help you to properly compose a storage cluster solution to meet all your requirements. It will produce the calculations necessary to account for capacity overheads for the ESA to make the sizing process easy and predictable. The ReadyNode sizer can estimate with reasonable levels of accuracy what you may see in real world scenarios.

Much like vSAN HCI cluster sizing, vSAN storage clusters will provide and advertise its capacity in raw form, where the total capacity advertised by a cluster is the aggregate total of all the storage devices claimed by vSAN ESA. Since different levels of resilience can be applied using storage policies, this means that the amount of data consumed on the datastore will be based on the storage policy, and other overheads. For example, a 100GB virtual disk with an assigned storage policy of FTT=2 using RAID-6 will consume about 150GB of raw capacity in the cluster. Since we recommend FTT=2 using RAID-6 for all vSAN storage clusters, the resilience overheads should be easier to estimate since this aspect of the data overheads remains consistent. See the posts "Demystifying Capacity Reporting in vSAN" and "Capacity Overheads for the ESA in vSAN 8" and "Improved Capacity Reporting in VMware Cloud Foundation 5.1 and vSAN 8 U2" for more information.

Note that **a vSAN storage cluster is primarily intended for processing and storing data.** There may be system instantiated VMs like vSphere Cluster Services (vCLS) VMs and agent VMs used to power protocol service containers for vSAN File Services. There is no software restrictions that prevents VM instances from running on a vSAN storage cluster, but vSAN storage clusters are specifically tuned for the task of processing and storing data. Running a large number of VM instances on vSAN storage clusters may conflict with your design goals. If one would like to run guest VM workloads while providing storage to another vSAN cluster, one can use a vSAN HCI with datastore sharing.

Design the vSAN storage cluster with the intention of resource symmetry across hosts in the vSAN storage cluster.
While vSAN HCI and vSAN storage clusters can accommodate for asymmetry host configurations, **the best cluster designs should strive for uniformity of all resources**, including CPU, memory, network, and storage capacity across each host that comprises the vSAN storage cluster. vSAN storage clusters should strive for consistent hardware resources across the cluster. Mixing ReadyNode Profiles certified for vSAN storage clusters in the same cluster is not advised, or supported. See the post, "Asymmetrical vSAN Clusters – What is Allowed, and What is Smart" for more information.

Cluster Design and Sizing

vSAN storage cluster single site cluster must consist of at least 4 hosts.
vSAN storage clusters with as few as 4 hosts are now allowed. This will better accommodate smaller environments, and allow the cluster to use space-efficient RAID-5 erasure coding, yet still have enough hosts to regain the prescribed level of resilience in the event of a sustained host failure.

Recommendation: Do not enable the "Host Rebuild Reserve" capacity management mechanism in a vSAN storage cluster that consist of only 4 hosts. When paired with the Auto-Policy Management feature, this will prevent vSAN storage clusters from being able to use space-efficient RAID-5 erasure coding in this configuration. See the post: "Auto-Policy Management Capabilities with the ESA in vSAN 8 U1" for more details.

vSAN storage clusters consisting of 7 or more hosts will provide two benefits.

- **Optimal Resilience.** This will ensure that the cluster can support FTT=2 using RAID-6. While FTT=2 using RAID-6 only requires 6 hosts, a sustained failure of a host in a cluster consisting of just 6 hosts would result in an insufficient number of hosts to regain its prescribed level of resilience. 7 hosts would be able to automatically regain its prescribed level of resilience upon a sustained host failure. 7 hosts are also the minimum required host count for vSAN's Auto-Policy Management feature to use RAID-6 when the cluster capacity management setting of Host Rebuild Reserve is enabled. For more information, see the post "Auto-Policy Management Capabilities with the ESA in vSAN 8 U1."

- **Reduce percentage of impact with a sustained host failure.** As shown in the illustration below, percentage of impact of a host failure becomes much smaller as the cluster host count increases. As illustrated in the graph below, a cluster with a minimum of 7 hosts would impact no more than about 14% of the storage resources (capacity and performance) in the event of a sustained host failure. Increase the host count reduces the percentage of impact even more – as low as about 4% for a 24 host vSAN storage cluster.

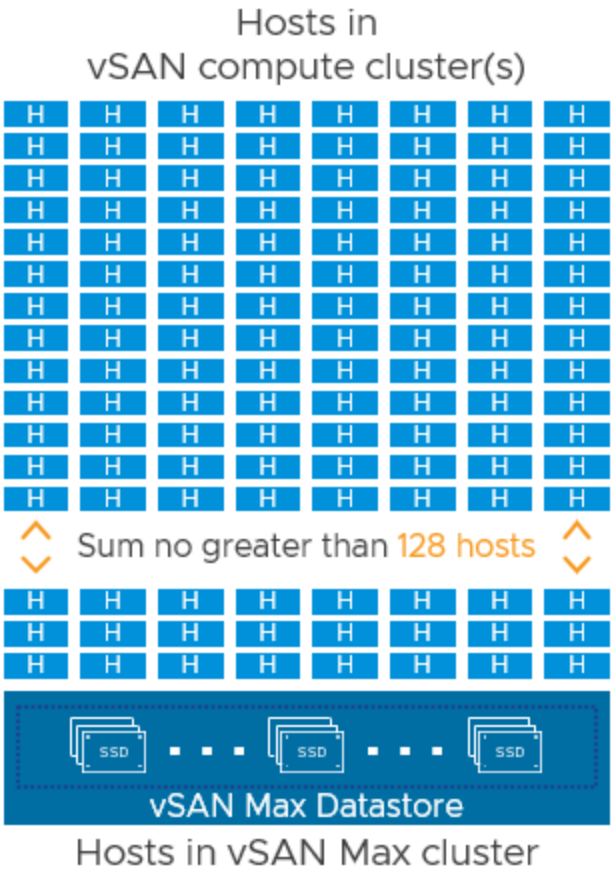

Figure. Recommended host count for a vSAN storage cluster.

vSAN storage clusters in a stretched cluster should consist of a minimum of 8 data hosts across two data sites.
A stretched cluster with at least 4 hosts in each of the two data sites would ensure the cluster could support a storage policy with a secondary level of resilience of FTT=1 using RAID-5 erasure coding, while being able to regain its prescribed level of resilience in the event of a sustained host failure in each one of the sites.

A stretched cluster with at least 7 hosts in each of the two data sites would ensure the cluster could support a storage policy with a secondary level of resilience of FTT=2 using RAID-6 erasure coding, and allow vSAN to regain its prescribed level of resilience in the event of a sustained host outage. An additional host beyond the minimum required allows vSAN to reconstruct the stripe and parity in the most efficient way, with the fewest performance implications. See the post "Using the vSAN ESA in a Stretched Cluster Topology" for more information. To ensure that the network connectivity between sites will meet the requirements of the workloads, see the "vSAN Stretched Cluster Bandwidth Sizing" document.

For all cluster topology types, the recommended maximum cluster size for a vSAN storage cluster is 24 hosts.
With the initial release of vSAN Max, VMware recommends a **maximum of 24 hosts in a cluster, but 32 hosts in a vSAN Max cluster is the supported limit.** The host count for the vSAN Max cluster and any vSAN compute clusters cannot exceed 128 hosts in total, as shown below.

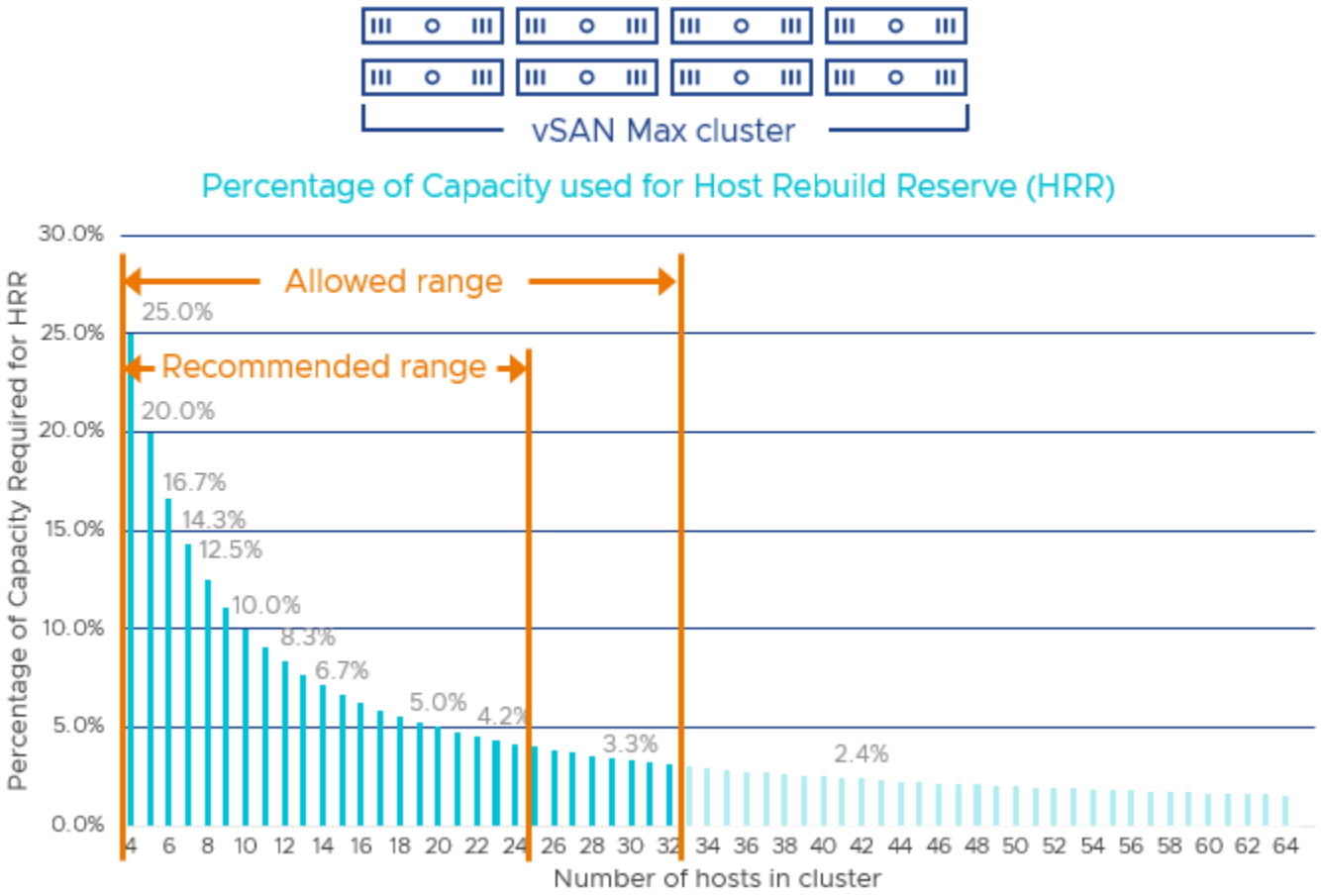

Figure. The maximum number of hosts participating in a vSAN Max cluster and client clusters.

Limiting the vSAN storage cluster size to 24 hosts will allow for up to 104 hosts from vSAN compute clusters to mount the datastore. A vSAN storage cluster size of 32 hosts would allow for up to 96 hosts from vSAN compute clusters to mount the datastore, still offering a very good compute to storage ratio of 3:1.

You may find that limiting vSAN storage cluster sizes to no more than 16 hosts make the most sense. 16 host is the approximate maximum number of 2U servers that can fit in a rack. Keeping the storage cluster to a single rack will allow backend storage traffic to be serviced by the top of rack (ToR) switches, and will minimize the amount of traffic traversing the network spine.

Consider your overall capacity needs when determining initial cluster size.
vSAN storage clusters provides tremendous flexibility in incremental scaling of capacity and performance through simply adding more hosts to an existing cluster. Consider your overall capacity requirements and forecasts when determining the ideal host count in a cluster. Recent updates allows for more flexibility with hardware and host counts in a vSAN storage cluster. See the post: "Greater Flexibility with vSAN Max through Lower Hardware and Cluster Requirements" for more information.

For example, in a single site environment, if you anticipate needing 4 PB of raw capacity immediately, and 4 additional PB in the next 18 months, consider creating vSAN storage cluster with 12 hosts to address the initial need, and a second vSAN storage cluster with 12 hosts for the additional expansion. This would allow for each respective cluster to easily grow by adding hosts because it is well under the recommended host count maximum for the vSAN storage cluster.

This approach would also allow more client clusters to mount the respective storage resources.

General client cluster compatibility considerations

vSphere clusters that wish to mount a datastore provided by vSAN storage clusters will have a thin layer of vSAN activated on the hosts to provide the connectivity. This step in the configuration of the client cluster is what makes a vSphere cluster a "vSAN compute cluster" as stated in the UI. Installation of the software is an automated one-time process that occurs on each hosts in a vSphere cluster.

vSAN HCI clusters can also act as a client cluster, connecting to a datastore provided by a vSAN storage cluster. At this time, only vSAN HCI clusters using the ESA can mount a datastore provided by vSAN storage clusters. This is because vSAN storage clusters is built using vSAN ESA.

Supported client cluster connectivity for vSAN storage clusters in a stretched topology.

There are certain limitations with client cluster connection types when mounting a datastore from a vSAN storage cluster or vSAN HCI cluster that is in a stretched cluster configuration. **In general, vSphere clusters stretched across two geographic sites is currently not supported when using a vSAN storage clusters or vSAN HCI cluster in a stretched cluster configuration.** For a detailed list of supported client configurations of vSAN storage clusters in a stretched topology, see the table below.

Client Cluster Type	Server Cluster Type	Supported	Notes
vSAN HCI clusters (ESA) in a stretched cluster configuration.	vSAN storage cluster or vSAN HCI cluster (ESA) in a stretched cluster configuration	Yes	Provides resilience of data and high availability of running VM instances.
vSAN HCI clusters (ESA) when it resides in one of the data sites where the vSAN storage cluster resides.	vSAN storage cluster or vSAN HCI cluster (ESA) in a stretched cluster configuration	Yes	Provides resilience of data but no high availability of running VM instances.
vSphere clusters stretched across two sites using asymmetrical* network connectivity.	vSAN storage cluster or vSAN HCI cluster (ESA) in a stretched cluster configuration	No	Not supported at this time.
vSphere clusters stretched across two sites using symmetrical* network connectivity.	vSAN storage cluster or vSAN HCI cluster (ESA) in a stretched cluster configuration	Yes	Supported, but less common, as it would require the same network capabilities (bandwidth and latency) between fault domains defining each site.
vSphere clusters when it resides in one of the data sites where the vSAN storage cluster resides.	vSAN storage cluster or vSAN HCI cluster (ESA) in a stretched cluster configuration	Yes	Provides resilience of data but no high availability of running VM instances.
Any client cluster running vSAN OSA	vSAN storage cluster or vSAN HCI cluster (ESA) in a single site or stretched cluster configuration	No	Not supported at this time.

Stretched Cluster Compatibility Matrix for vSAN storage clusters

*** Asymmetrical network connectivity would describe a topology where the network capabilities (latency & bandwidth) connecting the two sites (fault domains) would be less than the network capabilities between the client cluster and the server**

cluster within each site. This is most common with stretched cluster configurations using and inter-site link (ISL) between sites. **Symmetrical network connectivity would describe a topology where the network capabilities connecting the two sites would be the same as the network capabilities between the client cluster and server cluster within each site.** This configuration is less common, but might be found in environments where the two fault domains defining the sites in the stretched topology are simply server racks or rooms sitting adjacent to each other using the same network spine.

Understand cluster licensing

vSAN storage clusters require vSAN ReadyNodes certified specifically for vSAN Max. There will be no support for in-place upgrades/conversions from a vSAN HCI cluster deployment option to a vSAN storage cluster deployment option. See the document: "vSAN HCI or vSAN Max - Which Deployment Option is Right for You?" for more details.

Networking

Strive to use at least 25Gb networking between hosts that comprise the vSAN storage cluster.

The networking connectivity requirements between hosts that comprise a vSAN storage cluster will depend on the vSAN ReadyNodes uses for the storage cluster. For example, the networking requirements of a cluster using the entry-level "vSAN-Max-XS" ReadyNodes has a smaller network bandwidth requirement than the larger vSAN-Max-Med ReadyNodes. When possible, it will be best to specify faster networking for all connectivity between vSAN storage cluster hosts to ensure networking bandwidth between hosts in a vSAN storage cluster is not the bottleneck.

Even if a design and sizing exercise determines that 25Gb or even 100Gb is recommended for your vSAN storage cluster, this recommendation only applies to the hosts that make up the vSAN storage cluster. vSphere clusters that mount the datastore served by vSAN storage clusters do not need to meet this requirement. See the illustration below as an example.

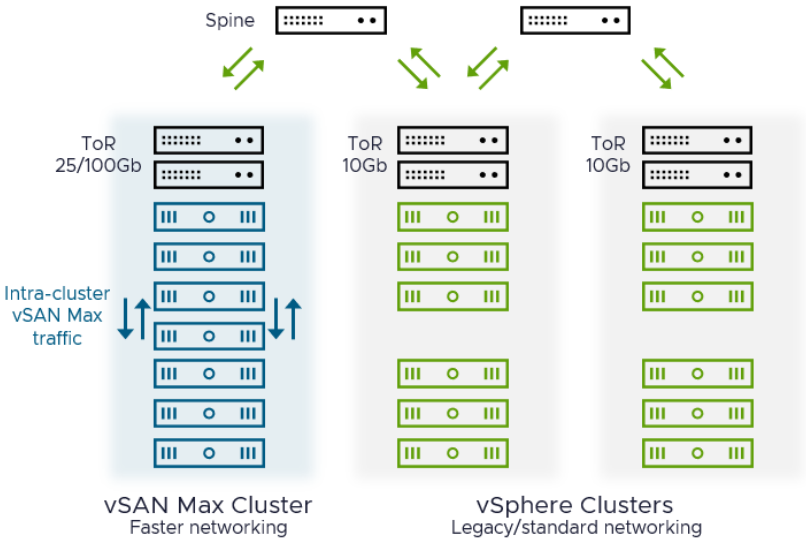

Figure. Discerning network requirements with vSAN storage cluster hosts versus client clusters.

See the post "Starting Small with vSAN Max" for more information.

Use reasonably fast networking from compute clusters to the vSAN storage cluster.

vSphere clusters that mount the datastore can do so with as little as 10Gb networking. This can be ideal for legacy environments comprised of vSphere clusters with older networking. Simply ensure that the vSAN Max cluster has met its networking requirements, which refers to the bandwidth needed for host-to-host communication within the vSAN Max cluster, While vSphere clusters can connect to a vSAN Max cluster with as little as 10Gb networking, using faster networking from the client clusters may provide better performance if those 10Gb links were a source of contention.

The mounting of a vSAN Max or vSAN HCI datastore will run a precheck, and look for network connectivity between the client clusters and the server cluster of 5ms or less. Since this connection represents storage traffic, it is best to strive for 1ms or less for network connectivity between client clusters and a server cluster.

RDMA is supported in vSAN HCI clusters powered by the ESA, but it is not supported in disaggregated environments at this time.

Understand configuration options in the hypervisor for networking between vSAN Max hosts within a cluster

ReadyNodes that comprise a cluster can support several different network configuration types, as long as they meet the minimum requirements assigned to the ReadyNode profile. Below are some examples

- **Single vSAN VMkernel port using Active/Standby configuration.** This configuration is the **preferred configuration for all vSAN cluster deployments**, and typically uses two or more uplinks in a VMkernel port, where one uplink configured as "Active" and the other(s) are configured as "Standby." This approach is extremely simple to configure and maintain, but will only use one uplink for vSAN traffic. Typically the uplink assigned as "Standby" with this VMkernel port will be assigned as "Active" with some other VMkernel port providing other services, such as vMotion so that links are utilized efficiently under normal operating conditions.
- **Single vSAN VMkernel port with two active uplinks using Load Based Teaming (LBT).** This would choose uplink using "Route based on Physical NIC load. The benefits to this are relatively minor, and can be problematic when providing a deterministic path for high levels of consistent storage performance. While it is currently the default for VCF**, it is not recommended for vSAN HCI or vSAN storage clusters.** You can change the VMkernel port tagged for vSAN to an Active/Standby arrangement described above without issue.
- **Single vSAN VMkernel port using Link Aggregation (LACP).** This configuration will use two or more uplinks paired with advanced assist with balancing multiple network connection sessions across the links. This may provide some levels of improved throughput, but requires configuration on the network switches and the host to operate properly. It is not as commonly used as the options above, and **not supported** as an option when using VMware Cloud Foundation.

Note that in VCF 5.2, it will set the teaming policy for vSAN traffic to active/active using LBT. Any VMkernel port tagged for vSAN traffic should be changed to active/standby using “Route based on originating virtual port ID” for optimal performance and consistency. This is a fully supported configuration change in VCF.

For more information on minimum requirements of ReadyNodes certified for vSAN storage clusters, see the post, "ReadyNode Profiles Certified for vSAN Max."

Recommendation: As with other vSphere networking, make sure that the uplinks assigned to vSwitches and VMkernel portgroups are come from different NICs in the host. This ensures that if there is a NIC failure, that the services can fail over to the other NIC(s) available on the host.

Understand how topologies can change traffic in a spine and leaf network.

With smaller vSAN HCI clusters, vSAN storage traffic typically stayed within the Top-of-Rack (ToR) leaf switches. With larger vSAN HCI clusters, clusters using vSAN’s Fault Domains feature, or vSAN HCI clusters using cluster-capacity sharing, this traffic will traverse across the network spine.

Depending on the size of the environment, vSAN storage clusters may also affect how traffic traverses across a network. For example, several other vSphere clusters residing in other racks will travers the network spine to connect to the vSAN storage cluster providing storage resources. Ensuring sufficient resources at the network spine will improve performance and resilience.

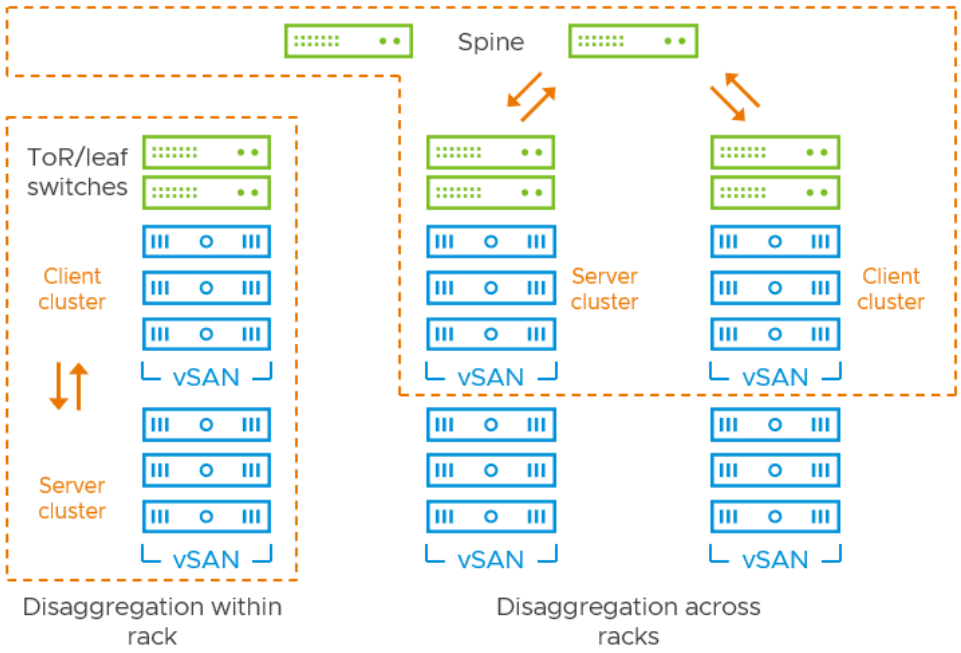

Figure. How topologies may affect network traffic.

Maintain a 1:1 oversubscription ratio in your network spine-leaf topology

Network "oversubscription" refers to the amount of bandwidth services by leaf switches (at the ToR) to the amount of bandwidth provided by the spine switches. This is expressed as a ratio (e.g. 1:1, 2:1, etc.). A lower ratio represents a less constrained spine, and will offer the most predictable storage performance when storage is traversing the spine.

Ensure proper networking connectivity between vSAN compute clusters and vSAN storage clusters.

Since the communication between compute clusters and vSAN storage clusters is latency sensitive storage traffic, we recommend simplified network connectivity between clusters. **This means avoiding Firewalls and IDS/IPS systems that may inadvertently block this mission critical storage I/O in a manner that could cause substantial disruption.** Network overlays are supported, but if one runs vSAN traffic through a VDS that is managed by NSX-T, use VLAN-backed port groups to prevent a loss of access to the host or the availability of VMs. For more information, see the video "vSAN Quick Questions – Can I run vSAN traffic through a network overlay, Firewall, IDS or NSX?"

With any type of storage traffic, redundancy of connectivity from end to end is important to ensure I/O is transmitted in a timely and reliable manner even in the event of a single network connection failure. While teaming multiple NICs in a host is common practice for all vSphere and vSAN environments, ensuring redundancy from the hosts in the server cluster (vSAN storage cluster) to the client clusters (vSphere clusters) and the connecting switch fabric will provide a robust environment.

Day-0 Initial Deployment and Configuration

Cluster Services Configuration

Preparing the vSAN storage cluster for its initial configuration.

Even though a vSAN storage cluster will typically not host any user-created guest VMs, some vSphere configuration settings are necessary for proper functionality. Prior to initiating any vSAN storage cluster configuration workflow, please ensure the following:

- **Use vDS in cluster configuration.** Ensure that virtual Distributed Switches (vDS) are used with all relevant VMkernel ports configured in the cluster. vDS functionality is available as a part of vSphere. Recommendations on network configuration choices such as NIC teaming generally align with guidance provided for vSAN clusters. The vSAN

Design Guide has a “Network Design Considerations” section that provides an overview of recommendations, with more extensive information provided in the vSAN Network Design Guide.

- **Ensure that DRS and HA are configured.** These services are available as a part of vSphere.
- **Ensure that a vMotion is configured.** The configuration of VMkernel ports with vMotion traffic tagged will help ensure mobility of management VMs.

Configure a new cluster as a vSAN storage cluster.
Configuring a new cluster to serve as a vSAN storage cluster is easy. Simply create a new cluster and name it as desired to complete the workflow. Do not enable vSAN in this initial workflow. Once the cluster is created, highlight the cluster, and click Configure > vSAN > Services. You will be presented with three options.

- **vSAN HCI.** This creates a traditional vSAN HCI cluster.
- **vSAN Compute Cluster.** This creates a vSphere cluster that can be used to connect to a vSAN storage cluster.
- **vSAN storage cluster.** This creates a vSAN storage cluster.

Simply select “vSAN storage cluster” and choose if you want it to be a single site vSAN storage cluster, or a stretched vSAN storage cluster, as shown below.

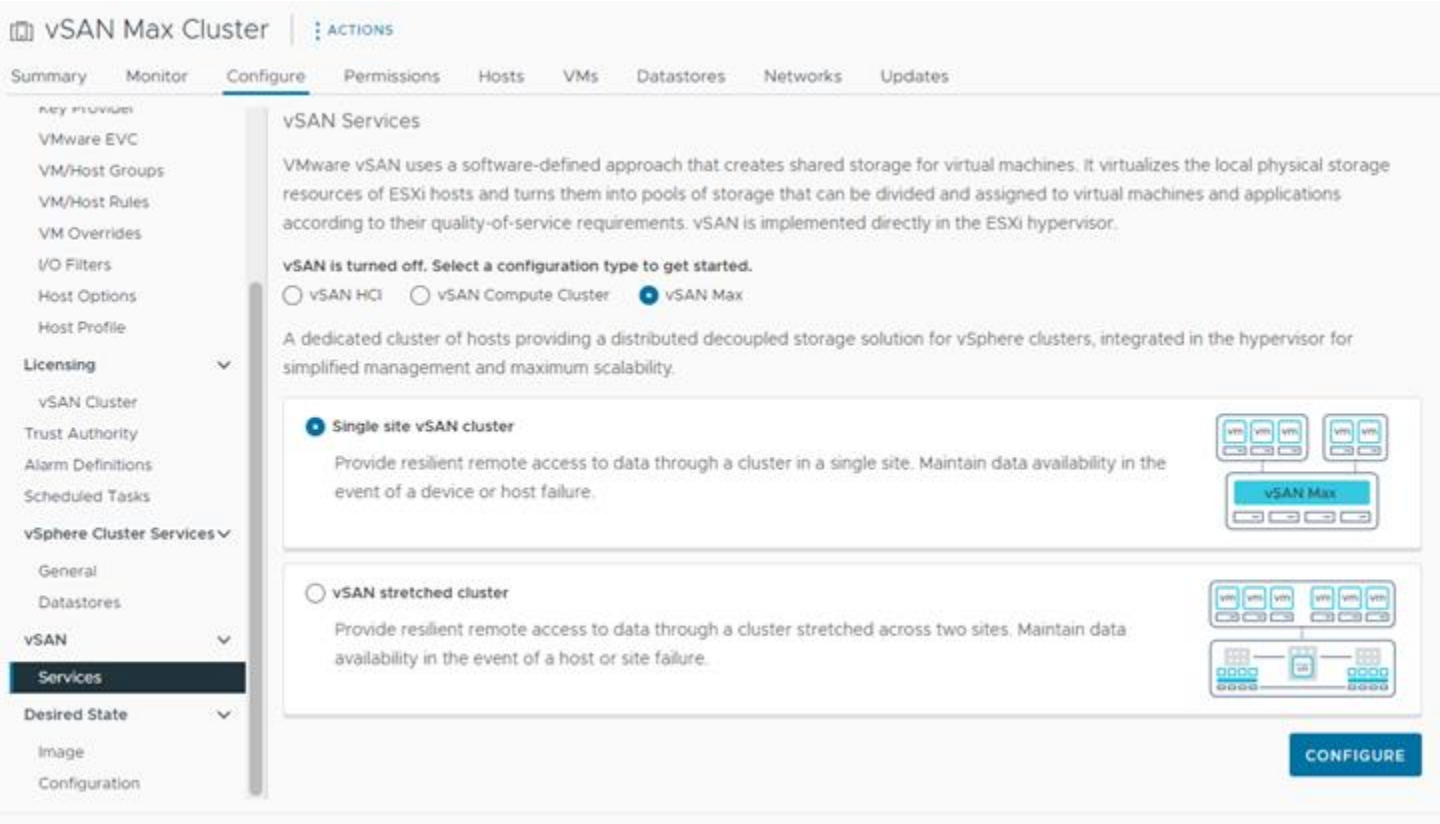

Figure. Initial configuration of a vSAN storage cluster.

Proceed with completing the workflow, which will present a few more options, such as Encryption services, and Auto-Policy management.

If desired, enable the vSAN “Operations Reserve” and “Host Rebuild Reserve” toggles for single site vSAN storage clusters.
When enabled, this helps ensure there is sufficient free space in the cluster for internal operations and to rebuild data in the event of a sustained host failure. Note that when Host Rebuild Reserve is enabled, and paired with the Auto-Policy Management feature, it will require one additional host beyond the absolute minimum required by the storage policy. This is

why we recommend 7 hosts at minimum for a single site vSAN cluster, where data can be stored using the highly resilient and space efficient FTT=2 using RAID-6, while still having a spare fault domain to regain prescribed levels of resilience in the event of a sustained host failure. See the post "Understanding 'Reserved Capacity' Concepts in vSAN" for more information. It is perfectly acceptable to not use the "Operations Reserve" and the "Host Rebuild Reserve" feature if you choose.

Note that the "Operations Reserve" and "Host Rebuild Reserve" toggles **cannot be enabled** when configuring vSAN storage clusters in a stretched topology, or when using the vSAN Fault Domains feature.

Recommendation: Do not enable the "Host Rebuild Reserve" capacity management mechanism in a vSAN storage cluster that consist of only 4 hosts. When paired with the Auto-Policy Management feature, this will prevent vSAN storage cluster from being able to use space-efficient RAID-5 erasure coding in this configuration. See the post: "Auto-Policy Management Capabilities with the ESA in vSAN 8 U1" for more details.

The Operations Reserve and Host Rebuild Reserve toggles can be enabled by highlighting the cluster, clicking **Configure > vSAN > Services > Reservations and Alerts** as shown in the image below.

Enabling the Host Rebuild Reserve and Operations Reserve Toggles in vSAN

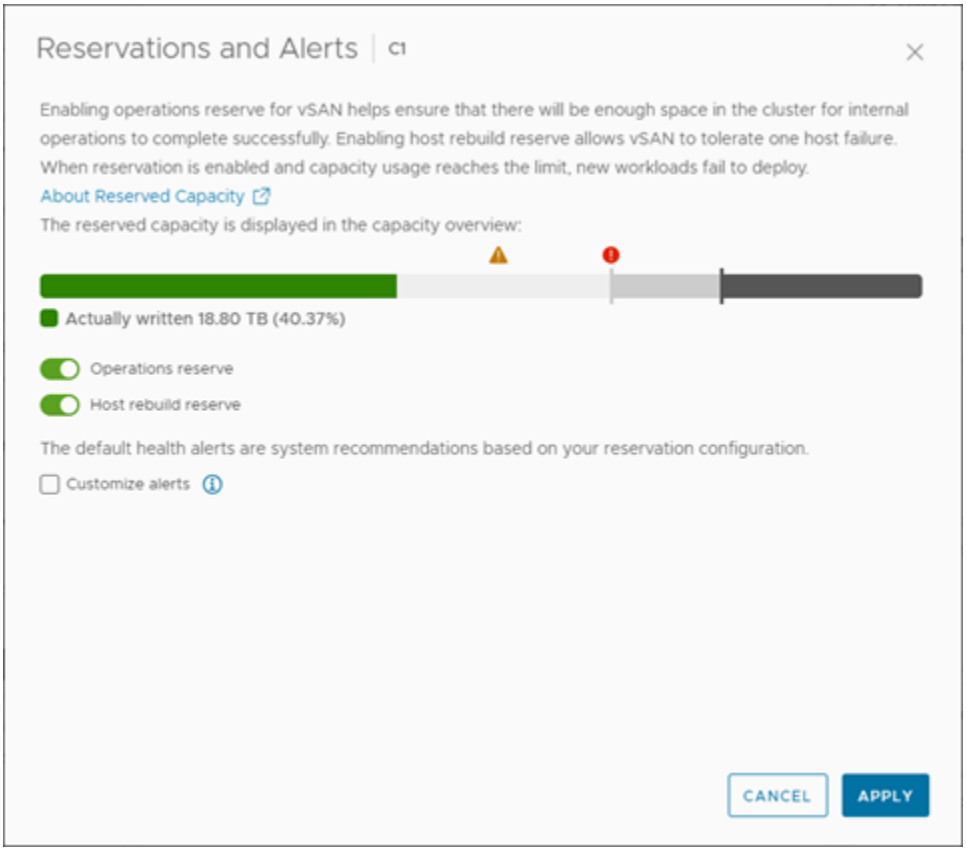

Figure. Enabling the Host Rebuild Reserve and Operations Reserve Toggles in vSAN.

Enable the vSAN "Auto-Policy" management feature on all topology types when using vSAN storage clusters.
This will ensure optimal levels of resilience and space efficiency for data stored on a vSAN storage cluster. A **cluster-specific default storage policy** will be created and tuned for the cluster based on the host count, topology type (ex: single site, stretched, vSAN Fault Domains), and if the Host Rebuild Reserve is enabled or not. For stretched clusters, Auto-Policy Management will also ensure that a secondary level of resilience is applied to the default storage policy, improving resilience. Be aware that Auto-Policy Management may suggest a new storage policy rule setting that may impact many VMs. Skyline Health for vSAN will provide a health finding that will assist in the change of this default storage policy, and vSAN will manage the rate that the VM objects will be changed to the new policy setting.

See the post "Auto-Policy Management Capabilities with the ESA in vSAN 8 U1" for more information.

Recommendation: Do not enable the "Host Rebuild Reserve" capacity management mechanism in a vSAN storage cluster that consist of only 4 hosts. When paired with the Auto-Policy Management feature, this will prevent vSAN storage clusters from being able to use space-efficient RAID-5 erasure coding in this configuration. See the post: "Auto-Policy Management Capabilities with the ESA in vSAN 8 U1" for more details.

The Auto-Policy Management feature can be enabled by highlighting the cluster, clicking **Configure > vSAN > Services > Storage > EDIT** as shown in the image below.

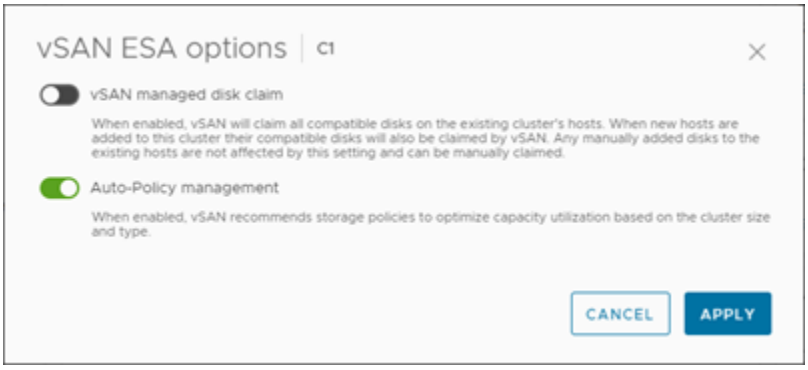

Figure. Enabling Auto-Policy management in vSAN storage clusters.

Note that if vSAN storage clusters are used across multiple vCenter Server instances, where the client cluster managed by a different vCenter Server than the vSAN storage cluster, the object's storage policy assignment is controlled by the vCenter Server the object is managed from (e.g. the compute cluster). Therefore, **the cluster-specific storage policy created by Auto-Policy Management will not be available for use in this circumstance.** See the Storage Polices section for more information on this topic.

Enable the "Automatic Rebalance" cluster setting on all topology types when using vSAN storage clusters.
This toggle will tell vSAN to rebalance data to reasonable levels of symmetry if a host or device if capacity disparities exceed its thresholds. A more evenly balanced distributed storage system like vSAN storage clusters will perform more consistently when resources are consumed in a balanced manner. See the post "Should Automatic Rebalancing be Enabled in a vSAN Cluster?" for more information.

The Automatic Rebalance feature can be enabled by highlighting the cluster, **clicking Configure > vSAN > Services > Advanced Options > EDIT** as shown in the image below.

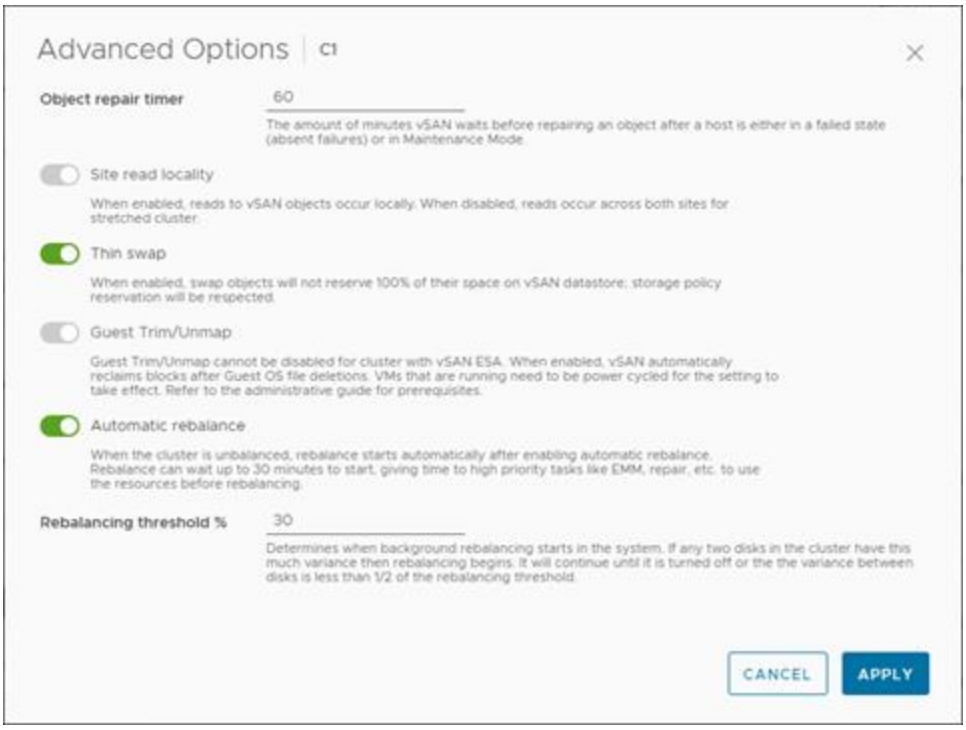

Figure. Enabling Automatic rebalance in vSAN storage clusters.

Ensure the Customer Experience Improvement Program (CEIP) is enabled.

The CEIP enables VMware to provide additional benefits to its customers through anonymized telemetry data. While it has been enabled by default for many versions of vSphere, double checking that this is enabled will help VMware provide the highest levels of product support. See the document "vSAN Support Insight" for more information.

The status of the CEIP can be viewed in the vSphere Client by clicking on **Administration > Deployment > Customer Experience Improvement Program**. Skyline Health for vSAN will also produce a health finding alert if it is not enabled on the vCenter Server instance, as shown below.

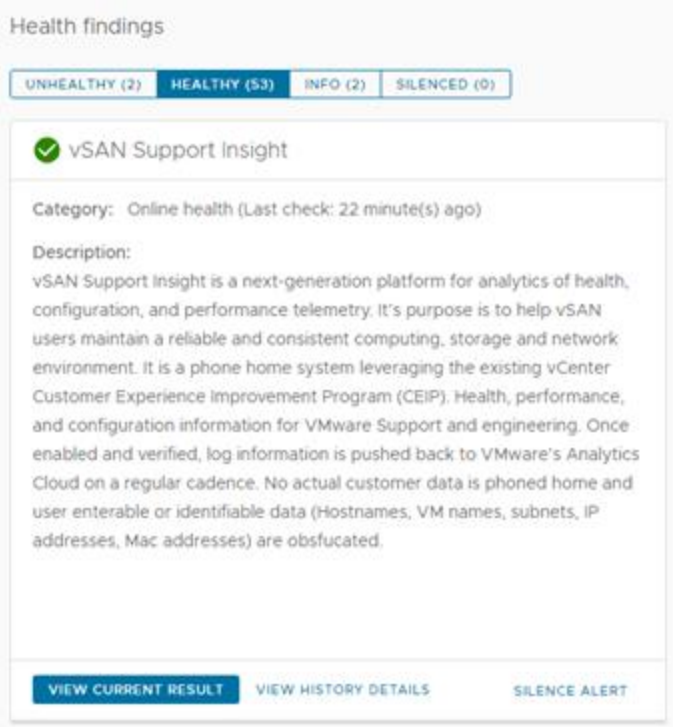

Figure. CEIP enabled verification using the vSAN Support Insight health check.

Client clusters connecting to vSAN storage clusters

Creating a vSAN Compute cluster.

A "vSAN compute cluster" is simply a **vSphere cluster that has a thin layer of vSAN activated** for the purposes of mounting the remote vSAN storage cluster datastore. Once a vSphere cluster is created, highlight the cluster, and click **Configure > vSAN > Services**. You will be presented with three options.

Select "vSAN Compute Cluster" and complete the workflow, as shown below.

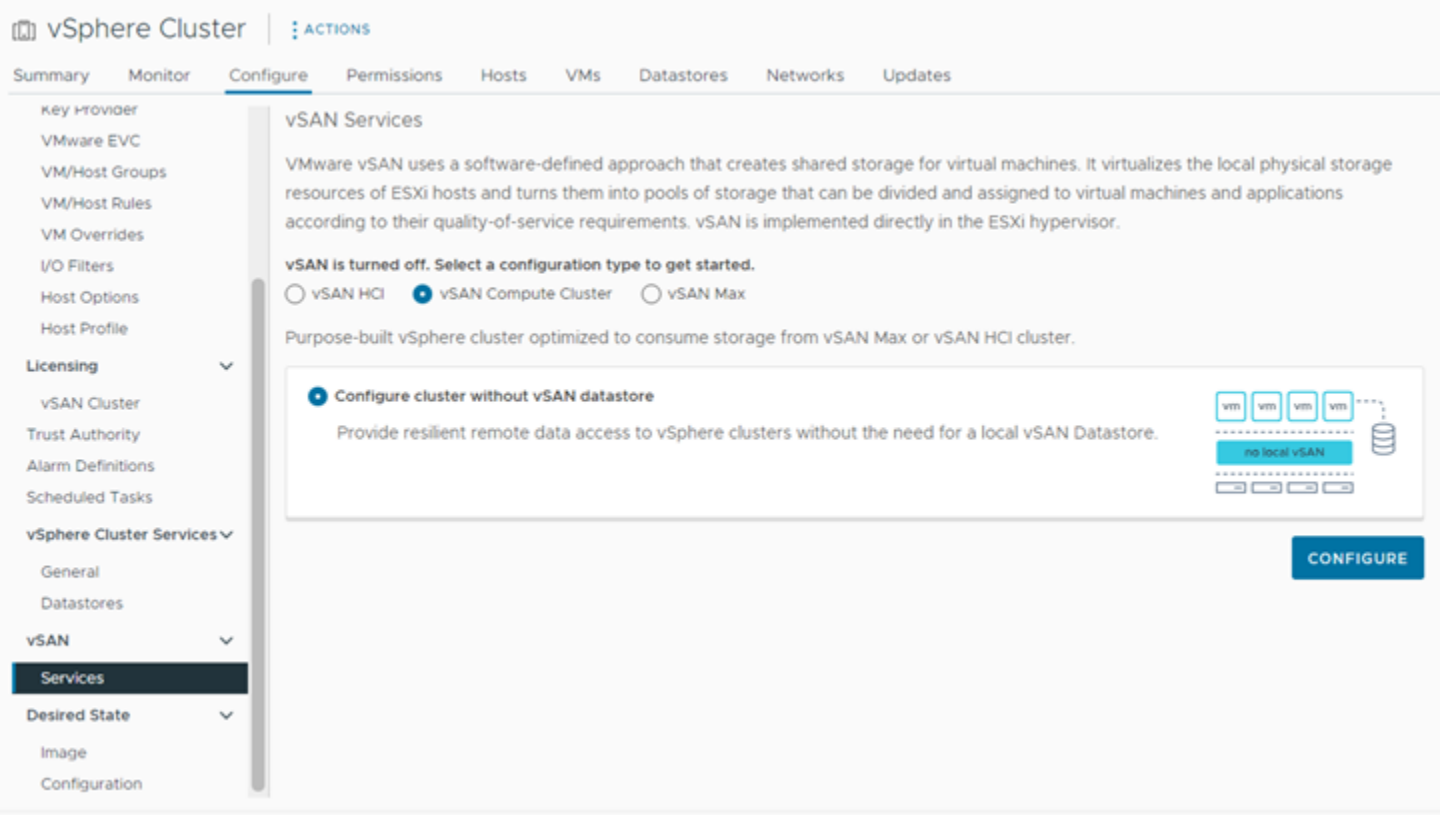

Figure. Enabling connectivity from a vSAN storage cluster to a vSphere cluster.

The hosts in a vSphere cluster attempting to mount a vSAN storage cluster datastore must be running vSphere 8 or later.

Mount a datastore from a vSAN storage cluster to a vSAN Compute Cluster.

Once a vSphere cluster is configured as a vSAN compute cluster as shown above, one can easily mount the remote datastore provided by the vSAN storage cluster. One can highlight the vSphere cluster, click **Configure > vSAN > Services > Mount Remote Datastores** as shown below, or find the same ability to mount the remote datastores in **Configure > vSAN > Services > Remote Datastores.**

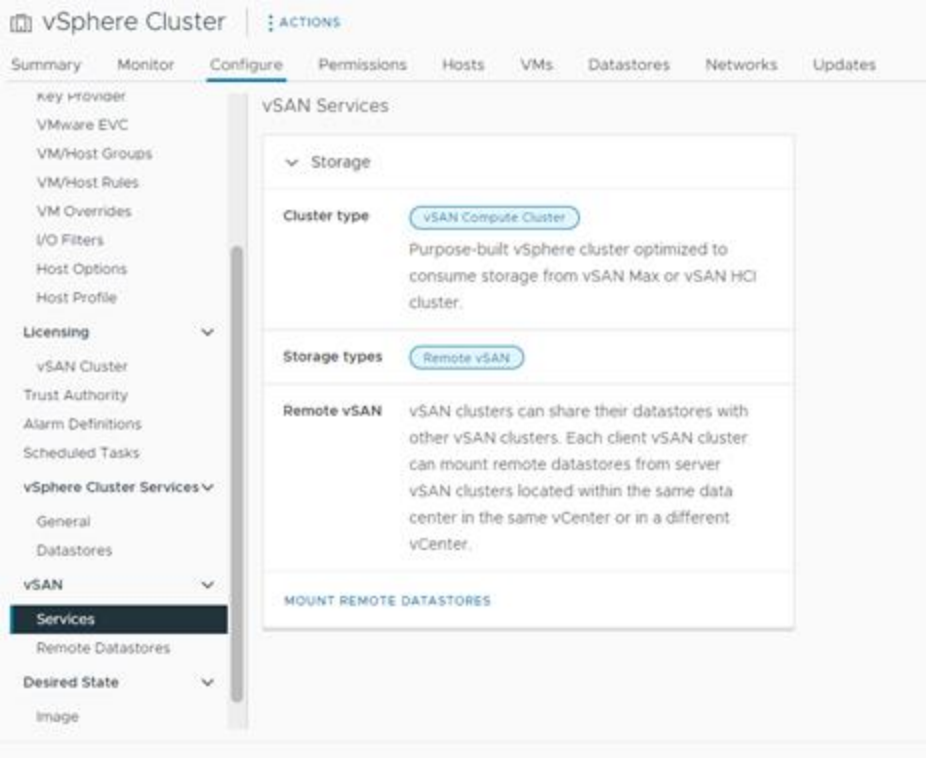

Figure. Mounting a datastore provided by vSAN storage cluster.

It will then present the available vSAN storage cluster datastore(s) that are eligible to mount. Select the desired remote datastore, and click "Next." A compatibility check will be performed before the workflow completes.

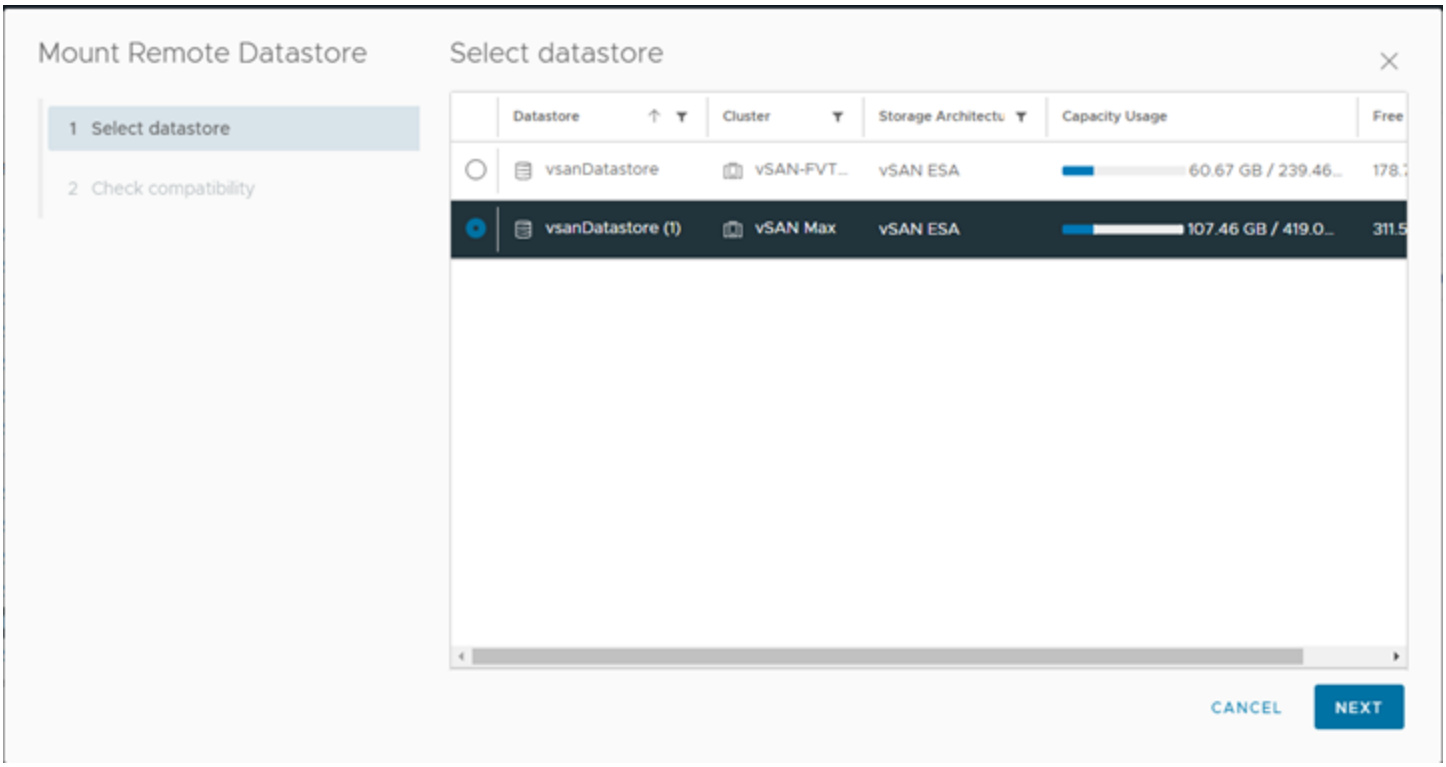

Figure. Selecting the desired vSAN storage cluster datastore for use with a vSphere cluster.

The datastore is now ready for use.

Be aware of the client cluster count limit for a vSAN storage cluster datastore.

Design your vSAN storage cluster with the knowledge that the maximum number of client clusters is 10, as shown below. Client clusters can be vSphere clusters, also known in this context as "vSAN Compute clusters," and vSAN HCI clusters.

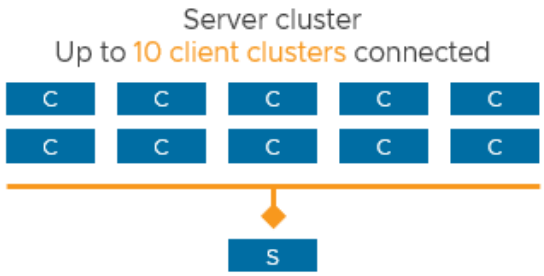

Figure. Maximum number of client clusters that can connect to a vSAN storage cluster.

The number of client clusters connecting to a vSAN storage cluster datastore may not exceed 10. This limit is tightly coupled with the total host count participating in a vSAN storage cluster any vSAN compute clusters. This limit is 128 hosts in total, as described earlier in this document.

A Few reminders about cluster types:

- **vSphere clusters.** These provide a collection of cluster-specific resources, such as compute and memory for VMs. Storage resources are provided by an external shared storage solution.
- **vSAN HCI clusters.** These provide a collection of cluster-specific resources, such as compute, memory, and storage for VMs. When one is not using any type of cluster capacity sharing capability within vSAN, the storage is treated as an exclusive resource of the cluster.
- **vSAN storage clusters.** These provide a collection of storage resources for VMs residing in other clusters, acting as a shared storage solution for vSphere clusters, and even vSAN HCI clusters.

vSphere clusters can be comprised of between 1 and 96 hosts. Since vSAN storage clusters disaggregates, or decouples storage from compute resources, one can create vSphere clusters a size that best meets the needs of the organization, and design the compute clusters to reflect the computational requirements of the applications, leaving the storage responsibilities up to vSAN storage clusters.

Design of compute clusters connected to vSAN storage clusters is no different than designing vSphere clusters using another external shared storage solution. vSphere cluster design is a lengthy topic with considerations outside of the scope of this document, many of those same principles apply.

Ensure proper APD failure response in vSphere HA configuration.

Any vSphere cluster and acting as a "client cluster" that mounts a datastore served by a vSAN storage cluster must have the proper response to an "All Paths Down" (APD) failure. **When enabled and configured correctly on the client vSphere cluster**, the isolation events related to the connectivity between the client and the server cluster, or within the client cluster will result in the VM on the client cluster being terminated and restarted by HA. The APD failure response can be set to "Power off and restart VMs -– Aggressive restart policy" or "Power off and restart VMs -– Conservative restart policy." This HA cluster setting is not required for vSAN clusters that do not participate in vSAN's disaggregation offerings.

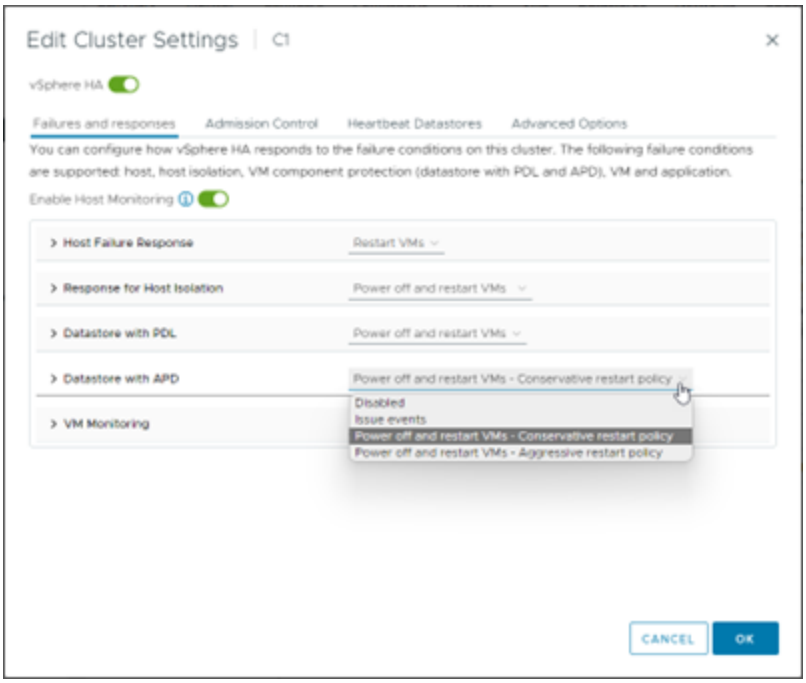

Figure. Configuring HA for a vSphere cluster using a vSAN storage cluster datastore.

While a vSAN storage cluster uses the vSAN network for HA heartbeats, the connecting compute clusters continue to use the vSphere management network for HA heartbeats, and not the vSAN network configured on the compute cluster.

Storage Policies

vSAN storage clusters should use storage policies that provide the highest levels of space-efficient resilience.
The recommendations below show the minimum number of hosts in the cluster to support those resilience levels. See the "Cluster Sizing Guidance" section in this document for more guidance on recommended cluster sizes for vSAN storage clusters.

- **Single site cluster consisting of 4 to 5 hosts,** Use a storage policy with FTT=1 using RAID-5. This recommendation also applies to clusters using the Fault domains feature where there are 4-5 fault domains.
- **Single site cluster consisting of 6 or more hosts.** Use a storage policy with FTT=2 using RAID-6. This recommendation also applies to clusters using the Fault domains feature where there are 6 or more fault domains.
- **Stretched clusters with 8-10 data hosts.** Use a storage policy that provides site mirroring for site-level resilience, paired with FTT=1 using RAID-5 for a secondary level of resilience.
- **Stretched clusters with 12 or more data hosts.** Use a storage policy that provides site mirroring for site-level resilience, paired with FTT=2 using RAID-6 for a secondary level of resilience.

Enabling Auto-Policy Management in the cluster will ensure that the default storage policy is automatically configured using the highest-level space efficient resilience possible for the cluster. For more information, see the post "Auto-Policy Management Capabilities with the ESA in vSAN 8 U1."

Recommendation: Do not enable the "Host Rebuild Reserve" capacity management mechanism in a vSAN storage cluster that consist of only 4 hosts. When paired with the Auto-Policy Management feature, this will prevent vSAN storage clusters from being able to use space-efficient RAID-5 erasure coding in this configuration. See the post: "Auto-Policy Management Capabilities with the ESA in vSAN 8 U1" for more details.

Do not use RAID-1 mirroring.

When using the vSAN ESA in vSAN HCI and vSAN storage clusters, RAID-6 erasure coding is faster than RAID-1 mirroring. RAID-1 mirroring will consume much more capacity than RAID-6. The only time RAID-1 is an acceptable option is with stretched clusters, where it is employed through a storage policy to mirror data across sites, while providing secondary levels using RAID-6.

Leave compression enabled in all vSAN storage clusters.

vSAN storage clusters use the ESA's compression mechanism, which is controlled by storage policy. Leaving compression enabled will yield additional capacity efficiency while having little to no impact on storage performance. See the post "Using the vSAN ESA in a Stretched Cluster Topology" for a better understanding of why compression should remain enabled. Compression also has a positive benefit on storage overhead considerations. See the post: "Improved Capacity Reporting in VMware Cloud Foundation 5.1 and vSAN 8 U2" for more information.

Understand storage policy behavior of VMs in vSAN compute clusters managed by different vCenter Server than vSAN storage cluster.

Storage policies are a construct of a vCenter Server instance. Currently there is not a way to provide storage policy management across multiple vCenter Server instances. When a VM on the client cluster managed by one vCenter Server is using the storage on a server cluster managed by a different vCenter Server, **only one SPBM policy will take effect: The policy that is being used by the vCenter Server the object is being managed from.** The SPBM engine on the other remote vCenter Servers will not see this VM so the policies on those vCenter servers will not impact the VM.

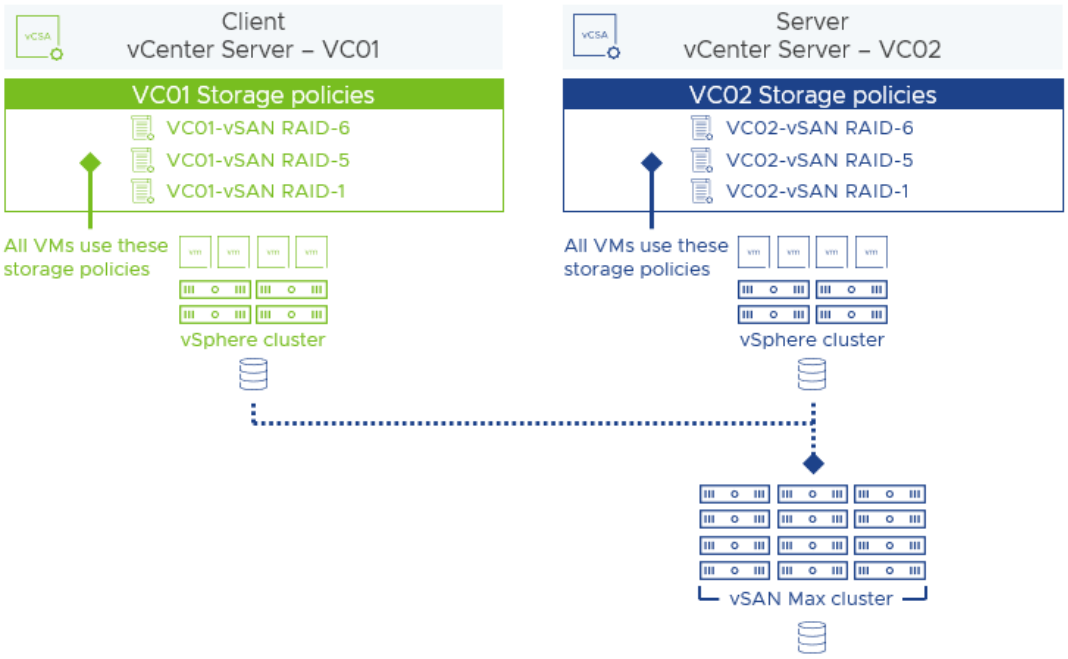

Figure. Storage policy usage in a cross-vCenter Server connection.

Auto-Policy Management enabled for a vSAN storage cluster **will not apply this cluster-optimized default storage policy to any client cluster managed by another vCenter Server instance.** One will need to select the appropriate RAID-6 storage policy on the vCenter Server instance managing the client cluster.

Day-2 Operations and Optimizations

Cluster Updates and Patching

Use "Ensure Accessibility" when entering a host in a vSAN storage cluster into maintenance mode.

vSAN storage clusters will support the use of durability components which will not only improve the availability of the most recently written data planned and unplanned host outages, but they dramatically reduce the time it takes to update data after the maintenance event completes. Given that all data in a vSAN storage clusters will be protected through storage polices with FTT=2 using RAID-6, this makes the data under maintenance events extremely resilient. Full evacuations of a host for maintenance purposes are largely unnecessary, and just consume valuable resources. Full evacuations may make sense if you are choosing to permanently remove a host from a vSAN storage cluster.

Become familiar with your server vendor's offering for VMware's vSphere Lifecycle Manager (vLCM).
Ensure that you have downloaded and installed your vendor's plugin for vLCM known as a Hardware Support Manager (HSM), so that the proper desired state image can be created using the appropriate combination of hypervisor version and vendor drivers and firmware. This will make maintaining a vSAN Max cluster easy and predictable. If you are unfamiliar with vLCM, see the topic "Introducing vLCM into an Existing Environment" in the vSAN Operations Guide.

Scaling

Understand how to add resources to an existing vSAN storage cluster.
Resources in a vSAN Max cluster can be scaled easily, and incrementally in ways that is not possible with traditional storage. Adding more resources to a vSAN storage cluster will involve one of the two methods below.

- **Scaling out.** This simply means adding more hosts to a vSAN storage cluster. More hosts is an easy way to distribute the existing workloads across more storage resources, and improve the aggregate capacity and performance resources to the data.
- **Scaling up.** This means adding more or higher density storage devices within the existing hosts that comprise a vSAN storage cluster. See the "Cluster Design and Sizing" section of this document to learn about strategies for growth by adding hosts to clusters.

Depending on your environment, existing hardware configurations, operational procedures, time, procurement processes and hardware availability, one option may be more suitable for you than another. As a vSAN storage cluster is scaled, we recommend striving for uniform levels of resources in both performance and capacity across all hosts in the vSAN storage cluster. See the post "Asymmetrical vSAN Clusters – What is Allowed, and What is Smart" for more information.

Performance Optimizations

If using multiple virtual disks in a VM, configure multiple paravirtual SCSI adapters in a VM's virtual hardware configuration.
This helps the guest operation systems ability to queue additional I/O. The use of multiple VMDKs using multiple paravirtual SCSI adapters in a VM's virtual hardware configuration has been a common recommendation by Independent Software Vendors (ISV) for running their applications optimally in VMs on most storage systems. This recommendation applies to vSAN storage cluster as well. See the "Applications" section in the Troubleshooting vSAN Performance guide for more information.

Monitoring and Event Handling

Learn how to view remote datastores connected to vSAN storage clusters.
For remote datastore connections between client clusters and vSAN storage clusters managed by the same vCenter Server, one can find this by highlighting the vSAN storage cluster, clicking **Configure > vSAN > Remote Datastores**, as shown below.

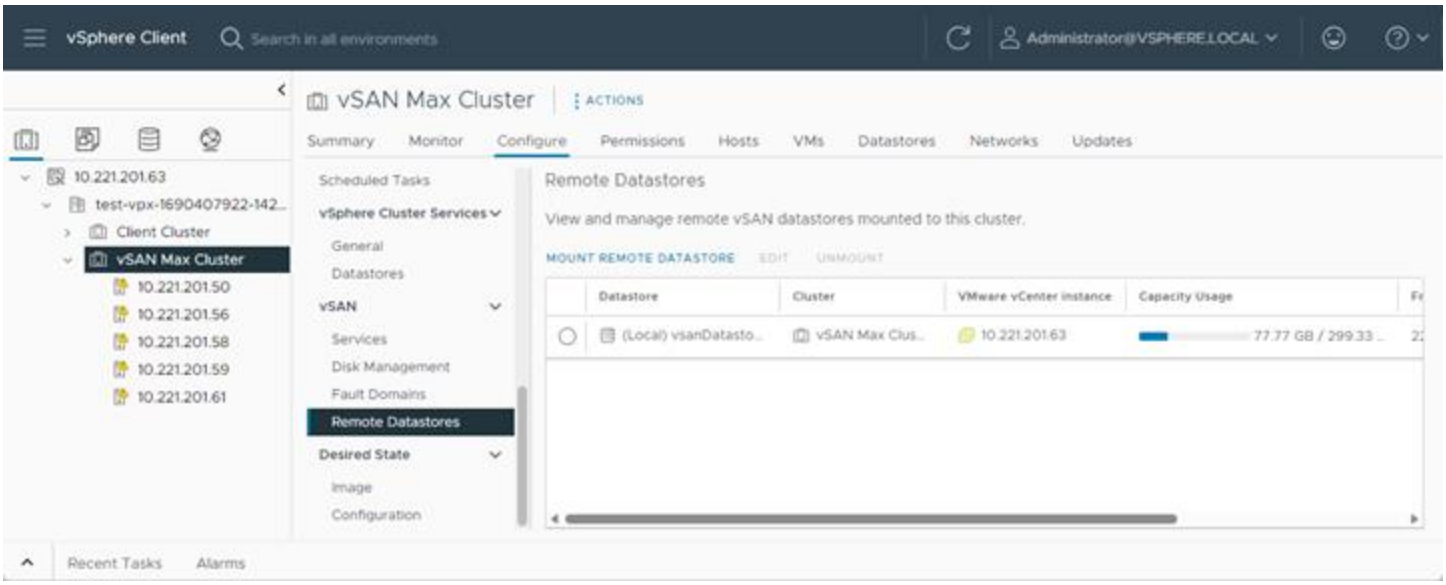

Figure. Viewing remote datastores connected to vSAN storage clusters.

Remote datastores can also be viewed by highlighting the vCenter Server instance, clicking on **Configure > vSAN > Remote Datastores**, as shown below. This view is helpful for connections between vSAN compute clusters and vSAN storage clusters managed by different vCenter Server instances.

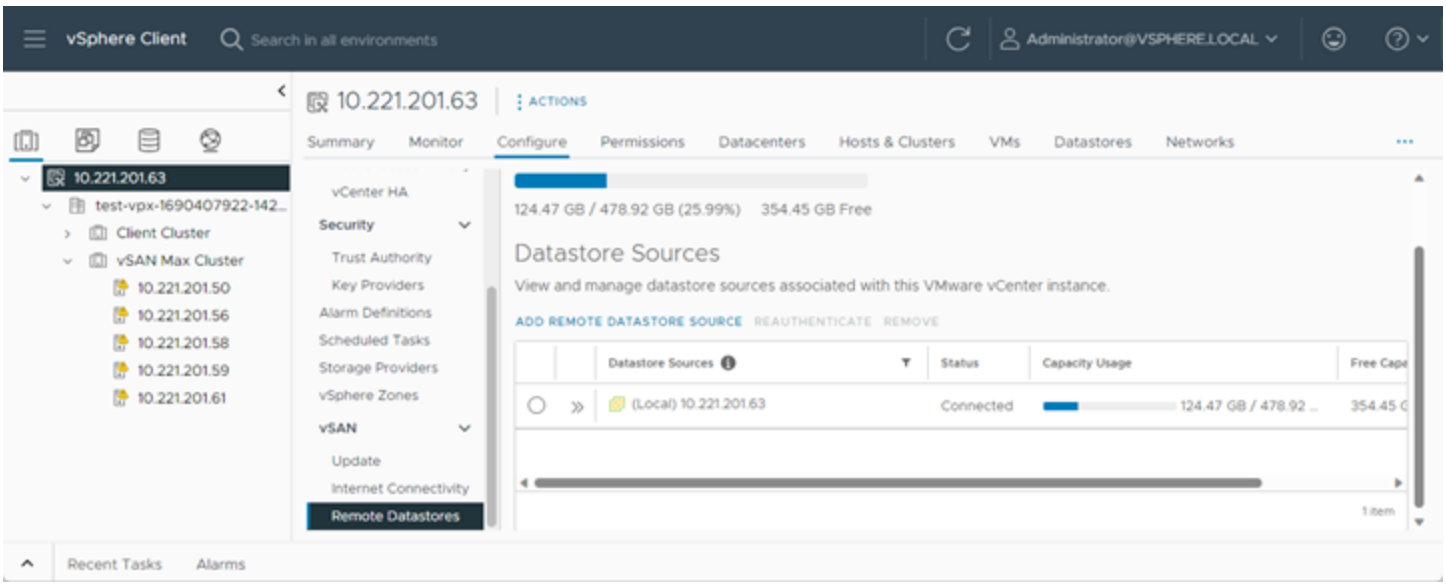

Figure. Viewing remote datastores at the vCenter Server instance level.

Monitor Capacity usage to ensure sufficient capacity.

For single site vSAN storage clusters, enabling the vSAN “Operations Reserve” and “Host Rebuild Reserve” toggles dramatically improve the ability to ability to maintain sufficient free space for transient vSAN storage cluster operations, host failures, and incremental growth. One may wish to customize the capacity warning and error thresholds to suite the needs of the environment. For vSAN storage clusters consisting of just 4 hosts, do not enabled "Host Rebuild Reserve" in combination with the vSAN's Auto-Policy Management feature, as this will prevent it from using RAID-5 erasure coding on a small cluster. See "Auto-Policy Management Capabilities with vSAN ESA in vSAN 8 U1."

The post "Improved Capacity Reporting in VMware Cloud Foundation 5.1 and vSAN 8 U2" also demonstrates how to understand the respective capacity overheads of ESA, which powers vSAN storage clusters.

Host Failures and Remediation.

Much like vSAN HCI clusters, vSAN storage clusters have mechanisms in place to ensure that the data stored maintains availability, and durability as prescribed by the applied storage policies.

Upon a host failure or isolation event, vSAN storage clusters will wait before rebuilding the data elsewhere to regain the prescribed levels of compliance of the data. By default, it waits for 60 minutes to determine if it is a transient event that corrects itself, or a sustained event that requires a rebuild elsewhere. **In most cases, leaving this timer at its default is advised.** However, two considerations may influence a desire to adjust this setting.

- **Resource density of the host configuration**. Hosts configured with higher capacities may take longer to resynchronize data elsewhere upon a failure. Using higher storage capacities may change the desired time you wish for it to wait prior to initiating a rebuild.
- **Failed host replacement workflows.** Some environments use operation run books to adhere to service level agreements (SLA). To adhere to these clearly defined SLAs, these organizations have procedures in place that will replace the entire host regardless of the failure type, such as a fan failure. If an environment uses this type of an approach, the object repair timer may need to be adjusted to fit this workflow and the operating SLAs.

The object repair timer can be adjusted by highlighting the cluster, clicking **on Configure > vSAN > Services > Advanced Options > Object repair timer**.

Note that **resynchronizations only are for the data stored, not the data capacity provided**. For example, if a host is storing 10TB of data but it can provide 100TB of capacity, the synchronization will only be for the 10TB of data.

Storage Device Failure and Remediation.

The impact of a storage device failure in vSAN storage clusters will be limited to the failed device. vSAN storage clusters will reconstruct, or resynchronize this data elsewhere in the cluster to regain its prescribed level of resilience. For this type of failure, since the boundary of failure is limited to just the storage device in question, a relatively small amount of resynchronization will be performed. See the post: "The Impact of a Storage Device Failure in vSAN ESA versus OSA" for more information.

When replacing the device, it will be important to understand how to identify the physical location of the failed device so that the correct device is replaced. vCenter Server allows you to highlight a desired storage device and turn on the device locator LED to correctly identify a specific device within a server. For any type of vSAN cluster, this can be found by highlighting the cluster, clicking **Configure > vSAN > Disk Management**, highlighting the host, clicking **View Disks**, highlighting the desired disk, and clicking on "**Turn on LED**" as shown below.

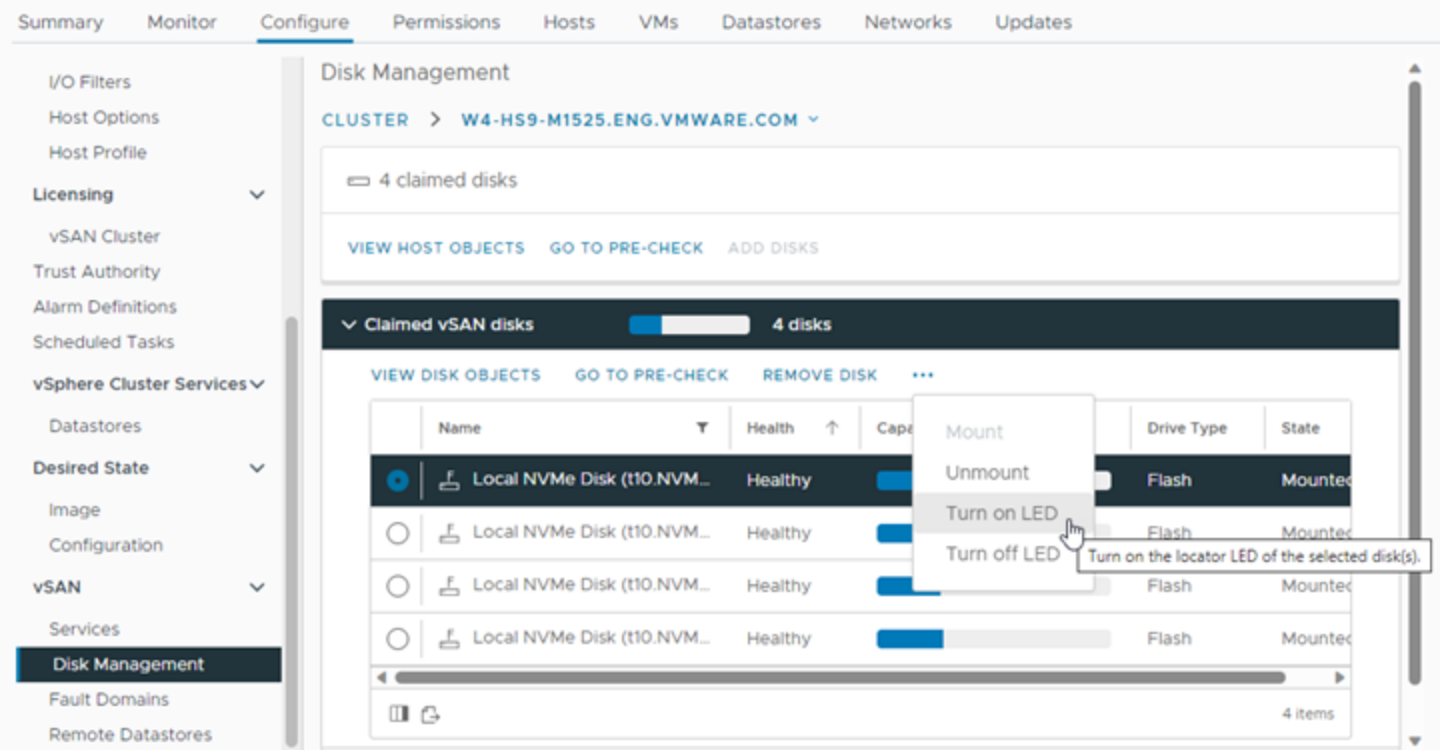

Figure. Managing storage devices in a vSAN storage cluster.

This functionality may be dependent on the capabilities of the server, the storage device, and prerequisite software from the server manufacturer to work. It is recommended to test this functionality in a server prior to entering it into a production environment.

Summary

We believe that aggregated vSAN HCI clusters and disaggregated vSAN storage clusters serving vSphere clusters can provide a powerful combination to your enterprise needs, and can serve as the unified storage platform for all workloads running on VMware Cloud Foundation. The recommendations above will help customers achieve the highest levels of performance, resilience, and operational simplicity for their environments powered by vSAN storage clusters.

Additional Resources

The following are a collection of useful links that relate to vSAN storage 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

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