vSAN Stretched Cluster Guide

Recommendations for vSAN 8 U3 and VMware Cloud Foundation 5.2

January 2, 2025

vmware by Broadcom

Table of Contents

Introduction

vSAN stretched clusters are a powerful option for environments that require the highest levels of data resilience and VM uptime. This guide was developed to provide additional insight and information for the design, installation and configuration of vSAN stretched clusters. It will also cover operational procedures and explain how it handles failure scenarios unique to this topology.

A VMware vSAN stretched cluster is a cluster configuration where the hosts that comprise a vSAN cluster reside in two geographic locations. The two sites typically have an equal number of hosts residing in each site, and are connected using a high bandwidth/low latency connection known as an inter-site link, or ISL. The third site, or tertiary site hosts a vSAN Witness Host that is connected to both data sites. This connectivity can be via low bandwidth/high latency links.

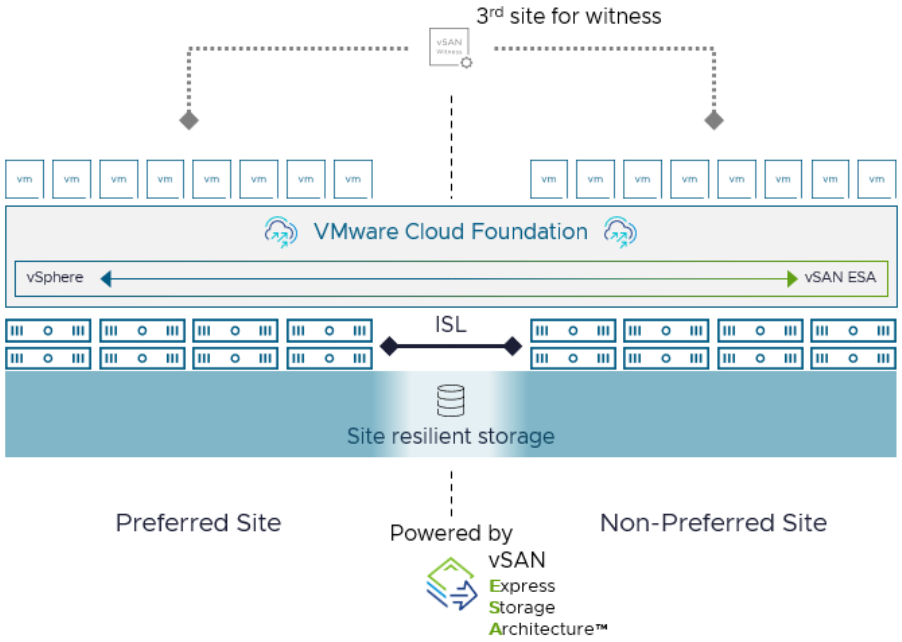


Figure. A vSAN stretched cluster topology in an environment running VMware Cloud Foundation.

vSAN stretched clusters provide more than just data resilience across sites. Thanks to its integration with vSphere HA, and vSphere DRS, it provides a comprehensive solution of high availability of mission critical workloads and the data it serves. vSAN provides this site-level resilience all as a cluster configuration type that is relatively easy to install and manage.

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. The guidance provided here will apply to clusters running vSAN's Original Storage Architecture (OSA), as well as vSAN Express Storage Architecture (ESA). Items specific to one architecture or the other will be noted accordingly.

For simplicity, the information and examples assume vSAN is deployed as an aggregated vSAN HCI cluster. As of vSAN 8 U3 and VCF 5.2, disaggregated vSAN storage clusters (previously known as vSAN Max) have limited support in stretched cluster configurations, as noted later in this document.

vSAN Stretched Cluster Concepts

vSAN Stretched Clusters and Fault Domains

vSAN uses a construct known as a "fault domain" to help it distribute data in a resilient way. By default, vSAN treats each host as a fault domain, which helps keep data available in the event of a discrete host failure. vSAN has an optional feature known

as "vSAN Fault Domains." When one defines a group of hosts as a "fault domain" it identifies this an additional logical boundary of failure. It can use this information to write the data in such a way that if that defined fault domain is offline, the data will remain available. The Fault Domains feature will commonly be used to ensure rack-level resilience in a data center and is compatible with both vSAN OSA, and ESA.

vSAN stretched clusters are a cluster deployment option within vSAN that use this same concept of fault domains. Instead of a fault domain representing a single rack or room in a data center, it represents a physical site. This ensures that the virtual machines and the data they consume will maintain availability in the event of a site outage. These sites are known in the product as:

- **Preferred.** This represents the data site assigned as the primary owner of the objects. In cases of isolation between the Preferred and the Non-Preferred sites, the Preferred site will always take precedence.
- **Non-Preferred.** This represents the data site assigned as the alternate, or second fault domain.
- **Witness.** This represents the witness site that contains only the witness host.

Even though a stretched cluster configuration is most often deployed across geographic sites, these fault domains could represent rooms in a large building or perhaps a campus. It would not represent as much of a geographically dispersed environment to account for natural disaster scenarios, but may be suitable for some customer requirements.

vSAN Witness Host

vSAN determines the availability of VM objects (such as VMDKs) using metadata embedded in every small chunk of data that makes up the object. These small chunks of data are called components and are dispersed across hosts in a way that makes the data resilient, and helps vSAN determine availability of that object. See the post: "vSAN Objects and Components Revisited" for more information.

But stretched clusters need additional help to account for their topology. A witness host is used in these configurations, where it stores a very small amount of metadata associated with the object on the witness host. The metadata helps vSAN understand if there is enough resilient data active to keep the data available or not. The witness host is critical in in site isolation conditions, where one data site could be completely isolated from another site due to network connectivity between sites. vSAN can determine quorum on if and where the data should remain available or unavailable. This is what prevents "split-brain" conditions where a VM and its data are updated independently.

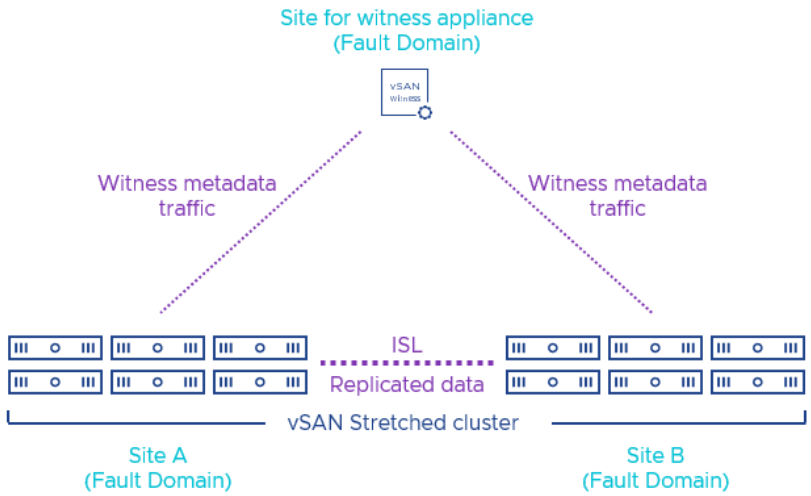


Figure. Relationship of the witness host and witness site to the data sites in a vSAN stretched cluster.

The witness host must be managed by the same vCenter Server managing the vSAN stretched cluster. It will reside in the hosts and clusters inventory within vCenter Server, but will not be a member of a specific cluster. The managing vCenter

Server must be able to communicate with all of the hosts in the vSAN cluster, as well as the witness host. See the vSAN Witness Host Requirements section of this document for more information.

Data Path Optimizations for vSAN Stretched Clusters

vSAN has several data path optimizations built into the product that work well for stretched cluster topologies, as well as other settings that will activate when a stretched cluster is deployed.

Proxy Communication

vSAN's distributed object manager (DOM) uses the concept of a proxy to communicate replica traffic across the ISL efficiently. For example, if one configures a storage policy for site level resilience, using a secondary level of resilience of RAID-5 erasure coding, the DOM owner will send the replica writes to the DOM proxy on the other site. The owner and the proxy will then be responsible for distributing the writes as a RAID-5 erasure code.

Data Compression and Encryption

vSAN ESA will compress and encrypt the data prior to the data being transmitted across the ISL, which will essentially increase the effective bandwidth of the ISL, and minimize the effort of the hosts at each site. For more information, see the post: "Using the vSAN ESA in a Stretched Cluster Topology."

Read Locality

In traditional vSAN clusters, a virtual machine's read operations are distributed across all replica copies of the data in the cluster. vSAN stretched clusters typically have limited bandwidth between data sites. In these configurations, read operations will always come from the site of the VM instance requesting the read. This is enabled as a setting for the cluster, and helps reduce traffic across the inter-site link (ISL). This behavior can be overridden using a storage policy rule.

DRS after Recovery from a Failed Site

In scenarios where a site is offline after a failure, or isolation event, some VMs will be vMotioned back to the failed site in accordance with the DRS "should" rules. DRS is smart enough to wait for the object data of a VM to be fully resynchronized after a failure before it initiates this activity. This ensures that VM read operations are using an optimal data path.

Requirements of a Stretched Cluster

To provide a level of service and supportability for the data that is stored on a stretched cluster, additional requirements extend beyond the minimum requirements for a standard vSAN cluster. They include:

Witness Host Requirements

This is typically in the form of a prepackaged virtual appliance known as a *Witness Host Appliance*. Provided in an OVA format, it can be deployed like other virtual appliances in an infrastructure. The deployment of the appliance will allow you to select a predefined appliance size (e.g. Tiny, Medium, Large, etc.) that best represents the number of VMs in you expect in your cluster.

A physical host can be used as a vSAN witness host, but this does introduce additional complexities, including:

- **Licensing.** Additional licensing will be required when using a physical host, versus no licensing required when using the virtual witness host appliance.
- **Versioning.** A physical host will be required to run the same build of vSphere as the stretched cluster as the hosts in that cluster.
- **Inefficient.** A physical witness host is not as cost effective and agile as a virtual appliance, which can be managed and even redeployed easily.

Witness Host Location

The witness host appliance **must reside at a third, tertiary location that is not the location of either one of the data sites for the stretched cluster it is associated with**. The deployment of virtual witness host appliance will not consume any licensing. If one chooses to use a physical ESXi host to act as the witness, it will be subject to license consumption rules.

The witness host appliance can run on any vSphere environment with any supported storage, such as a VMFS datastore, NFS datastore, or a vSAN cluster. A witness host appliance may run on another vSAN stretched cluster, **but only if that other stretched cluster does not reside in the same sites for the cluster the witness host appliance is serving.** This will help avoid an unintended situation in which one site failure initiates another site failure for another stretched cluster.

Witness Host in the vSphere Client UI

When deployed, the virtual witness host appliance will show up in the "Hosts and Clusters" view of the vSphere client with a special color. A dedicated witness host appliance will be needed for every vSAN stretched cluster deployed. The witness host appliance should not be added as a member of any cluster.

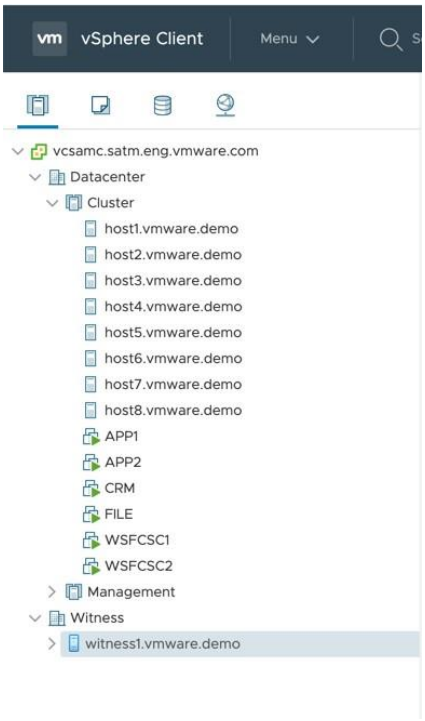


Figure. The virtual witness host appliance in the Hosts and Clusters view.

While each vSAN stretched cluster must have its own dedicate witness host appliance, this third, tertiary site can host several witness host appliances that can be used for other stretched clusters. They may even use the same VLAN if desired.

Witness Host Lifecycle Management

Even though the witness host is deployed as a virtual machine, it can be managed using VMware vSphere Lifecycle Manager (vLCM). This means you can upgrade the stretched cluster and the witness host appliance using a single method. Alternatively, one can simply deploy a new witness host appliance and replace through an easy workflow in the UI.

Networking Requirements

Since vSAN is a distributed storage system, networking plays a critical role in any vSAN environment. When a cluster is stretched across two physical data sites, there are additional networking requirements to consider beyond what is required for a standard vSAN cluster.

Data Site to Data Site Network Bandwidth and Latency

The minimum bandwidth required between data sites is **10Gbps**. This applies to both OSA and ESA clusters. While this is the supported minimum, as the performance capabilities of vSAN increase, the demands on the ISL, as well as the expectations with performance may also increase. After the deployment of a stretched cluster, monitor the effective bandwidth used across

the ISL to ensure it is sized correctly. This will help ensure that performance expectations and data availability requirements are met.

The minimum latency required between the data sites hosting virtual machine objects should not exceed **5 milliseconds** (ms) in round trip time (RTT), or 2.5ms one way. The latency minimums exist largely to ensure that applications can write data resiliently across two sites in a timely manner. This latency requirement will dictate the realistic geographic distance between sites.

Data Site to Witness Site Network Bandwidth and Latency

Since the witness site only stores small amounts of metadata, and is not a part of the data path, the network requirements are much less compared to the requirements between data sites. The amount of network bandwidth between the data site and the witness site will depend on the number of objects components in vSAN. Approximately **2Mbps of bandwidth is needed for every 1,000 components on vSAN**. The Design Considerations section as well as the vSAN stretched cluster Bandwidth Sizing document will cover this in more detail.

The minimum latency required between witness site and the data sites should not exceed **200ms RTT**, or 100ms one way.

Firewalls, IDS and WAN Optimization

While front-end VM traffic can run through network overlays, we highly recommend that all VMkernel traffic (including vSAN traffic) has as simple and unmanipulated of a path as possible. Firewalls, NAT and IDS/IPS systems can inadvertently block this mission critical storage I/O in a manner that could cause substantial impacts on the performance or availability of data. WAN optimization appliances can also be problematic.

Witness Traffic Separation (Optional)

In the simplest of stretched cluster configurations, all members of a vSAN stretched cluster will have a VMkernel interface tagged for vSAN, for use with all vSAN-related traffic. vSAN stretched clusters and vSAN 2-Node clusters support the ability to separate traffic for the witness host appliance from the traffic for the data sites. This is referred to as "witness traffic separation" (WTS), and is a configuration that would be applied to all hosts in the two data sites that participate in the stretched cluster. It does not apply to the configuration of the witness host appliance. This customization provides more flexibility in accommodating network conditions or requirements, such as unique network characteristics to the witness site, or various security requirements.

MTU Sizes

vSAN clusters require hosts with unform MTU sizes across the cluster. This allows for data to be transmitted in a consistent way, and reduces packet fragmentation, dropped packets and retransmits.

When using vSAN's witness traffic separation capability, one can specify a different MTU size used for communication from the data site to the witness site, versus communication from the data site to the other data site. This is useful for ISLs that support larger MTU sizes that may not be possible on the links connected to the witness site.

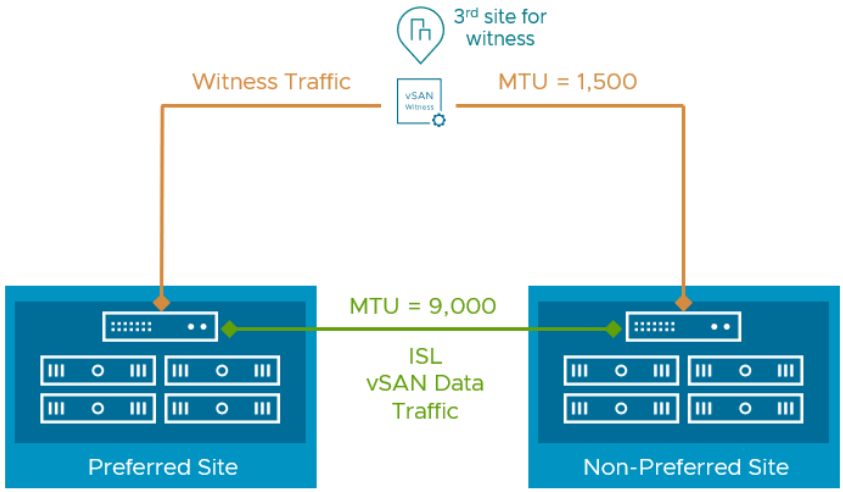
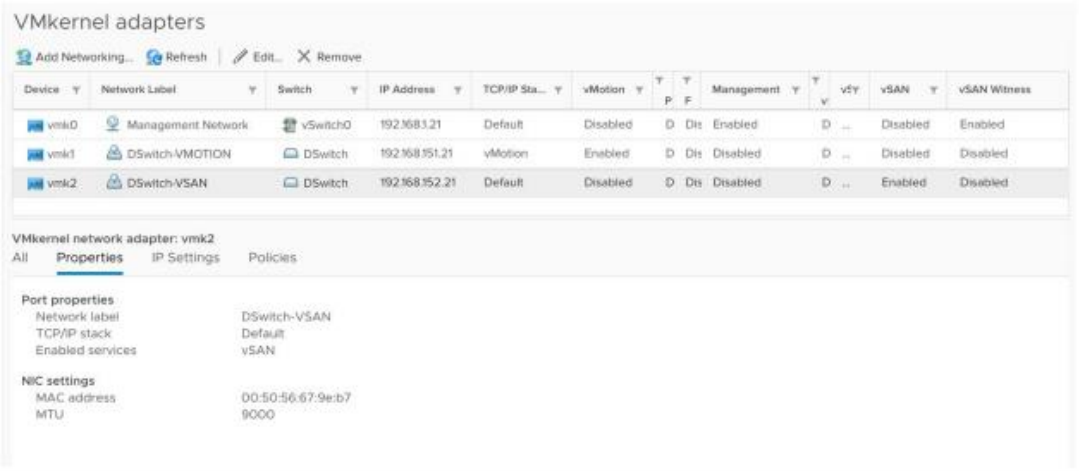


Figure. Illustrating how vSAN supports the use of mixed MTU sizes.

vSAN's set of health checks will recognize deployments using witness traffic separation and allow for different MTU sizes for the vSAN data traversing the ISL, and the witness traffic communicating with the witness site. The images below show the vSAN data network (vmk2) using an MTU size of 9,000 while the witness traffic is using an MTU size of 1,500.



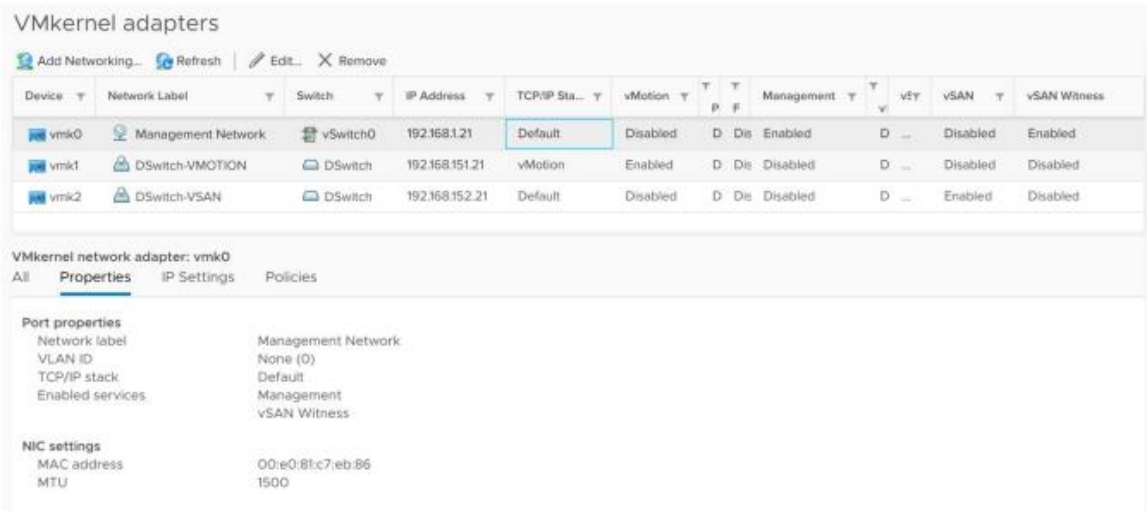


Figure. Configuration of witness traffic separation on the data hosts in a vSAN stretched cluster.

On the vSAN Witness, the Management VMkernel (vmk0) is tagged for vSAN Traffic with an MTU of 1500.

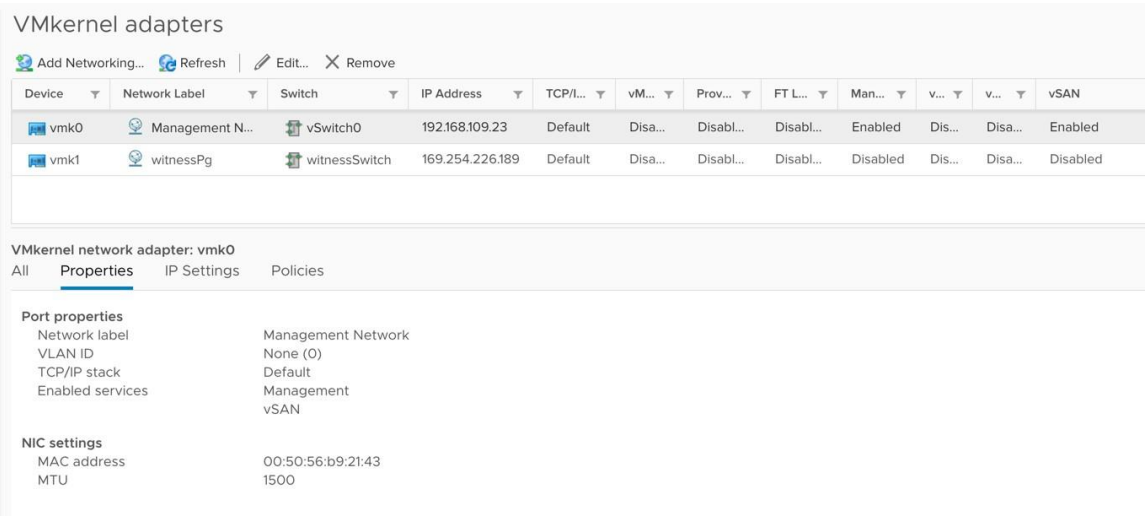


Figure. Configuration of the witness host appliance in a vSAN stretched cluster.

Configuration Minimums and Maximums

vSAN stretched clusters do have considerations and limitations that are unique to the topology.

Virtual Machines Per Host

The number of virtual machines supported in vSAN is unaffected by a vSAN stretched cluster deployment. The maximum number of VMs supported is the same as standard vSAN cluster deployments - up to 200 VMs per host when running vSAN OSA, and up to 500 VMs per host when running vSAN ESA. Since stretched clusters are designed to provide resources in the event of a full site failure, the following limits should be incorporated into your design exercise.

- **In a non-failure state, no more than 50% of the maximum number of VMs supported on a host.** This will help account for VMs restarting courtesy of HA on the remaining site during a site failure.
- **In a non-failure state, host compute and memory resources do not exceed an average of 50% utilization.** This will help account for VMs restarting courtesy of HA on the remaining site during a site failure.

Hosts Per Stretched Cluster

The minimum number of hosts in a vSAN cluster is 1 host in each data site, and one vSAN witness host appliance in a third site. This is sometimes referred to as a 1+1+1 deployment and is an atypical configuration since it has limited abilities in resilience. Most configurations will have at least 3 hosts in each data site and a witness host appliance in a third site, sometimes referred to as a 3+3+1 deployment. Three hosts in each site is the minimum required to use a secondary level of resilience in a stretched cluster configuration.

The maximum number of hosts in a vSAN stretched cluster is 40 hosts, plus the witness host appliance (20+20+1). In some high capacity or high VM count conditions, it may make sense to limit the cluster to no greater than 32 hosts, as that can work best with host component count limits for vSAN.

Symmetry vs Asymmetry

vSAN supports asymmetrical configurations in standard vSAN clusters, and vSAN stretched clusters. This can exist in the form of different amounts of resources within each host in the cluster, or a different number of hosts in each site of a stretched cluster. As with standard vSAN clusters, one should strive for reasonable levels of symmetry across a cluster, as taking asymmetry to extremes may prevent the ability to store the data resiliently on each site, or use resources in an efficient way.

For more information and guidance on asymmetry in vSAN, see the post: "Asymmetrical vSAN Clusters – What is Allowed, and What is Smart."

Witness Host

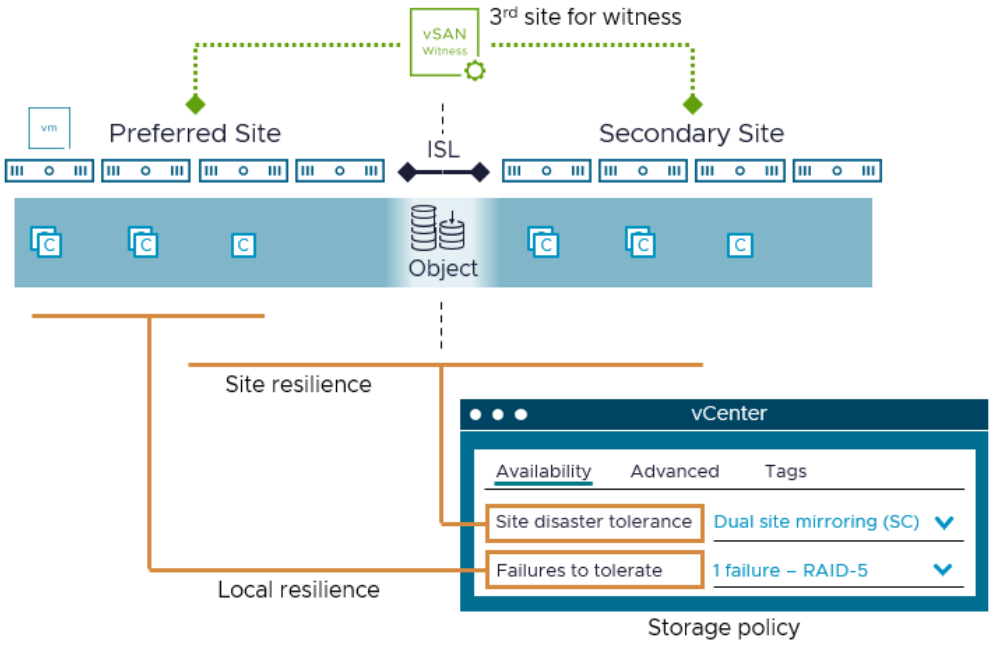
Each vSAN stretched cluster will use a vSAN witness host, typically using a vSAN virtual witness host appliance. This appliance will be installed at a third site, and used exclusively for the vSAN cluster for which it is assigned.

Design Considerations

Storage Policies with Stretched Cluster Topologies

Storage polices can account for the site-based topology of a stretched cluster. These storage policies will allow you to specify a "Site disaster tolerance" as well as a secondary "Failures to tolerate" or "FTT" for the environment. When the policy is applied to one or more data objects, vSAN will take care of the rest, ensuring to the best of its ability that it adheres to the defined outcome of that policy.

vSAN also allows you to apply storage policies that do not provide resilience across sites. This means that the VM data does not need to be synchronously written to both sites, but rather, only the site where the VM resides. When paired with DRS host and VM rules, this can be useful for applications that either do not need to be replicated across both sites, or use their own application-level replication. Even in cases where no site-level resilience is needed, one can still provide resilience within a site by applying the secondary level of failures to tolerate.



Stretched Cluster-Specific Storage Policy Rules

The "Site disaster tolerance" entry will have the following options available:

- **None – Standard Cluster.** Used for standard vSAN clusters that are not stretched.
- **Dual site mirroring (stretched cluster).** Maintains availability of data in the event of an entire site.
- **None – Keep data on Preferred.** Stores data only on the preferred site.
- **None – Keep data on Non-preferred.** Stores data only on the non-preferred site.

The "Failures to tolerate" entry will define the optional, secondary level of resilience. Assuming a dual site mirror, selecting "1 failure – RAID-5" will store the data in a space efficient RAID-5 erasure code within each site, in addition to the dual site mirroring. It assumes that one has enough hosts within each site to support the desired setting.

vSAN stretched clusters rely on DRS settings such as VM and host group assignments and "should" rules to help balance VMs across the hosts in a stretched cluster. For VMs do not need site-level resilience for the reasons stated above, DRS "must" rules help pin the VM instance same site that the VM data resides. A vSAN stretched cluster requires the configuration of these DRS rules for

How Stretched Cluster Storage Policy Rules Impact Capacity Requirements

The ability to make data resilient upon a host failure, or even a site failure means that the data must be written in a resilient way. The amount of raw storage capacity required will depend on the level of resilience you desire and specify in the storage policy. The table below shows how much raw capacity will be required for a fictional VM consuming 100GB storage as seen by the guest OS. For simplicity, the examples do not include opportunistic space efficiency features like compression or deduplication, and only the most common levels of resilience were chosen. Note that vSAN ESA uses a different RAID-5 erasure coding scheme than vSAN OSA. This impacts storage capacity consumption, as noted below. For more information, see the post: "Adaptive RAID-5 Erasure Coding with the Express Storage Architecture in vSAN 8."

Storage Policy Rule	vSAN Architecture	Capacity Consumption Multiplier in Accordance to Applied Policy	Capacity Consumption Result	Minimum Hosts Required in Each Data Site
Dual site mirroring without secondary level of resilience	Both	2x	200GB	1
Dual site mirroring with secondary FTT=1 using RAID-1	Both	4x	400GB	3
Dual site mirroring with secondary FTT=1 using RAID-5	OSA	2.66x	266GB	4
Dual site mirroring with secondary FTT=1 using RAID-5	ESA	3x for clusters with fewer than 12 hosts 2.5x for clusters of 12 or more hosts	300GB or 250GB respectively	3 6
Dual site mirroring with secondary FTT=2 using RAID-6	Both	3x	300GB	6
Preferred Site (no site mirroring) with secondary FTT using RAID-1	Both	2x	200GB	3 (in one site)
Preferred Site (no site mirroring) with secondary FTT using RAID-5	OSA	1.33x	133GB	4 (in one site)
Preferred Site (no site mirroring) with secondary FTT using RAID-5	ESA	1.5x for clusters with fewer than 12 hosts	150GB or 125GB respectively	3 (in one site) 6 (in one site)

		1.25x for clusters with 12 or more hosts		
Preferred Site (no site mirroring) with secondary FTT using RAID-6	Both	1.5x	150GB	6 (in one site)

Storage policies will impact the amount of data that is traversing the ISL. For more information, see the vSAN Stretched Cluster Bandwidth Sizing Guide.

Network Design

The three fault domains that comprise a stretched cluster must maintain network connectivity to each other during a normal operating state. This configuration will allow the loss of any one site and still maintain availability of the data.

Connectivity and Network Types

	Preferred Site	Non-Preferred Site	Witness Site
Management Network	Layer 2 or 3 to vCenter/vSAN Hosts	Layer 2 or 3 to vCenter/vSAN Hosts	Layer 2 or 3 to vCenter
VM Network	Recommend Layer 2	Recommend Layer 2	No requirement for a VM Network if using the vSAN Witness Appliance. Running VMs on the vSAN Witness Appliance is not supported. Running VMs on a Physical Witness Host is supported.
vMotion Network	If vMotion is desired between Data Sites, Layer 2 or Layer 3 are supported vMotion is not required between this Data site & the Witness Site	If vMotion is desired between Data Sites, Layer 2 or Layer 3 are supported vMotion is not required between this Data site & the Witness Site	There is no requirement for vMotion networking to the Witness site.
vSAN Network	To the Secondary Site: Layer 2 or Layer 3	To the Preferred Site: Layer 2 or Layer 3	To the Preferred Site: Layer 3 To the Secondary Site: Layer 3

Layer 2 versus Layer 3 Networking

While vSAN does provide a lot of flexibility with networking connectivity, we **recommend that stretched cluster topologies use Layer 3 (routed) networking between the two data sites, and the witness site.** Layer 3 will help avoid spanning tree protocol (STP) from redirecting traffic across an undesirable link, such as the more bandwidth constrained link to the witness site. At minimum, if using Layer 2 between data sites, ensure that the connectivity from each data site to the witness site is using Layer 3 routing.

When networking is properly configured, the following characteristics will exist

- The data sites will only be able to communicate with each other using the ISL

- The witness host should only be able to communicate with hosts in each data site directly, and not through another data site.

vSAN VMkernel interfaces use the same default gateway as the Management VMkernel interface. **The default gateway for the VMkernel adapter can be overridden in the UI to provide a different gateway for the vSAN network**. This feature simplifies routing configuration that previously required manual configuration of static routes in the vSphere CLI or PowerCLI. These options are still available, but more tedious.

Witness traffic separation allows the ability to use an interface different than that of the Management VMkernel interface. This can provide additional levels of security isolation if desired. When using witness traffic separation:

- If another VMkernel interface is tagged as "witness" traffic (other than the Management VMkernel interface is used which is typically vmk0) static routes will be required to communicate with the vSAN Witness Host VMkernel interface tagged for vSAN traffic.
- If the management VMkernel interface is tagged with "witness" traffic, static routes are not required if the host can already communicate with the vSAN Witness Host VMkernel interface using the default gateway.
- If only a single subnet is available for the vSAN Witness Host, it is recommended to untag vSAN traffic on vmk1 and tag vSAN traffic on vmk0 on the vSAN Witness Host.

Note that the ability to change the default gateway applies to all services on the specific VMkernel on the host. If a VMkernel interface is providing multiple services (vSAN, mgmt, vMotion, etc.) it will change the default gateway for everything on that VMkernel interface.

Network Port Communication

vSAN requires the following ports to be open, both inbound and outbound

	Port	Protocol	Connectivity To/From
vSAN Clustering Service	12345, 23451	UDP	vSAN Hosts
vSAN Transport	2233	TCP	vSAN Hosts
vSAN VASA Vendor Provider	8080	TCP	vSAN Hosts and vCenter Server
vSAN Unicast Agent (to Witness host)	12321	UDP	vSAN Hosts and vSAN Witness Appliance

Default Gateway on ESXi hosts in a vSAN Stretched Cluster

ESXi hosts come with a default TCPIP stack. As a result, hosts have a single default gateway. This default gateway is associated with the Management VMkernel interface (typically vmk0). For vSAN clusters, we recommend using another VMkernel interface with its own IP subnet, and tagging this network for use with vSAN.

Since VMkernel ports tagged for vSAN use the same TCP/IP stack as the management VMkernel interface, the traffic will attempt to use the same default gateway. When vSAN is on its own subnet, this presents a challenge, because the default gateway specified is not on the same subnet as the vSAN traffic. This can be addressed in one of two ways.

- Overriding the default gateway in the vSphere client UI
- Establish static routes on each host within a cluster.

Overriding the default gateway in the vSphere client UI is the easiest of the two options. It also allows you to quickly identify and correct misconfigurations.

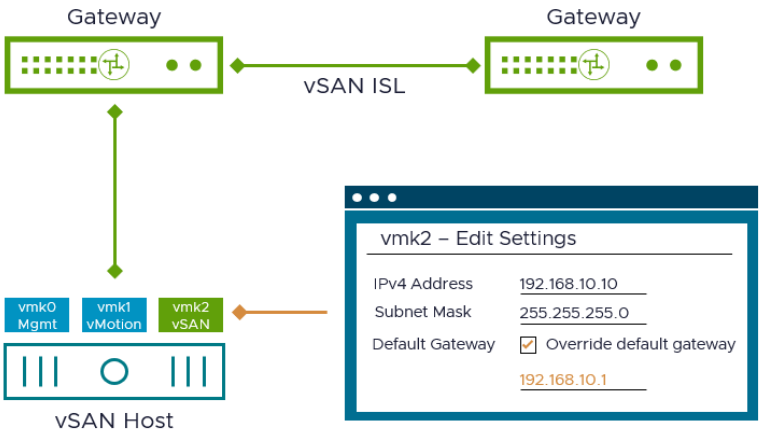


Figure. Changing the default gateway for VMkernel traffic tagged for vSAN

Static routes can also be established on each host of the vSAN cluster, including the witness host. This option predates the easy method found in the UI, but remains available using ESXCLI or PowerCLI. Static routes are added via the esxcli network IP route or esxcfg-route commands. Refer to the appropriate vSphere Command Line Guide for more information.

Routing of other Networks

For vSAN stretched clusters, there is typically no need to configure service networks such as vMotion or VM port group networks with static routes. The use of default gateway overrides, or static routing generally applies to the communication to and from the witness host. The witness host does not use these other networks. Note that the witness host will need to communicate with vCenter server, so in some topologies, a static route may need to be added on the witness host.

Custom TCP/IP Stacks

vSAN traffic uses the default TCP/IP for ESXi. It does not have a dedicated TCP/IP stack, and custom TCP/IP stacks cannot be used for vSAN traffic

Network Bandwidth

Network bandwidth and latency plays an important role in the effective performance capabilities of a vSAN stretched cluster. The amount of bandwidth needed between data sites is ultimately determined by the workload characteristics or requirements of your environment. For more information on calculating bandwidth requirements, see the “vSAN Stretched Cluster Bandwidth Sizing” guide.

Cluster Settings - vSphere HA

VMware vSAN is fully integrated and supported with vSphere HA, and should be enabled on all vSAN clusters. Some additional setting within the vSphere HA configuration are recommended, and detailed below:

Host monitoring

Host monitoring should be enabled on vSAN stretch cluster configurations. This feature uses network heartbeats to determine the status of hosts participating in the cluster, and if corrective action is required, such as restarting virtual machines on other nodes in the cluster.

Virtual Machine Response for Host Isolation

This setting determines what happens to the virtual machines on an isolated host, suche as a host that can no longer communicate to other nodes in the cluster nor reach the isolation response IP address. For stretched clusters, setting the “Response for Host Isolation” to “Power off and restart VMs” is recommended. This is because a clean shutdown will not be possible as on an isolated host. The VM would be unable to write any data to disk in this condition.

Admission Control

Admission control ensures HA has sufficient resources to restart virtual machines after a failure. As a complete site failure is one scenario that needs to be considered in a resilient architecture, VMware recommends enabling vSphere HA Admission Control. Availability of workloads is the primary driver for most stretched cluster environments. Sufficient capacity must, therefore, be available for a total site failure. Since ESXi hosts will be equally divided across both sites in a stretched cluster, it is recommended that Admission Control be set to 50% for both memory and CPU.

Admission control does not apply to storage resources. Assuming all VM data is synchronously replicated between sites, in the event of a site failure, there is no need to consume any more storage capacity, since the data already exists in the site.

Host Hardware Monitoring – VM Component Protection

This setting can be left disabled, as it was designed primarily for traditional three-tier storage to help accommodate for "All Paths Down" (APD) and "Permanent Device Loss" conditions.

Heartbeat Datastores

vSphere HA can optionally use another heartbeat mechanism to determine the state of the hosts in a cluster. In a vSAN cluster, the vSAN datastore is NOT used for heartbeats. It is recommended to disable this option, but if another datastore is available, it can be used for this purpose. vSphere HA may produce a warning message if no additional heartbeat datastores are available. This can be suppressed by following Broadcom Article ID: 318871.

Advanced Options

The use of host isolation addresses is particularly important to ensure the proper behavior of vSAN stretched clusters during a failure condition, such as a site partiion. While vSAN clusters shift vSphere HA communication from the VMkernel interface tagged for management to the VMkernel interface tagged for vSAN, vSphere HA continues to use the default gateway of the management network. One must change the isolation addresses from the management network to the vSAN network to ensure that HA is testing against the network providing data communication between sites, and not the management network or any loopback address that may provide a false-positive response. In the advanced settings, one can add the following options:

- das.isolationaddress0 = <IP address on the vSAN network in the Preferred Site>
- das.isolationaddress1 = <IP address on the vSAN network in the Non-Preferred Site>
- das.usedefaultisolationaddress = false

An example of these advanced settings is provided below.

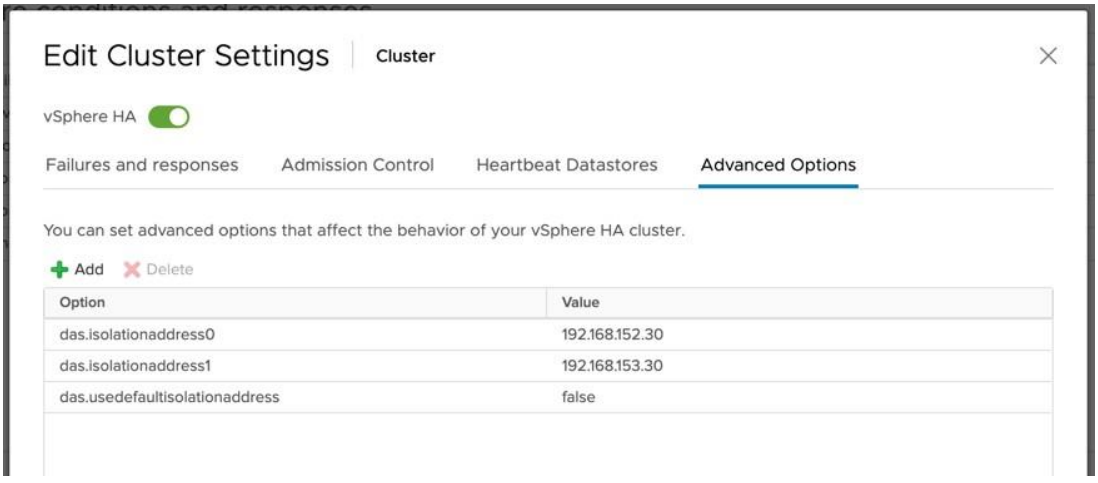


Figure. Setting isolation addresses for a vSAN stretched cluster

Cluster Settings - DRS

vSphere DRS is used to intelligently distribute VM instances across a cluster. It is fully integrated with vSAN its use is encouraged. It works extremely well without any additional modifications, but with vSAN stretched clusters, a few simple configurations will need to be made for it to operate as desired. The additional settings **help provide DRS the awareness necessary of one site versus another site** so that vSAN can distribute VMs across both sites, and pin a VM instance to the same site that the data resides if the VM uses a storage policy that only stores the data in one site.

These Host Groups, VM Groups, and VM/Host Rules should align with your storage policies to ensure that the VMs and data resilience outcomes you desire can be achieved.

Host Groups

For a vSAN stretched cluster, two host groups should be created. Each host group should represent a given geographic site. Within that host group, add all vSAN hosts for that stretched cluster that reside in the site.

VM Groups

In stretched cluster configurations, one will want to create at least two VM groups. A very simple example would consist of two VM groups, that each contain a roughly equal number of VMs in each group. Although, if site affinity is used, one will want to create additional VM groups that contain those VMs not protected across sites, and which site the VM instance may run.

VM/Host Rules

Host groups and VM groups are used to create rules within DRS to suggest or force a behavior of VM placement in a vSAN stretched cluster. These rules generally fall into one of two categories.

- **"Should" rules.** These rules tell DRS that under normal operating conditions, a set of VMs will run on one site, while another set of VMs run on another site. These "should" rules help accommodate for site failure conditions, where the rule allows the VMs to run in the other site in the event of an HA event like a site outage. **These rules will likely be used for most of your workloads.**
- **"Must" rules.** These rules tell DRS that under normal operating conditions, a group of VM will only run on a given site (the same site as the data) within a stretched cluster. If there is a site outage, HA will not restart those VMs in the other site, as the data would be unavailable. **These rules will typically be used for VMs in your stretched cluster that do not need site-level resilience or use their own application-level replication for resilience.**

Recommendation: Make sure your host group and VM group names are descriptive and reflect the name of the site or intention. E.g. "Hosts-Site-A" "Replicated-VMs-Site-A" "Affinity-VMs-Site-A" etc. This will help define and manage the VM/Host Rules in a simple and clear way.

Partially Automated Mode vs Fully Automated Mode in vSphere DRS

Both "partially automated" and "fully automated" modes in DRS are supported in vSAN. In partially automated mode, DRS will handle the initial placement of virtual machines. However, any further migration recommendations will be surfaced to the administrator to decide whether or not to move the virtual machine. The administrator can check the recommendation and may decide not to migrate the virtual machine. Recommendations should be for hosts on the same site. DRS will take care of virtual machines' initial placement and ongoing load balancing in fully automated mode. DRS should adhere to the Host/VM groups and rules and never balance virtual machines across different sites. This is important as virtual machines in a vSAN stretched cluster will use read locality, which implies that they will cache locally. If DRS migrates the virtual machine to the other site, the cache must be warmed on the remote site before it reaches its previous performance levels.

Storage Policies used in Conjunction with VM/Host Groups and Rules

With the DRS rules in place, one must ensure that storage policies are available that reflect the intention of the DRS rules. For example, for VMs that use the DRS "should" rules, any storage policy that provides site-level resilience should be fine. For VMs that use DRS "must" rules, a storage policy must be created that reflect the fact that the data may only be on one site. Creating a storage policy named something like "No Site Resilience, Site A VMs - FTT=2") or something similar would make it easy to align the VMs that use this storage policy to the VMs in each site affinity VM group.

vSphere Clustering Service (vCLS)

The vSphere Clustering Service (vCLS) was introduced in vSphere 7 U1 and aims to decouple services such as DRS and HA from vCenter Server to provide better scalability and autonomy of these cluster services in the event that vCenter Server is unavailable. For more information, see the post: vSphere 7 U1 – vSphere Clustering Service (vCLS). vCLS will create self-managed VM instances across vSAN clusters in standard and stretched cluster configurations. Ideally at least one vCLS VM should be running in each site. For more information on vCLS impact on vSAN workflows, see KB Article ID: 326879

Initial Deployment

An initial deployment of a vSAN stretched cluster assumes

- All physical hosts are installed with ESXi, and configured for base levels of connectivity in their respective sites.
- A virtual witness host appliance (OVA) is downloaded and ready for deployment.
- Networking matters (subnetting, IP address assignments, have been established.
- You have proper credentials to log in and deploy a new vSAN cluster

The guidance below **is not intended to be step-by-step instructions**, but general guidance on the initial deployment of a vSAN stretched cluster.

Install a vSAN Witness Host Appliance

A vSAN virtual witness host appliance must be installed at a third location, and hosted in a vSphere or vSAN cluster, or stand-alone ESXi host. The witness host appliance will contain two virtual network adapters connected to separate vSphere Standard Switches (VSS).

The vSAN Witness Appliance Management VMkernel is attached to one VSS, and the WitnessPG is attached to the other VSS. The Management VMkernel (vmk0) is used to communicate with the vCenter Server for appliance management. The WitnessPG VMkernel interface (vmk1) is used to communicate with the vSAN Network. This is the recommended configuration. These network adapters can be connected to different, or the same, networks. As long as they are fully routable to each other, it's supported, separate subnets or otherwise.

The Management VMkernel interface could be tagged to include vSAN Network traffic as well as Management traffic. In this case, vmk0 would require connectivity to vCenter Server and the vSAN Network.

A virtual witness host appliance is essentially a nested ESXi instanced running on and environment powered by vSphere. You may find that promiscuous mode is required to allow all Ethernet frames to pass to all VMs attached to the port group, even if it is not intended for that particular VM. Promiscuous mode is enabled in these environments to prevent a virtual switch from dropping packets for (nested) vmnics that it does not know about on nested ESXi hosts.

The installation of the witness host appliance OVA will step you through a deployment wizard allowing you to select the name, resource, networks, credentials, and so on. It will also allow you to deploy the witness host appliance that is sized correctly for your environment. Choose the size as desired. Once the installation completes, power on the witness and begin the configuration process.

Configure a vSAN Witness Host Appliance

One can open up the Direct Console User Interface (DCUI) with its infamous yellow and black background. Hit F2 to customize the configuration, including VMkernel IP addresses, VLANs, DNS addresses and default gateways. You will find this step similar to configuring any other physical ESXi host.

Adding the witness host to the vCenter Server Inventory

This virtual ESXi host can now be added to a vCenter Server inventory, so that it can be easily accessible for management, and recognizable by the forthcoming stretched cluster. Once added, you can place them in the root of a data center, or in a folder. Remember that these virtual witness host appliances cannot be a member of a cluster. You will also note that the virtual witness host appliances will appear with a special blue shading. This will not happen if a physical ESXi host is being used as a witness.

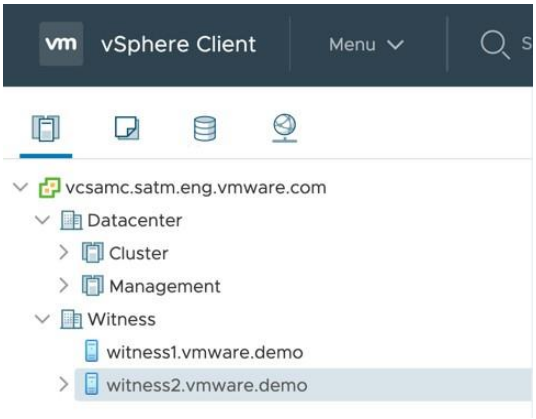


Figure. Viewing witness host appliances in the vSphere Client.

Configure Witness Host Networking

Next, configure the vSAN network correctly on the vSAN Witness Appliance. When the Witness is selected, navigate to Configure > Networking > Virtual switches. The Witness has a port group pre-defined called witnessPg. Here the VMkernel port to be used for vSAN traffic is visible. If there is no DHCP server on the vSAN network (which is likely), then the VMkernel adapter will not have a valid IP address. Select VMkernel adapters > vmk1 to view the properties of the witnessPg. Validate that "vSAN" is an enabled service as shown below.

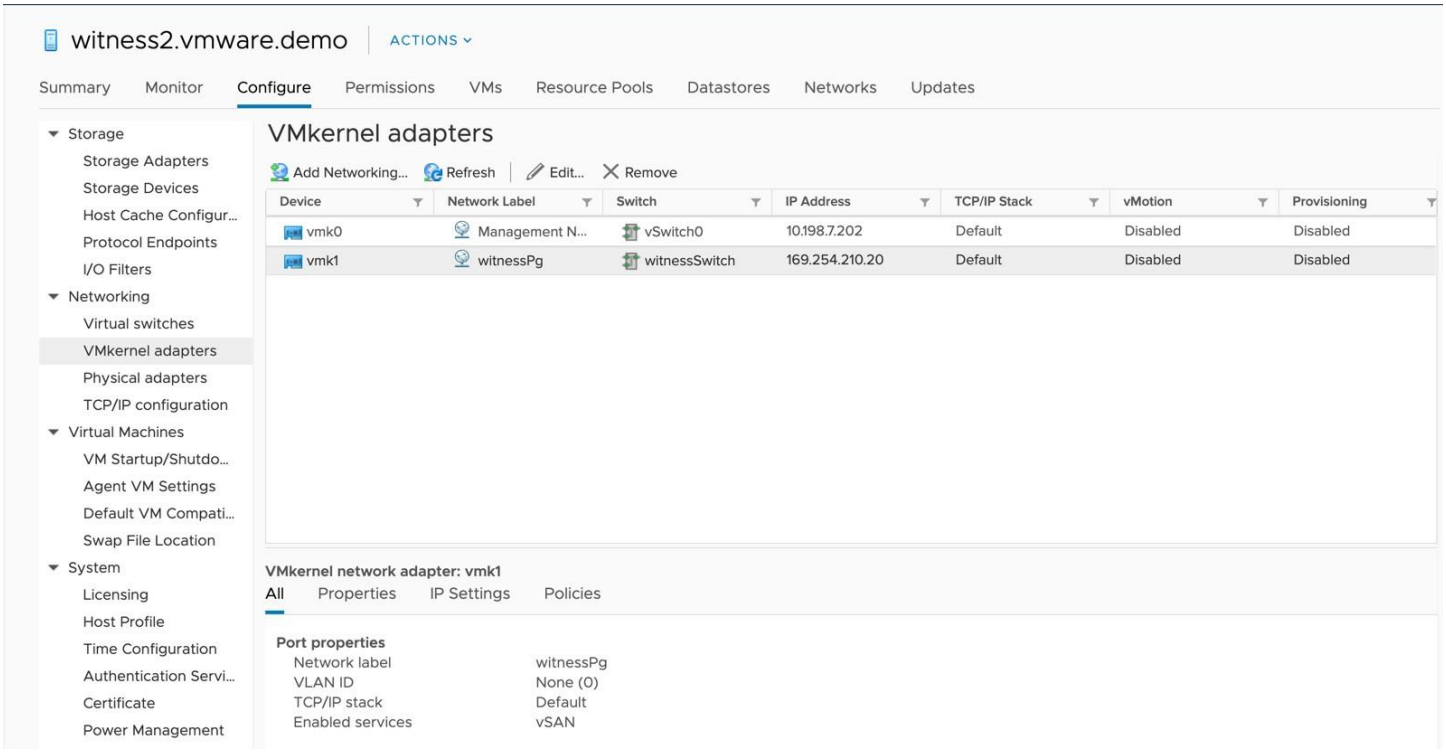


Figure. Listing of VMkernel adapters on witness host appliance.

Simply assign the IP addressing information, as well as tagging the correct data services for your respective VMkernel interfaces. Static routes are still required by the witnessPg VMkernel interface (vmk1). This is because vSAN uses the default TCP/IP stack, just as the Management VMkernel interface does, which is typically on a different network than vmk1. The "Override default gateway for this adapter" setting is not supported for the witness VMkernel interface (vmk1). Once the witnessPg VMkernel interface address has been configured, click OK.

Validate Networking

Before continuing with other configuration steps for a vSAN stretched cluster, we want to ensure there is connectivity among the hosts in each site and the Witness host. It is important to verify connectivity before attempting to configure vSAN stretched clusters.

By default, traffic destined for the vSAN Witness host have no route to the vSAN networks from hosts. As a result, pings to the remote vSAN networks fail.

As highlighted previously, static routes tell the TCPIP stack to use a different path to reach a particular network. Now we can tell the TCPIP stack on the data hosts to use a different network path (instead of the default gateway) to reach the vSAN network on the witness host. Similarly, we can tell the witness host to use an alternate path to reach the vSAN network on the data hosts rather than via the default gateway.

Even if Layer 2 is used to communicate between the data sites, Layer 3 must be used to communicate from the data sites to the witness site. Therefore, either a default gateway override, or static routes are needed between the data hosts and the witness host for the vSAN network, but they are not required for the data hosts on different sites to communicate to each other over the vSAN network.

Hosts in Site A

Looking at **host esxi01-sitea.rainpole.com** , initially, there is no route from the vSAN VMkernel interface (vmk1) to the vSAN Witness Appliance vSAN VMkernel interface with the address of **147.80.0.15.** Notice that there is no communication when attempting to ping the vSAN Witness Appliance from host esxi01-sitea.rainpole.com's vSAN VMkernel interface (vmk1).

The command **vmkping -I vmk1 <target IP>** uses vmk1, because the -I switch specifies using the vmk1 interface.

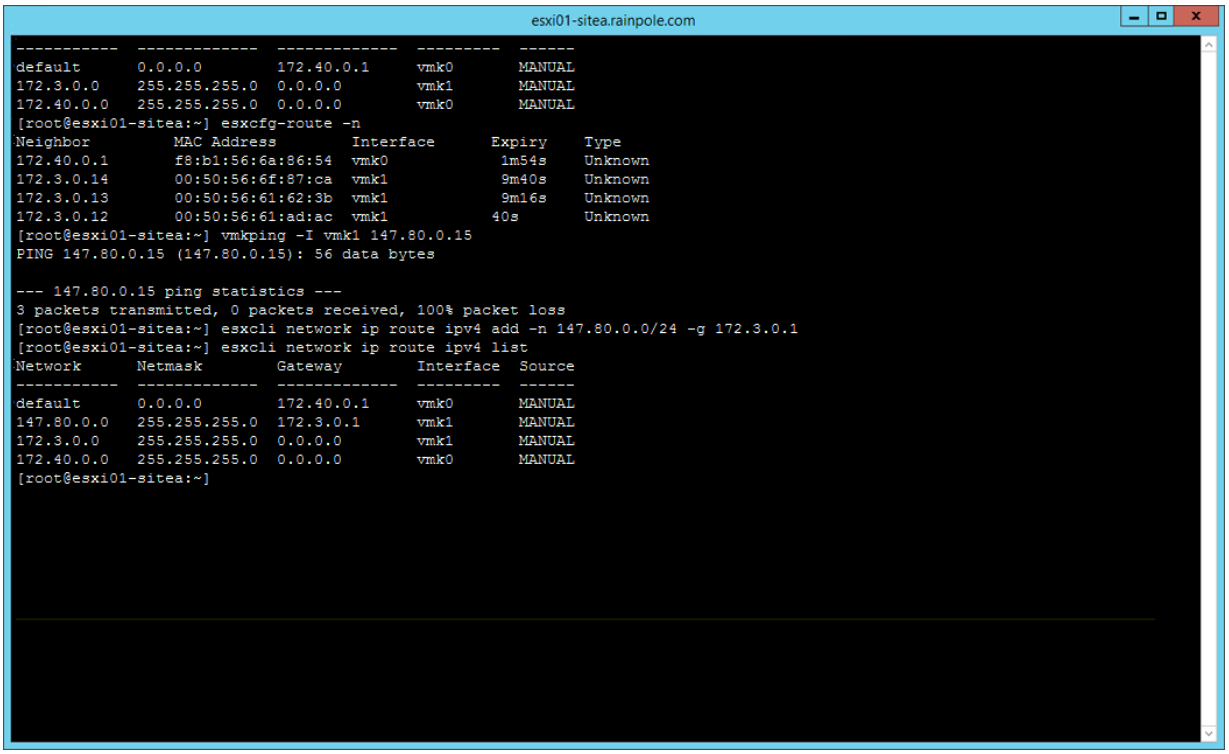


Figure. Testing continuity for a stretched cluster

Add a static route for each host. The esxcli commands used to add a static route is:

esxcli network ip route ipv4 add –n <remote network> -g <gateway to use>

The command used above for the hosts in Site A **is esxcli network ip route ipv4 add -n 147.80.0.0/24 -g 172.3.0.1.** This is because the hosts in Site A have a gateway to the Witness vSAN Appliance vSAN VMkernel interface through 172.3.0.1. Other useful commands are esxcfg-route –n , which will display the network neighbors on various interfaces, **and esxcli network ip route ipv4 list**, to display gateways for various networks. Make sure this step is repeated for all hosts.

Hosts in Site B

Looking at **esxi02-siteb.rainpole.com ,** it can also be seen that there is no route to the vSAN VMkernel Interface (vmk1) to the vSAN Witness Appliance VMkernel interface with the address **147.80.0.15**. The issue is the same **as esxi01-sitea.rainpole.com** on **esxi02-siteb.rainpole.com.**

The route from Site B to the vSAN Witness Appliance vSAN VMkernel interface, is different however. The route from Site B (in this example) is through **172.3.0.253**.

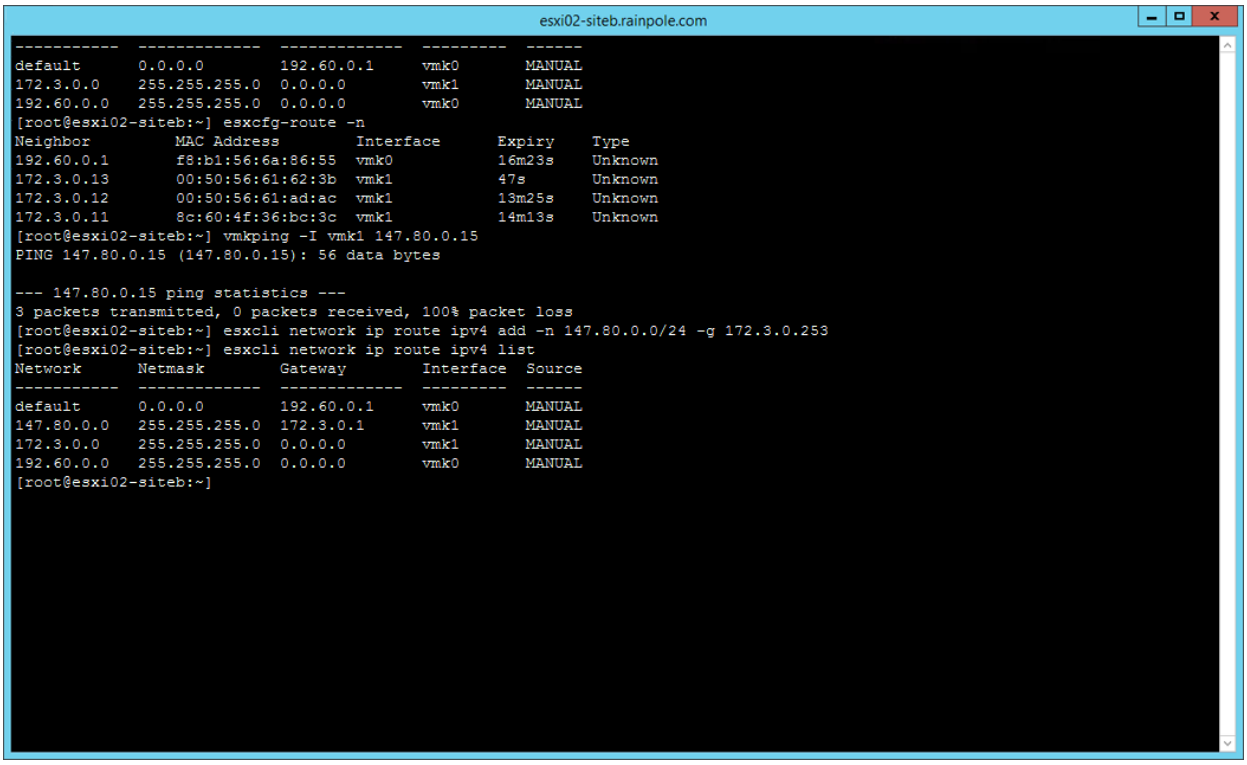


Figure. Testing continuity for a stretched cluster.

The command used above for the hosts in Site B is esxcli network ip route ipv4 add -n 147.80.0.0/24 -g 172.3.0.253.

The vSAN Witness Appliance in the 3rd Site

The vSAN Witness Appliance, in the 3rd Site, is configured a bit different. The vSAN VMkernel interface (vmk1) must communicate across different gateways to connect to Site A and Site B. Communication to Site A in this example must use 147.80.0.1, and communication to Site B must use 147.80.0.253.

Routes must be added for each vSAN VMkernel interface for each host in Site A and Site B on the vSAN Witness Appliance.

To do this individually for each host in Site A, the commands would be:

esxcli network ip route ipv4 add -n 172.3.0.11/32 -g 147.80.0.1

esxcli network ip route ipv4 add -n 172.3.0.12/32 -g 147.80.0.1

To do this individually for each host in Site B, the commands would be:

esxcli network ip route ipv4 add -n 172.3.0.13/32 -g 147.80.0.253

esxcli network ip route ipv4 add -n 172.3.0.14/32 -g 147.80.0.253

With proper routing for each site, connectivity can be verified. Before verifying, let's review the configuration.

Configuration Summary

Below represents a summary of the steps performed above, and a template for how a vSAN stretched cluster can be configured.

Host	VMkernel	IP	Site	Static route to witness	Static Route to Site A	Static Route to Site B	Fault Domain
esxi01-sitea.rainpole.com	vmk1	172.3.0.11	A	172.3.0.1	N/A	N/A	Preferred
esxi02-sitea.rainpole.com	vmk1	172.3.0.12	A	172.3.0.1	N/A	N/A	Preferred
esxi01-siteb.rainpole.com	vmk1	172.3.0.13	B	172.3.0.253	N/A	N/A	Non-Preferred
esxi02-siteb.rainpole.com	vmk1	172.3.0.14	B	172.3.0.253	N/A	N/A	Non-Preferred
witness-01.rainpole.com	vmk1	172.3.0.15	3	N/A	147.80.0.1	147.80.0.1	witness-01.rainpole.com

Configuring a vSAN Stretched Cluster

The simplest of configurations, a new stretched cluster can be created from a group of hosts that does not already have vSAN enabled. In this example, we will use the Cluster QuickStart option in the UI, but the cluster can be created in other ways. For this example, there are 6 nodes available: esx01-sitea, esx02-site a, esx03-sitea, esx01-siteb, esx02-siteb, and esx03-siteb. All six hosts reside in a vSphere cluster called stretched-vsan. The seventh host witness-01, which is the witness host, is in its own data center and is not added to the cluster.

Begin the cluster creation by clicking on Cluster QuickStart. Simply create a new cluster with the name desired, and make sure vSAN is slider toggle is enabled.

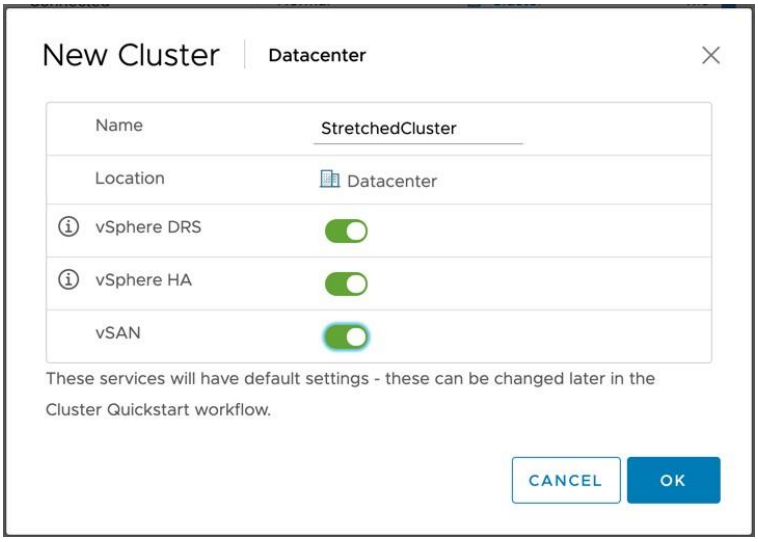


Figure. Enabling vSAN

Initiate the Cluster QuickStart, which will guide you to the next step of adding hosts.

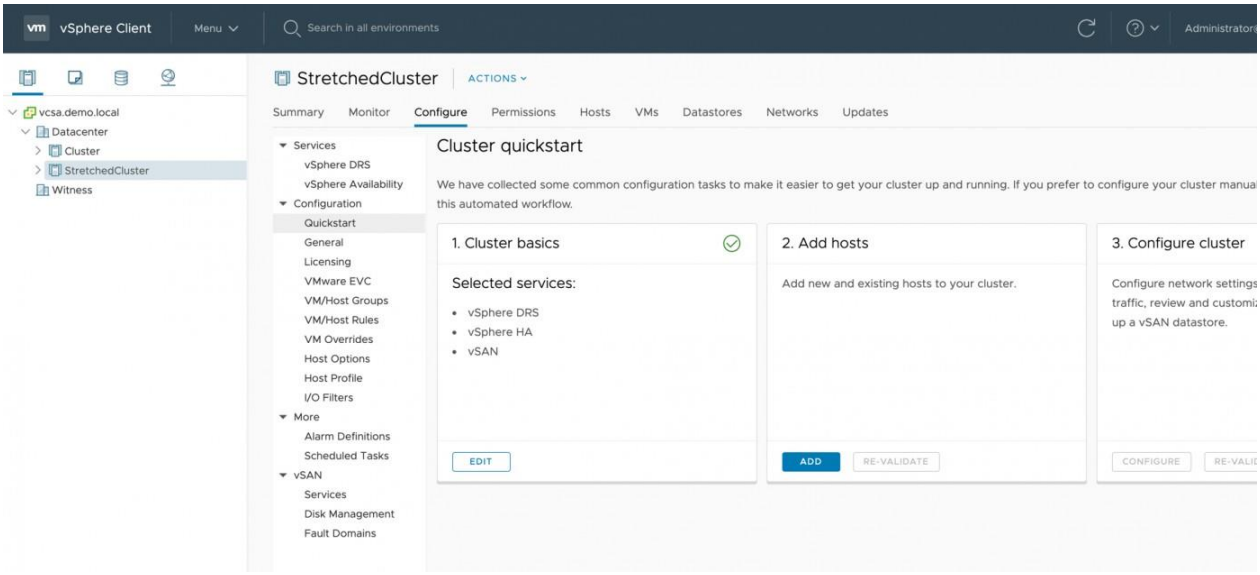


Figure. Creating a cluster using the Cluster QuickStart.

The next step will be to add the hosts desired

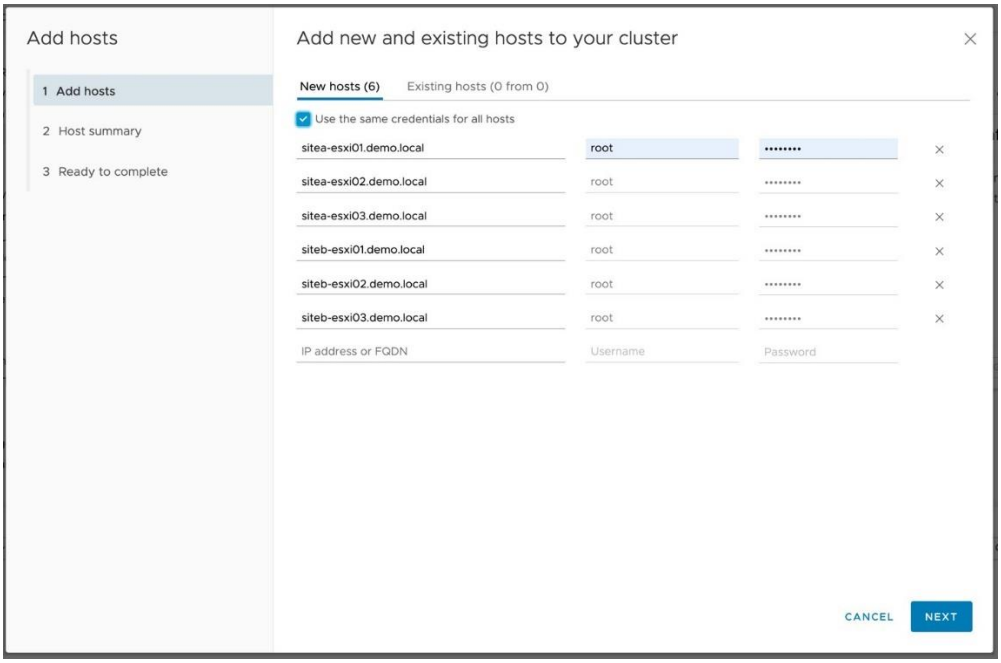


Figure. Adding new hosts to a vSAN cluster

Complete the addition of hosts in the new cluster, and proceed to the “Configure” portion of the Cluster QuickStart. This wizard starts with the ability to configure one or more vSphere Distributed Switches for use by vMotion and vSAN. If one or more vSphere Distributed Switches is already created, they may be used instead of creating new vSphere Distributed Switch(es). Choosing "Configure network settings later" will bypass the networking configuration. Use this setting if vMotion and vSAN configurations are already configured.

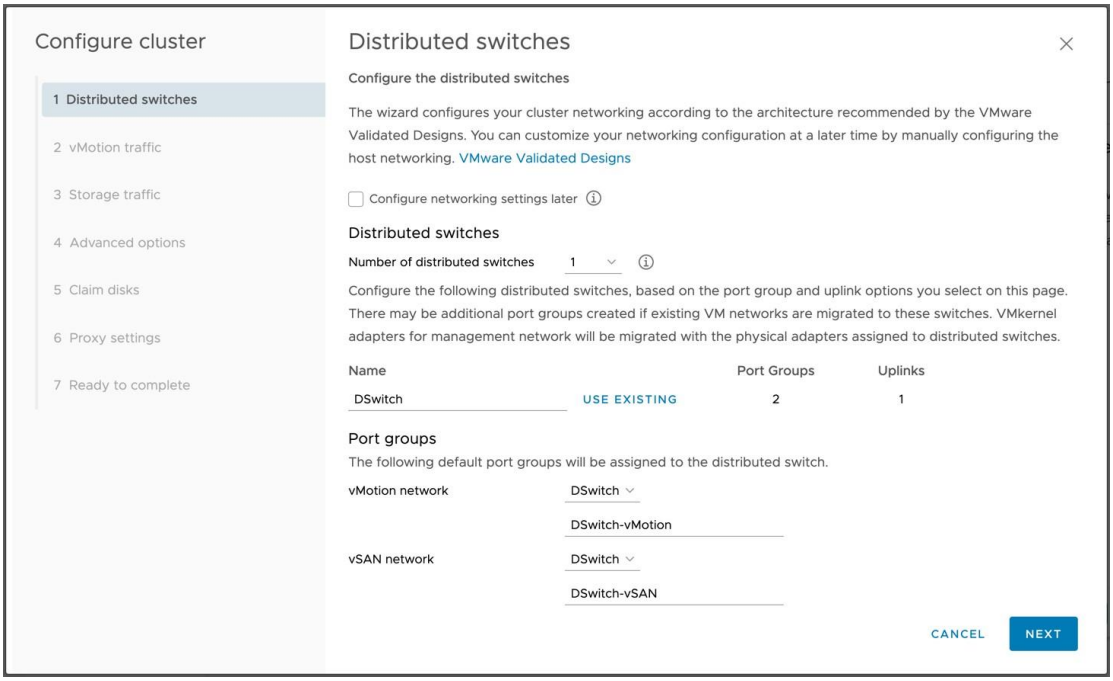


Figure. Distributed switch configuration in Cluster QuickStart.

Next, configure vMotion traffic. Enter in all of the specifics as necessary, such as VLAN tags, network addresses, etc.

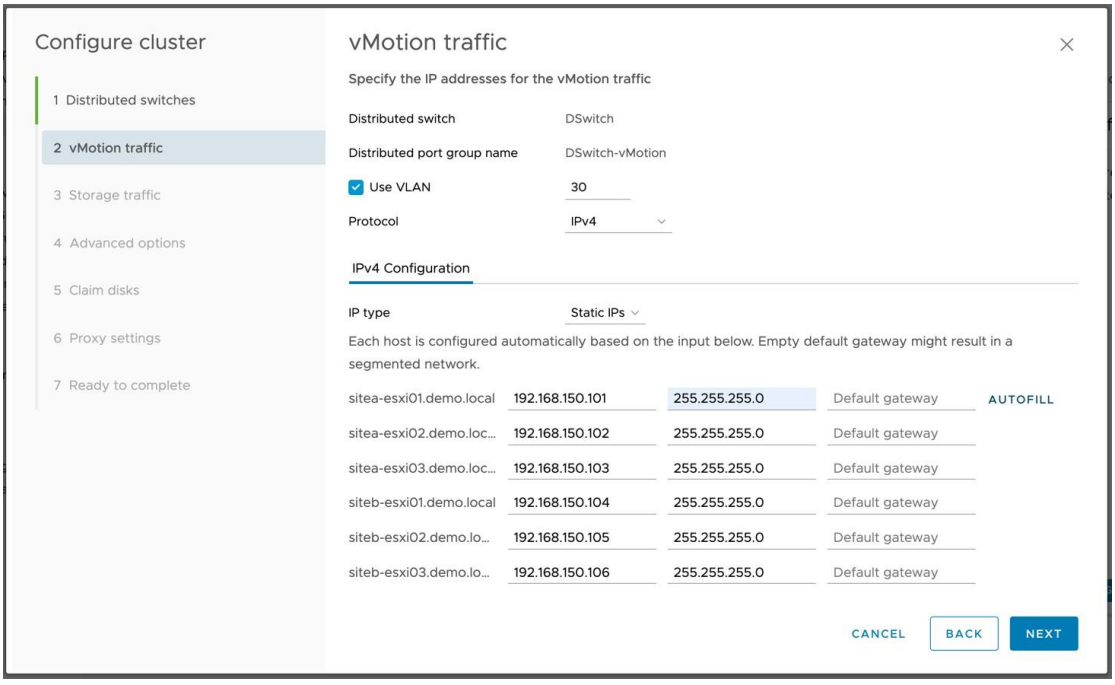


Figure. vMotion configuration in Cluster QuickStart.

Next up is to configure vSAN specific information, including VLAN, IP addresses, default gateways, etc. You can use the "Autofill" button to expedite this process.

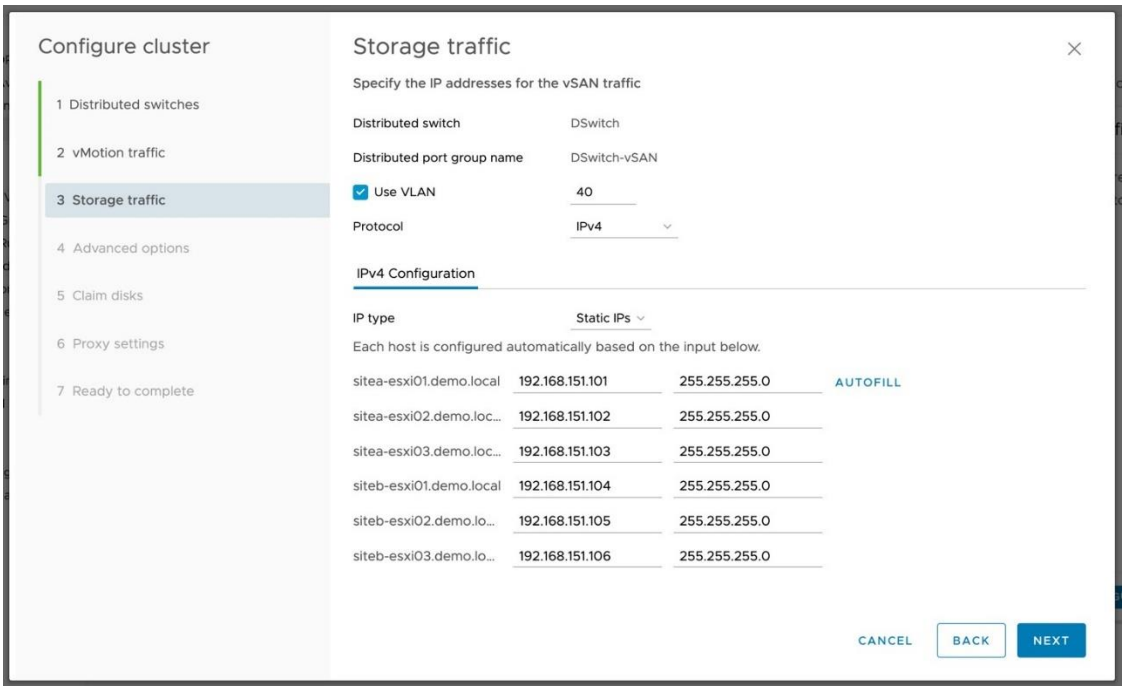


Figure. vSAN configuration in Cluster QuickStart.

Once completed, one can set HA with its appropriate settings (e.g. Host Failure Monitoring, Admission Control, etc.) as well as setting DRS to "Fully Automated" or "Partially Automated" based on your preference. The wizard will then step you through the process of claiming the storage devices, which is significantly easier in vSAN ESA than it was in vSAN OSA. The wizard will then present the options for hosts and their membership into one of two fault domains: The "Preferred" and "Non-Preferred" fault domains. These represent the geographic sites where the hosts reside.

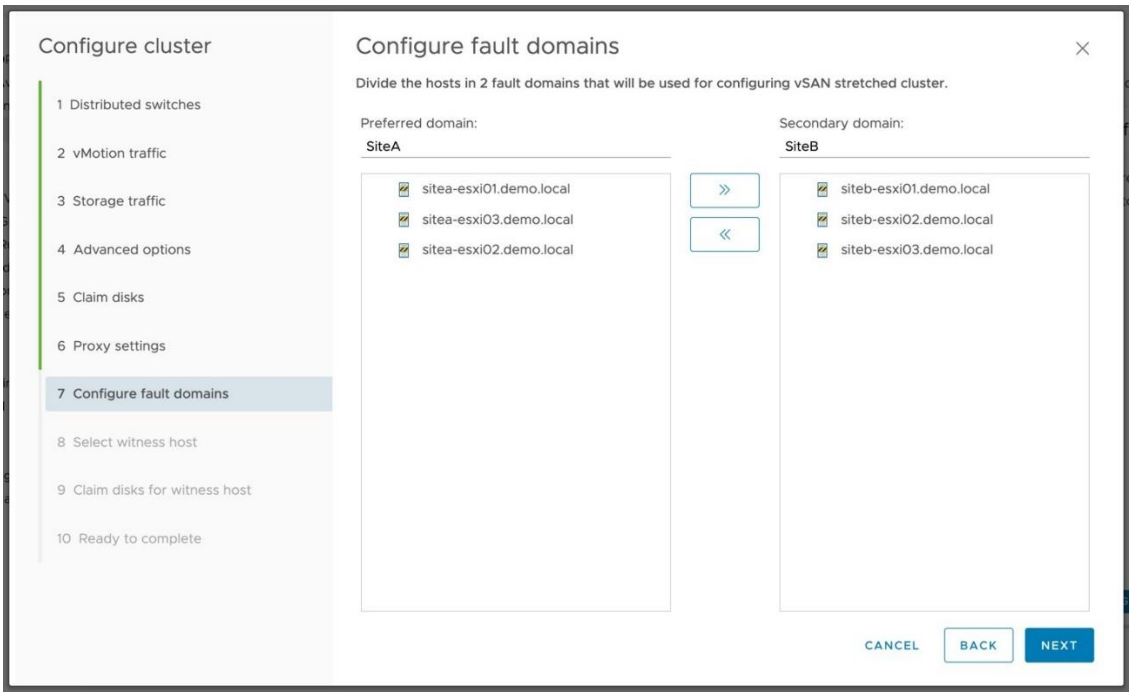


Figure. Assigning hosts to their respective fault domains (sites) in a stretched cluster.

After a few additional steps one can select the witness that will be participating in this vSAN stretched cluster, and complete the wizard.

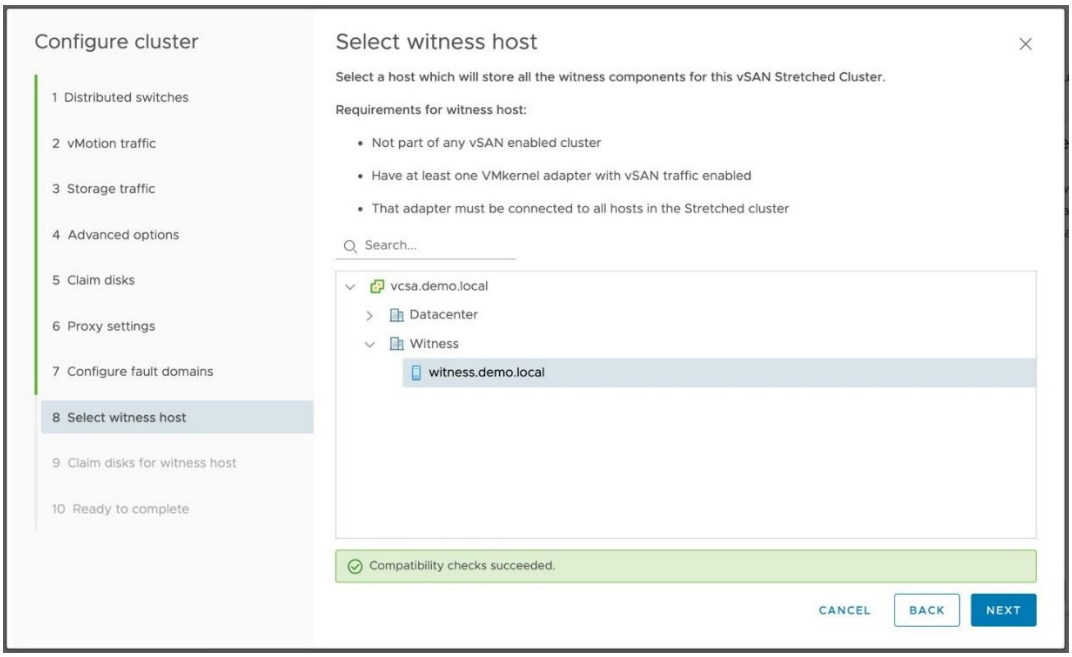


Figure. Completing the wizard by selecting the witness host appliance for use with the stretched cluster.

Converting a standard vSAN Cluster to a vSAN Stretched Cluster

A standard, single site vSAN cluster can be converted to a stretched cluster. Knowing what has been shown in this document, we see that a stretched cluster is nothing more than a vSAN cluster with 3 fault domains – two fault domains representing the data sites, and one fault domain representing the site for the witness.

The steps involve highlighting the vSAN cluster, clicking on the Fault Domains feature. One can select the respective hosts in the two fault domains, and the wizard will then prompt for the witness host to be used, since a cluster consisting of two fault domains would not be complete or a supported arrangement without a witness.

Verify Cluster Settings, and Configure DRS Rules and Storage Policies

Once completed with the initial configuration of a stretched cluster, a quick verification of steps may be helpful. They would include, but are not limited to:

- Ensure Site Read Locality is enabled on the vSAN stretched cluster. This can be found in the settings of the cluster.
- Verify proper HA settings. The recommended settings are described earlier in this document.
- Verify proper DRS settings, and the creation of VM/Host groups and rules. The recommended settings are described earlier in this document.
- Create storage policies that reflect your resilience outcomes
- Check Skyline Health for vSAN. This can be the best way to discover and correct any outstanding issues.

Testing

Once completed, you can create a VM and look at its layout of data components. Simply highlight the VM in question, click **Monitor > vSAN > Physical Disk Placement**, and you will see various components distributed across the hosts in the cluster. It will show the state of the components that comprise a given object for a VM.

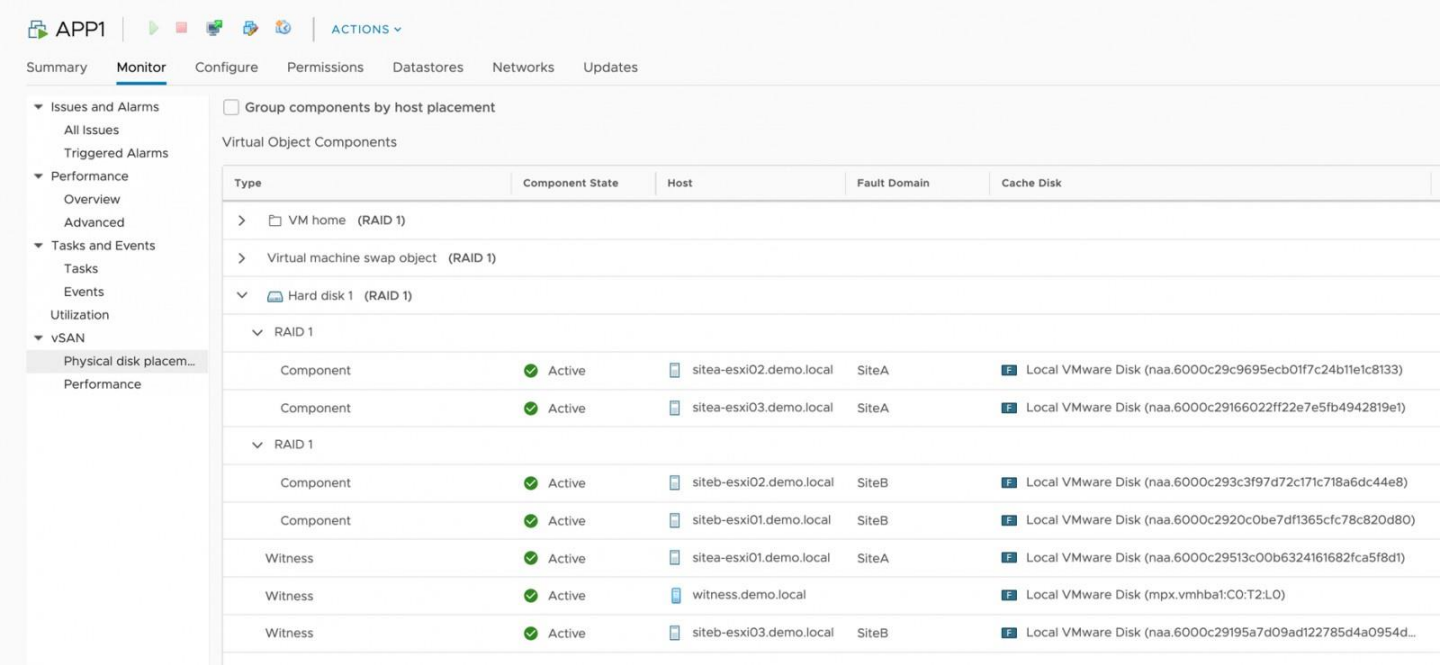


Figure. Viewing the components of an object in a VM.

Recommendation: Simply verify through the UI that the VM is compliant with the assigned storage policy. Data components, and voting mechanisms that determine quorum are an implementation detail of vSAN, and do not need to be understood by an administrator in order to manage a vSAN stretched cluster.

Management and Maintenance

This section covers some basic operational activities for vSAN stretched clusters. The vSAN Operations Guide provides additional information on ongoing operations of vSAN stretched clusters.

Lifecycle Management

The process of patching and upgrading vSAN stretched clusters is very similar to traditional, single site vSAN clusters. The vSphere Lifecycle Manager (vLCM) is now responsible for all lifecycle management duties with the hypervisor, including the core hypervisor, software drivers, and some hardware firmware.

Updating a Cluster

vLCM will not only orchestrate all of the updates for the physical hosts in the data site, and in more recent versions, the virtual witness host appliance in the third site. After vCenter Server is updated to the latest version, simply initiate an update of the cluster using vLCM, and it will roll through the physical hosts in the order that it determines best.

Since vSAN is a distributed storage system, it limits host updates to one host at a time. However, several optimization within vSphere have reduced the number of host restarts. The new architecture with vSAN ESA also speeds up host restart times in clusters running vSAN. The combination of these improvements will help reduce the time it takes to complete a cluster upgrade.

Recommendation: Focus on efficient delivery of services during cluster updates, as opposed to speed of update. vSAN restricts parallel host remediation. Updates using vLCM maintain full serviceability of the cluster while the update is in progress.

If a cluster update results in health check notifications that the on-disk and/or object format needs to be updated, look for an opportunity to perform those updates. On-Disk format upgrades are usually quick and effortless. For more information, see the post "Upgrading On-Disk and Object Formats in vSAN."

Updating the Witness Host Appliance

Prior to vSAN 7 U3, the witness host appliance was not capable of being updated using vLCM. For many, the upgrade process for the witness host appliance involved a manual update of its hypervisor, or simply a redeployment of a witness using a newer version. To complicate matters, in versions prior to vSAN 7 U1, the witness host appliance required that it be upgraded after to the rest of the cluster. Beginning in vSAN 7 U1, the witness host appliance required that it be upgraded prior to the rest of the cluster. This manual method of upgrading the witness host appliance is still supported, but vLCM takes the complexity out of the process by upgrading it for you at the appropriate time.

Healthy State of the Cluster

Skyline Health checks are an ideal way to quickly check the health and well being of a vSAN cluster. You can look for the cluster health score, which categorizes and weights the triggered health checks to help you determine the severity of the condition, as well as prioritizing the most important health checks that have been triggered, and will provide actions for remediation. For more information, see the post: "Skyline Health Scoring, Diagnostics, and Remediation for vSAN 8 U1."

Maintenance of a Site (Fault Domain)

Occasionally there is a requirement to take an entire site offline for the purpose of maintenance. As of vSAN 8 U3 and VCF 5.2 there is no automated process to do this. However, the announcement of VCF 9 includes the ability of a fully automated "Site Maintenance" workflow that will make site maintenance easy. Until VCF 9 becomes available, see the section: "Non-Disruptive Maintenance of One Site in a Stretched Cluster Environment" in the vSAN Operations Guide on how you can perform these steps manually.

Failure Scenarios

vSAN clusters configured as a single site, vSAN HCI cluster can accommodate a wide variety of scenarios. vSAN configured in a stretched cluster topology aims to make data available under an even wider variety of circumstances, but the characteristics of the topology add to the list of potential failure conditions it must account for. An example might be a condition where both sites are up, but one of the data sites cannot connect to either the data site or the witness site. This is a type of "partition" that must be and is accounted for in a vSAN stretched cluster topology.

vSAN also has flexibility in terms of it storing data in a manner that is analogous to an object store. Instead of using a monolithic filesystem, on a storage array, vSAN uses a smaller boundary of data known as objects. These objects may represent aspect of a VM, such as a VMDK. Many times, vSAN can have several discrete failures in a stretched cluster and maintain availability. vSphere clusters using two storage arrays across two sites (known as a metro storage cluster), and be challenging during failure conditions because it will treat any failure as a monolithic failure.

Time that it takes for failures to be recognized

vSAN uses a variety of time-out values to declare when something is unavailable. This is to help account for the wide variety of ways that some aspect of a system can fail. For example, if vSAN receives a SCSI sense code from a device that it has failed, then it will make that as failed, and act immediately, but a host may become unavailable because it actually failed, or there was a transient network error that will correct itself. Therefore, vSAN not only has multiple timeout values to account for this, but even has logic to wait a period of time (60 minutes by default) before the data begins to resynchronize itself to regain its prescribed level of resilience.

For vSAN stretched clusters, a failure of an object spanning both data sites make take about 5-7 seconds prior to declaring the object "degraded" or "absent." This may vary depending on the conditions of the failure. **When a failure is substantial enough that an object is marked as "absent" this is actually a mechanism to protect the integrity of your data.** It helps avoid conditions where insufficient nodes exist to complete the write commands, and to prevent updating the same data in two different locations – known as a split brain. The goal in any recovery of the system should be to get the system up to a condition where it can achieve quorum, and make the data available again.

Failure Scenario Matrices

The following details specific failure scenarios and the observed behavior found in a vSAN stretched cluster. It covers the most likely scenarios, but is not necessarily an exhaustive list.

Scenario	vSAN Behavior	Impact/Observed VM behavior
Cache device failure **(OSA only. Not relevant for ESA)**	Disk Group is marked as failed, and all components present on it will rebuild on another Disk Group.	VM will continue running.
Capacity disk failure (Dedupe and Compression ON) **(OSA only. Not relevant for ESA)**	Disk Group is marked as failed, and all components present on it will rebuild on another Disk Group.	VM will continue running.
Capacity disk failure (Dedupe and Compression OFF**) (OSA only. Not relevant for ESA)**	Disk marked as failed, and all components present on it will rebuild on another disk.	VM will continue running.
Disk Group failure/offline **(OSA only. Not relevant for ESA)**	All components present on the Disk Group will rebuild on another Disk Group.	VM will continue running.
RAID/HBA card **failure (OSA only. Not relevant for ESA)**	All Disk Groups backed by the HBA/RAID card will be marked absent and all components present will rebuild on other Disk Groups.	VM will continue running.
Host failure	Component on the host will be marked as absent by vSAN – component rebuild will be kicked off after 60 minutes if the host does not come back up.	VM will continue running if on another host. If the VM was running on the same host as the failure an HA restart of the VM will take place.
Host isolation	Components present on the host will be marked as absent by vSAN – component rebuilds will be kicked off after 60 minutes if the host does not come back online.	VM will continue running if on another host. If the VM was running on the same host as the failure an HA restart of the VM will take place when the Host Isolation response is accurately configured.
Witness loss / failed / isolated from one or both sites	Witness loss counts as a site (PFTT) failure, as such the cluster will be placed in a degraded state until the witness comes back online or is redeployed.	VM will continue running.
Data site failure or partition, ISL failure / connectivity loss	Site is declared lost, quorum is established between the witness and the remaining data site.	VMs running on the partitioned/failed site are powered off by vSAN. If there is an ISL loss, HA will restart the VMs from the secondary site on the preferred site. If the preferred site has failed, the VMs will be restarted on the secondary site.
Sites are partitioned due to ISL loss and VM(s) with storage policies configured for no site-level resilience	Storage policy places VM components in non-preferred site. No change in quorum, and no component	VM continues running.

and data affinity to the non-preferred site. DRS rules keeping VM instance in non-preferred site.	reconfiguration across sites is required due to no site resilience for the VM(s). DRS rules keep VM instance in non-preferred site	
Loss of both sites connecting to each other or the witness site.	Cluster offline. Quorum of components unable to be met.	VMs stop running. HA cannot restart VMs until quorum is re-established.

Storage policies play an important part in the types and severity of failures. The tables below provide the behavior of various failure conditions when running VMs that use the following storage policy rules:

- Site disaster tolerance: Dual site mirroring (stretched cluster)
- Failures to tolerate: 2 failures – RAID-6

It assumes HA and DRS are enabled with the appropriate DRS Host/VM groups and rules as recommended in this document..

Scenario	vSAN Behavior	Impact/Observed VM behavior
Single site failure	Site marked as failed and rebuild of components will begin when the failed site comes online again. This is also triggered in the event that the witness site is lost as it is viewed as a discrete site.	VMs running on the partitioned/failed site are powered off. HA will restart the VMs from the secondary site on the preferred site. If the preferred site has failed, they will restart on the secondary site.
Single disk, disk group (OSA only), host failure on one site.	All components present will rebuild on their respective fault domain.	Disk and disk group (OSA only) failures will not affect VM running state. VMs will continue running if on a host other than the one that failed. If the VM was running on the failed host an HA restart of the VM will take place.
Dual disk, disk group (OSA only), host failure on one site.	Site marked as failed by vSAN, component rebuilds will begin when the site comes online again.	The site is marked as failed from a vSAN standpoint. VMs will continue running if on a host other than the one that failed. If the VM was running on the failed host an HA restart of the VM will take place.
Single site failure and single disk, disk group (OSA only), host failure across remaining sites.	Site marked as failed, disk/disk group/host also marked as failed. Components present on the failed site will wait for the site to come online again in order to rebuild. Components present on the failed disk/disk group/host will rebuild on their respective fault domain within the same site	Disk and disk group failures will not affect VM running state. VMs will continue running if they are running on a host/site other than the ones that failed. If the VM was on the failed host/site an HA restart of the VM will take place.

Single site failure and dual disk, disk group (OSA only), host failure across remaining sites.	Site marked as failed, both disks/disk groups/hosts also marked as failed. Components present on the failed site will wait for the site to come online again in order to rebuild. Components present on the failed disks/disk groups/hosts will rebuild on their respective fault domain within the same site.	Disk and disk group failures will not affect VM running state. VMs will continue running if they are running on a host/site other than the ones that failed. If the VM was on the failed hosts/site an HA restart of the VM will take place.
Single site failure and triple disk, disk group (OSA only), host failure across remaining sites	Cluster offline. Both sites marked as failed by vSAN. A site will need to come back online to bring vSAN online. This is as a result of the site-level resilience failure (for example the witness) and the dual failure in a remaining site. The policy specifies "FTT=2" which, site-level resilience violation is counted globally across sites due to quorum implications. The cluster can be brought up again by bringing the failed site back online or replacing the failed devices on remaining sites.	VMs will stop running and the cluster will be offline until a site is brought back online. Due to the site resilience and secondary resilience violations (single site failure and triple failure in remaining site) the cluster's objects will have lost quorum and will not be available.
Dual site failure	Cluster offline. A site will need to come back online to bring vSAN back online	VMs will stop running and the cluster will be offline until a site is restored.
Single disk, disk group, host failure on one site and dual disk, disk group (OSA only), host failure on another site.	The site with a dual failure will be marked as failed by vSAN, components residing on the site will need to wait for it to come online to rebuild. The site with the single failure will have its components rebuilt on their respective fault domain within the site.	Disk and disk group (OSA only) failures will not affect VM running state. VMs will continue running if on a host other than the one that failed. If the VM was on the failed host as the failure an HA restart of the VM will take place. VMs running on the failed site are powered off, HA will restart the VMs from the secondary site on the preferred site. If the preferred site has failed, they will restart on the secondary site.
Dual disk, disk group (OSA only), host failure on one site and dual disk, disk group (OSA only), host failure on another site.	Cluster offline. A site will need to come back online to bring vSAN online. This results from the dual secondary level of resilience failures and the	VMs will stop running, and the cluster will be offline until a site is back online.

	policy specifying “FTT=2", after which the site is marked as failed. This can be achieved by replacing the failed devices on either site.	
Component failure of any sort when insufficient fault domains are available for a rebuild	The component out of compliance will not rebuild until adequate failure domains are available.	VM will continue to run as long as the policy has not been violated.

The table below provides another way of reviewing the failure scenarios. For clarity, the table will use the terms of “Site A” configured in vSAN as the “Preferred Site” and “Site B” configured as the “Non-Preferred Site.”

Site Disaster Tolerance	Secondary Failures to Tolerate (FTT)	VM Location	Failure	vSAN Behavior	VM Behavior
None - Preferred	No data redundancy	Site A or B	Host failure in Site A	Objects are inaccessible if the failed host contains on or more components of an object	VM cannot be restarted as the object is inaccessible.
None – Preferred	RAID-1/5/6	Site A or B	Host failure in Site A	Objects are accessible as there is site-local resilience	VM does not need to be restarted unless VM was running on the failed host.
None – Preferred	No data redundancy	Site A	Full failure Site A	Objects are inaccessible as the full site failed	VM cannot be restarted in Site B, as all objects reside in Site A
None – Preferred	No data redundancy	Site B	Full failure Site B	Objects are accessible, as only Site A contains objects	VM can be restarted in Site A, as that is where all objects reside
None – Preferred	No data redundancy	Site A	Partition Site A	Objects are accessible as all objects reside in Site A	VM does not need to be restarted
None – Preferred	No data redundancy	Site B	Partition Site B	Objects are accessible in Site A. Objects are not accessible in Site B as the network is down.	VM is restarted in Site A and powered off by vSAN in Site B

None – Non-Preferred	No data redundancy	Site B	Partition Site B	Objects are accessible in Site B	VM resides in Site B, does not need to be restarted
None – Preferred	No data redundancy	Site A	Witness host failure	No impact, witness host is not used as data is not replicated	No impact
None – Non-Preferred	No data redundancy	Site B	Witness host failure	No impact, the witness host is not used as data is not replicated	No impact
Site Mirroring	No data redundancy	Site A or B	Host failure Site A or B	Components on failed hosts are inaccessible, read and write IO across ISL without local redundancy and rebuild across ISL.	VM does not need to be restarted unless VM was running on the failed host
Site Mirroring	RAID-1/5/6	Site A or B	Host failure Site A or B	Components on failed hosts are inaccessible. Read IO locally due to RAID, and rebuild locally.	VM does not need to be restarted unless VM was running on failed host
Site Mirroring	No data redundancy	Site A	Full failure Site A	Objects are inaccessible in Site A as full site failed	VM restarted in Site B
Site Mirroring	No data redundancy	Site A	Partition Site A	Objects are inaccessible in Site A as the full site is partitioned, and the quorum is lost.	VM restarted in Site B
Site Mirroring	No data redundancy	Site A	Witness host failure	Witness object inaccessible, VM remains accessible	VM does not need to be restarted
Site Mirroring	No data redundancy	Site B	Full failure Site A	Objects are inaccessible in Site A as the full site failed	VM does not need to be restarted as it resides in Site B

Site Mirroring	No data redundancy	Site B	Partition Site A	Objects are inaccessible in Site A as the full site is partitioned, and the quorum is lost.	VM does not need to be restarted as it resides in Site B
Site Mirroring	No data redundancy	Site B	Witness host failure	Witness object inaccessible, VM remains accessible	VM does not need to be restarted
Site Mirroring	No data redundancy	Site A	Network failure between Site A and B (ISL down)	Site A binds with the witness, and objects in Site B become inaccessible	VM does not need to be restarted
Site Mirroring	No data redundancy	Site B	Network failure between Site A and B (ISL down)	Site A binds with the witness, and objects in Site B become inaccessible	VM restarted in Site A
Site Mirroring	No data redundancy	Site A or B	Network failure between Witness and Site A/B	Witness object inaccessible, VM remains accessible	VM does not need to be restarted
Site Mirroring	No data redundancy	Site A	Full failure Site A and simultaneous Witness Host Failure	Objects are inaccessible in Site A and Site B due to quorum being lost	VM cannot be restarted.
Site Mirroring	No data redundancy	Site A	Full failure Site A followed by Witness Host Failure a few minutes later	Pre vSAN 7.0 U3: Objects are inaccessible in Site A and Site B due to quorum being lost	VM cannot be restarted.
Site Mirroring	No data redundancy	Site A	Full failure Site A followed by Witness Host Failure a few minutes later	Post vSAN 7.0 U3: Objects are inaccessible in Site A, but accessible in Site B as votes have been recounted	VM restarted in Site B
Site Mirroring	No data redundancy	Site B	Full failure Site B followed by Witness	Post vSAN 7.0 U3: Objects are inaccessible in	VM restarted in Site A

			Host Failure a few minutes later	Site B, but accessible in Site A as votes have been recounted	
Site Mirroring	No data redundancy	Site A	Full failure in Site A and simultaneous host failure in Site B	Objects are inaccessible in Site A. If components reside on the failed host then the object is inaccessible in Site B	VM cannot be restarted
Site Mirroring	No data redundancy	Site A	Full failure in Site A and simultaneous host failure in Site B	Objects are inaccessible in Site A. If components do not reside on the failed host, then the object is accessible in Site B	VM restarted in Site B
Site Mirroring	No data redundancy	Site A	Full failure in Site A and simultaneous host failure in Site B	Objects are inaccessible in Site A, accessible in Site B as there's site-local resiliency	VM restarted in Site B

Adaptive Quorum Control

As noted in a few of the failure scenarios listed above, vSAN 7 U3 introduced new failure handling techniques to improve data availability during failure conditions. Adaptive Quorum Control (AQC) is a technique that maintains data availability of objects during a site failure (or maintenance) followed by subsequent unavailability of the witness host. In a fully operational stretched cluster, quorum (determining availability of an object) is the result of account for object components in both sites and the witness host appliance. This is achieved through a simple voting mechanism. With this feature, when a data site has a planned or unplanned outage, vSAN will adjust the votes to favor the active site that still has quorum. This will allow sufficient votes to maintain quorum, which will keep the data available during times of a planned or unplanned outage of the witness host appliance. Depending on the size of the cluster, it may take a few seconds to a few minutes to adjust all of the component votes. As it completes each object, then that object is able to tolerate the failure of a witness host and still maintain availability. This capability will not protect against a simultaneous double failure of a data site and a witness.

Recovery from a Double Failure

In conditions of a double site failure, where one data site fails at the same time as a witness site, the data and the VMs served will be unavailable as they do not achieve quorum. As mentioned previously, this is a protection mechanism to prevent updating like data in two different locations. There may be a need to recover the data in the one remaining site in the even that you know the other data site and the witness site are not coming back. For all versions up to and including vSAN 8 U3 (VCF 5.2), this means contacting Global Support (GS) and they can assist in manually recovering the object data. The

announcement of VCF 9 includes a new “Site Takeover” capability that will allow for this action to be performed by the administrator. More information will be shared as VCF 9 becomes available.

Support Statements

vSAN Storage Clusters and its Support of in a Stretched Cluster Configuration

As of vSAN 8 U3, and VCF 5.2, there is limited support of vSAN storage clusters when deployed at a stretched cluster. For more details, see the post: “Flexible Topologies with vSAN Max.”

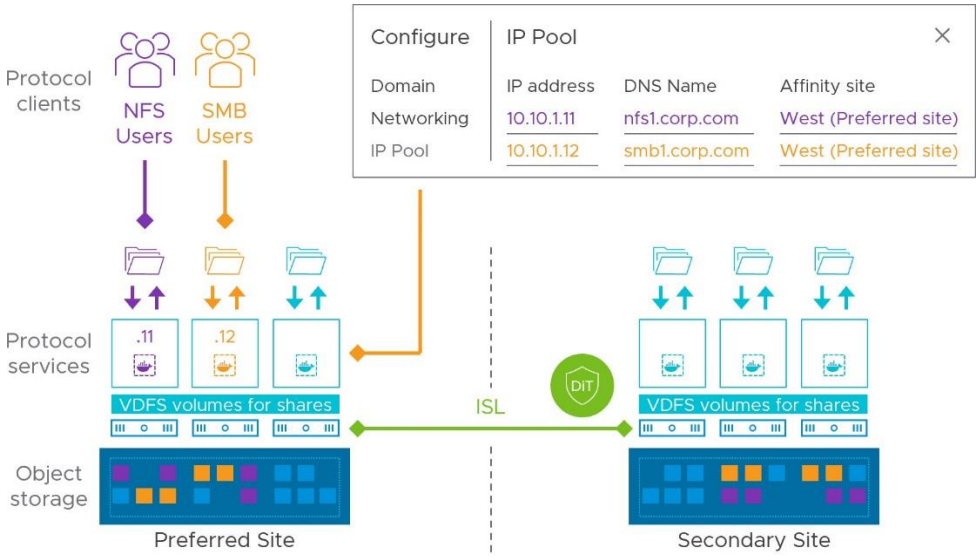
vSAN File services support for vSAN Stretched cluster

File services is supported in vSAN stretched clusters. Even if the file share is protected across sites via site-level protection, the file server is a single entity responsible for the connection from the client system. Thus, one would want this to be on the same site as the file server connection to provide an optimal data path. This mechanism will maintain the colocation of the client to the protocol services used, the VDFS proxy, VDFS server, and at least one of the backing vSAN objects.

The options for the affinity site sets are Preferred, Non-Preferred, or Either.

The site affinity setting for a file share is defining where the presentation layer (NFS or SMB services) resides. It does not relate to if or how the data is placed across sites. This is defined by the storage policy associated with the share. The storage policy settings will be able to protect the share data in the same manner as any other type of object data, such as no site-level mirroring, site-level mirroring, and site-level mirroring with secondary levels of protection at each site, including RAID-1 and RAID-5/6.

When a site hosting an SMB share fails, a failover to the alternate site will occur as expected. When the recovery of the site failure of an SMB file share occurs, the failback of the SMB file share to the desired site will not be automatic. A rebalancing to regain compliance with site affinity will only occur due to the rebalance health check action.



vSAN Stretched Cluster

For standard vSAN clusters, the minimum number of hosts needed to provide file services in a non-error state is three hosts, but it can still provide file services in an error state with just two hosts. The minimum number of hosts needed to provide file services in a non-error state is 2. The minimum host count described here reflects the minimum number of “file service VMs” or “FSVMs” needed to provide the protocol services on each host in a non-error state.

Summary

vSAN stretched clusters provide an incredibly powerful and resilient topology for your mission critical workloads. Following the design and operational recommendations in this document will ensure that your experience with vSAN stretched clusters is successful for you and your organization.

Additional Resources

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

vSAN Interactive Infographic. This tool allows you to dynamically select deployment and failure scenarios to better understand how vSAN maintains availability and recovers from failure.

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.