



vSAN Stretched Cluster Guide

VMware Storage

Table of contents

| | |
|---|----|
| vSAN Stretched Cluster Guide | 6 |
| Overview | 6 |
| Introduction | 6 |
| vSAN Stretched Cluster Concepts | 7 |
| vSAN Stretched Clusters vs. Fault Domains | 7 |
| The vSAN Witness Host | 7 |
| Read Locality in vSAN Stretched Clusters. | 7 |
| Witness Traffic Separation (WTS) | 7 |
| Per Site Policies | 8 |
| Site disaster tolerance | 10 |
| Failures to tolerate | 10 |
| vSAN File services support for vSAN Stretched cluster | 11 |
| Support Statements | 12 |
| vSphere Versions | 12 |
| vSphere Distributed Resource Scheduler (DRS) | 12 |
| vSphere Availability (HA) | 12 |
| On-disk Formats | 12 |
| vSAN Witness Host | 12 |
| Hybrid and All-Flash Support | 12 |
| vSAN 8 using the Express Storage Architecture (ESA) | 12 |
| Requirements | 13 |
| VMware vCenter Server | 13 |
| A Witness Host | 13 |
| Networking Requirements | 14 |
| Layer 2 and Layer 3 Support | 14 |
| vSAN Witness Host Networking | 15 |
| Supported Geographical Distances | 15 |
| Bandwidth and Latency Requirements | 15 |
| Data Site to Data Site Network Latency | 15 |
| Data Site to Data Site Bandwidth | 15 |
| Data Site to Witness Network Latency | 15 |
| Data Site to Witness Network Bandwidth | 16 |
| Support for Mixed MTU for Witness Traffic separation | 16 |
| Multiple vSAN Witness Hosts sharing the same VLAN | 18 |
| Configuration Minimums and Maximums | 19 |

| | |
|--|----|
| Virtual Machines Per Host | 19 |
| Hosts Per Cluster | 19 |
| Minimum Host Count | 19 |
| Maximum Host Count | 19 |
| Symmetrical Configurations | 19 |
| Asymmetrical Configurations | 19 |
| Witness Host | 19 |
| vSAN Storage Policies | 19 |
| Primary Number of Failures To Tolerate (PFTT) | 19 |
| Secondary Number of Failures To Tolerate (SFTT) | 19 |
| Failure Tolerance Method (FTM) | 19 |
| Affinity | 20 |
| Other Policy Rules | 20 |
| Fault Domains | 20 |
| Design Considerations | 21 |
| Witness Host Sizing | 21 |
| vSAN Witness Appliance (Virtual Machine) | 21 |
| Licensing | 21 |
| vSAN Witness Appliance Version | 21 |
| vSAN Witness Appliance Size | 21 |
| Compute Requirements | 21 |
| Memory Requirements | 21 |
| Storage Requirements | 21 |
| vSAN Witness Appliance Deployment Sizes & Requirements Summary | 21 |
| Where can the vSAN Witness Appliance run? | 22 |
| Physical Host as a vSAN Witness Host | 24 |
| Cluster Compute Resource Utilization | 25 |
| Network Design Considerations | 26 |
| Stretched Cluster Network Design Considerations | 26 |
| Connectivity and Network Types | 26 |
| Port Requirements | 26 |
| Configuration of the Network from the Data Sites to the Witness | 29 |
| Option 1: Physical vSAN Witness Host connected over L3 & static routes | 29 |
| Option 2: Virtual vSAN Witness Host connected over L3 & static routes | 30 |
| Bandwidth Calculation | 32 |
| Requirements Between Data Sites | 32 |
| Requirements when Read Locality is not Available | 32 |
| Requirements Between Data Sites and the Witness Site | 33 |

| | |
|--|----|
| Witness to Site Examples | 33 |
| The Role of vSAN Heartbeats | 34 |
| Host number calculation | 34 |
| Cluster Settings - vSphere HA | 35 |
| Turn on vSphere HA | 35 |
| Host Monitoring | 36 |
| Virtual Machine Response for Host Isolation | 37 |
| Admission Control | 38 |
| Host Hardware Monitoring - VM Component Protection | 39 |
| Datastore for Heartbeating | 40 |
| Advanced Options | 40 |
| Cluster Settings - DRS | 42 |
| Partially Automated or Fully Automated DRS | 43 |
| VM/Host Groups & Rules | 44 |
| Host Groups | 44 |
| VM Groups | 45 |
| VM/Host Rules | 46 |
| Per-Site Policy Rule Considerations | 47 |
| Installation | 51 |
| Before You Start | 51 |
| What is a Preferred Site? | 51 |
| What is Read Locality? | 51 |
| vSAN Health Check Plugin & Stretched Clusters | 52 |
| Details and setup of the vSAN Witness Appliance | 54 |
| vSAN Witness Appliance Details | 54 |
| Networking | 54 |
| A Note About Promiscuous Mode | 54 |
| Setup Step 1: Deploy the vSAN Witness Appliance | 55 |
| Setup Step 2: vSAN Witness Appliance Management | 65 |
| Setup Step 3: Add Witness to vCenter Server | 67 |
| Setup Step 4: Config vSAN Witness Host Networking | 72 |
| Configure the network address. | 74 |
| Setup Step 5: Validate Networking | 76 |
| Default Gateways and Static Routes | 76 |
| Configuring a vSAN Stretched Cluster | 80 |
| Creating a New vSAN Stretched Cluster | 80 |
| Create a Cluster using the Cluster QuickStart | 80 |

Converting a Cluster to a Stretched Cluster 88

Set vSphere HA Advanced Settings 91

Verifying vSAN Stretched Cluster Component Layouts 102

Upgrading an older vSAN Stretched Cluster 103

Management and Maintenance 107

 Maintenance Mode Consideration 107

 Monitoring 107

 Updates using vLCM 107

Failure Scenarios 108

 Failure Scenarios and Component Placement 108

 Individual Host Failure or Network Isolation 108

 Individual Drive Failure 109

 Site Failure or Network Partitions 110

 Multiple Simultaneous Failures 119

 Improved resilience for simultaneous site failures 121

 Recovering from a Complete Site Failure 123

 How Read Locality is Established After Failover 123

 Replacing a Failed Witness Host 123

 VM Provisioning When a Site is Down 127

 Efficient inter-site resync for stretched clusters 129

 Failure Scenario Matrices 129

Additional Resources 134

 Links to other vSAN resources 134

 Location of the vSAN Witness Appliance OVA 134

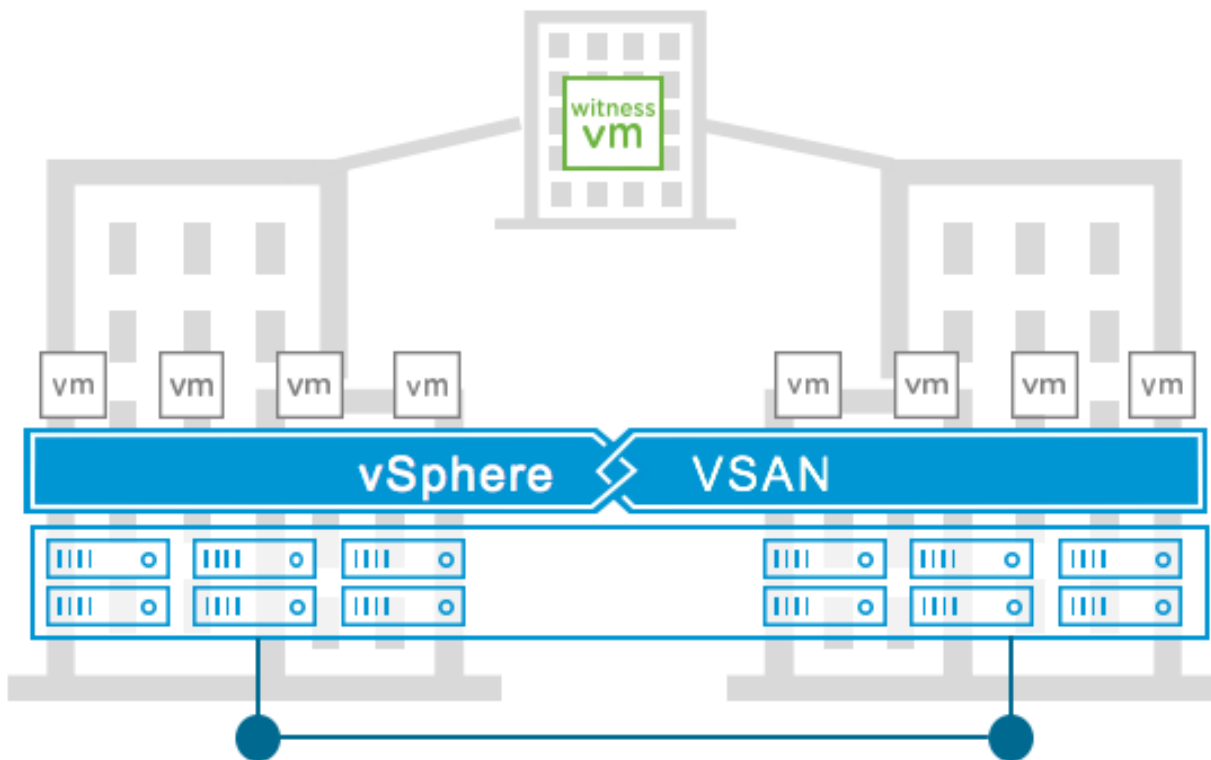
vSAN Stretched Cluster Guide

Overview

Introduction

A vSAN Stretched Cluster is a specific configuration implemented in environments where disaster/downtime avoidance is crucial. This guide was developed to provide additional insight and information for installing, configuration, and operating a vSAN Stretched Cluster infrastructure in conjunction with VMware vSphere. This guide will explain how vSAN Stretched Clusters handle specific failure scenarios and discuss various design considerations and operational procedures for Stretched Clusters using vSAN.

VMware vSAN Stretched Clusters refer to a deployment where a user sets up a vSAN cluster with two active/active sites with an equal number of vSphere hosts distributed evenly between the two sites. The sites are connected via a high bandwidth/low latency link. The third site hosting, the vSAN Witness Host, is connected to both data sites. This connectivity can be via low bandwidth/high latency links.



Each site is configured as a vSAN Fault Domain. The nomenclature used to describe a vSAN Stretched Cluster configuration is $X+Y+Z$, where X is the number of ESXi hosts at data site A, Y is the number of ESXi hosts at data site B, and Z is the number of witness hosts at site C. Data sites are where virtual machines are deployed. The minimum supported configuration is $1+1+1$ (3 nodes). The maximum configuration is $20+20+1$ (41 nodes). In vSAN Stretched Clusters, only one witness host exists in any configuration.

A virtual machine deployed on a vSAN Stretched Cluster will have one copy of its data on site A, a second copy on site B and any witness components placed on the witness host at site C. This configuration is achieved through fault domains, hosts and VM groups, and affinity rules. In the event of a single complete site failure, there will be a full copy of the virtual machine data and greater than 50% of the components available. This will allow the virtual machine to remain available on the vSAN datastore. If the virtual machine needs to be restarted on the other site, vSphere HA will handle this task.

vSAN Stretched Cluster Concepts

vSAN Stretched Clusters vs. Fault Domains

Fault domains enable what might be termed “rack awareness,” where the components of virtual machines can be distributed across multiple hosts in multiple racks. The virtual machine would remain available if a rack failure occurred. However, these racks would typically be hosted in the same data center. If there was a data center-wide event, fault domains could not assist with virtual machines' availability.

Stretched clusters essentially build on what fault domains did and now provide what might be termed “data center awareness.” VMware vSAN Stretched Clusters can now provide availability for virtual machines even if a data center has a catastrophic outage.

The vSAN Witness Host

The witness host is a dedicated ESXi host, or vSAN Witness Appliance, whose purpose is to host the witness component of virtual machine objects.

The vSAN Witness Appliance must have a connection to both the primary vSAN node and the backup vSAN node to join the cluster. In steady-state operations, the primary node resides in the “Preferred site”; the backup node resides in the “Secondary site.” Unless the witness host connects to the primary and the backup nodes, it will not join the vSAN cluster.

vSAN Witness Connectivity

The vSAN Witness Host must be managed by the same vCenter Server managing the vSAN Cluster.

There must be connectivity between vCenter Server and the vSAN Witness Host in the same fashion as vCenter controlling other vSphere hosts.

The vSAN Witness Host must also have connectivity between the vSAN Witness Host and the vSAN nodes. This is typically performed through connectivity between the vSAN Witness Host vSAN VMkernel interface and the vSAN data network.

These will be covered more thoroughly in a later section.

Updating or Upgrading the vSAN Witness Appliance

The vSAN Witness Appliance can easily be maintained and patched using the same methods as traditional ESXi hosts.

Deploying a new vSAN Witness Appliance is not required when updating or patching vSAN hosts. Standard upgrade mechanisms are supported on the vSAN Witness Appliance. If needed, it is easy to deploy a new vSAN Witness Appliance.

Note: When using an OEM-provided vSphere ISO for upgrading vSAN hosts, it is essential to remember that additional OEM specific drivers or software may be included. Using a VMware-provided vSphere ISO is vital to upgrade the vSAN Witness Appliance.

Read Locality in vSAN Stretched Clusters.

In traditional vSAN clusters, a virtual machine's read operations are distributed across all replica copies of the data in the cluster. In the case of a policy setting of `NumberOfFailuresToTolerate = 1`, which results in two copies of the data, 50% of the reads will come from replica 1 and 50% from replica 2. In the case of a policy setting of `NumberOfFailuresToTolerate = 2` in non-stretched vSAN clusters, it results in three copies of the data, 33% of the reads will come from replica 1, 33% of the reads will come from replica 2, and 33% will come from replica 3.

In a vSAN Stretched Cluster, we wish to avoid increased latency caused by reading across the inter-site link. The read locality mechanism was introduced to ensure that 100% of reads occur on the site the VM resides on. Read locality overrides the `NumberOfFailuresToTolerate = 1` policy's behavior to distribute reads across the components.

Another consideration with read locality is avoiding unnecessary virtual machine vMotion operations between sites. Since the read cache blocks are stored on one site, the cache will be cold after the move if the VM moves around freely and ends up on the remote site. (Note that this only applies to hybrid configurations, as all-flash configurations do not have an explicit read cache.) Now there will be sub-optimal performance until the cache is warm again. To avoid this situation, soft affinity rules keep the VM local to the same site/fault domain where possible. The steps to configure such rules will be shown in detail in the vSphere DRS section of this guide.

Witness Traffic Separation (WTS)

By default, when using vSAN Stretched Clusters, the Witness VMkernel interface tagged for vSAN traffic must have connectivity with each vSAN data node's VMkernel interface tagged with vSAN traffic.

Witness Traffic Separation is supported on Stretched Cluster configurations. An alternate VMkernel interface can be designated to

carry traffic destined for the Witness rather than the vSAN-tagged interface. This feature allows for more flexible network configurations by allowing for separate networks for node-to-node and node-to-witness traffic.

Mixed MTU (Jumbo Frames)

In addition to WTS, mixed MTU sizes (e.g., 9000 for vSAN data and 1500 for vSAN Witness traffic) are supported.

Per Site Policies

Per-site policy rules provide additional protection or flexibility for Stretched Cluster scenarios.

Primary Failures To Tolerate determines whether an object is protected on each site or only on a single site. A value of 1 enables protection across sites. A value of 0 keeps data at one or the other site only.

Secondary Failures To Tolerate defines the number of disk or host failures a storage object within a site can tolerate.

Affinity is only applicable when Primary Failures To Tolerate is 0. When Primary Failures To Tolerate is 0, this rule allows the administrator to choose which site the vSAN object should reside on, either the Preferred or Secondary Fault Domain.

Data access behavior using the new Policy Rules

vSAN Stretched Clusters have traditionally written a copy of data to each site using a Mirroring Failure Tolerance Method. These were complete writes to each site, with reads being handled locally using the Site Affinity feature. Write operations depend on VM Storage Policy rules.

Dual Site Mirroring/Primary Failures To Tolerate Behavior

When a Primary Failures to Tolerate rule equals 1, writes will continue to be written in a mirrored fashion across sites. When a Primary Failures to Tolerate rule equals 0, writes will only occur in the site specified in the Affinity rule. Reads continue from the site a VM resides on.

Local Protection/Secondary Failures To Tolerate Behavior

When a Secondary Failures-to-Tolerate rule is in place, the behavior within a site adheres to the Failure Tolerance Method rule. The number of failures to tolerate and the Failure Tolerance Method determine how many hosts are required per site to satisfy the rule requirements.

Writes and reads occur within each site in the same fashion as in a traditional vSAN cluster, but per site.

Data will be fetched from the alternate site only when data cannot be repaired locally, such as in cases where the only present copies of data reside on the alternate site.

Affinity

The Affinity rule only specifies which site a vSAN object, Preferred or Secondary, will reside on. It is only honored when a Primary Failures To Tolerate rule is set to 0. VMware recommends that virtual machines run on the same site their vSAN objects reside on.

Because the Affinity rule is a Storage Policy rule, it only pertains to vSAN objects, not virtual machine placement. This is because read and write operations will be required to traverse the inter-site link when the virtual machine and vSAN objects do not reside on the same site.

vSAN Stretched Cluster Capacity Sizing when using Per-Site Policy Rules

With Per-Site Policy Rules, capacity requirements can change entirely based on Policy Rule requirements.

| Availability | PFTT | SFTT | Capacity Required in Preferred Site in GB | Capacity Required in Secondary Site in GB | Capacity Requirement |
|---|------|------|---|---|----------------------|
| Dual Site Mirroring without Redundancy | 1 | 0 | 100 | 100 | 2x |
| Dual Site Mirroring with RAID1 (1 Failure) | 1 | 1 | 200 | 200 | 4x |
| Dual Site Mirroring with RAID1 (2 Failures) | 1 | 2 | 300 | 300 | 6x |
| Dual Site Mirroring with RAID1 (3 Failures) | 1 | 3 | 400 | 400 | 8x |
| Dual Site Mirroring with RAID5 (1 Failure) | 1 | 1 | 133 | 133 | 2.66x |
| Dual Site Mirroring with RAID6 (2 Failures) | 1 | 2 | 150 | 150 | 3x |
| Preferred Site Only RAID1 (1 Failure) | 0 | 1 | 200 | 0 | 2x |
| Preferred Site Only RAID1 (2 Failures) | 0 | 2 | 300 | 0 | 3x |
| Preferred Site Only RAID1 (3 Failures) | 0 | 3 | 400 | 0 | 4x |
| Preferred Site Only RAID5 (1 Failure) | 0 | 1 | 133 | 0 | 1.3x |
| Preferred Site Only RAID6 (2 Failures) | 0 | 2 | 150 | 0 | 1.5x |
| Non-Preferred Site Only RAID1 (1 Failure) | 0 | 1 | 0 | 200 | 2x |
| Non-Preferred Site Only RAID1 (2 Failures) | 0 | 2 | 0 | 300 | 3x |
| Non-Preferred Site Only RAID1 (3 Failures) | 0 | 3 | 0 | 400 | 4x |
| Non-Preferred Site Only RAID5 (1 Failure) | 0 | 1 | 0 | 133 | 1.3x |
| Non-Preferred Site Only RAID6 (2 Failures) | 0 | 2 | 0 | 150 | 1.5x |

vSAN Stretched Cluster Component Examples when using Per-Site Policy Rules

As Per-Site Policy Rules add local protection, objects are distributed into even more components.

Because the bandwidth requirements of the Witness Host are based on the number of components, using these policy rules will increase the overall component count.

Additional objects on vSAN, like snapshots, will add to the overall component count.

vSAN Stretched Cluster Witness Bandwidth considerations when using Per-Site Policy Rules

The witness bandwidth requirement is 2Mbps for every 1000 components. Using this formula, here are a couple examples:

200 virtual machines with 500GB vmdks using Policy Rules for Cross Site protection with local Mirroring would require:

- 7 for swap, 7 for VM home space, 18 for vmdks = 32
- 32 components X 200 VMs = 6,400 components 2Mbps for every 1000 is 6.4 X 2Mbps = **12.8Mbps**

The same 200 virtual machines with 500GB vmdks using Policy Rules for Cross Site protection with local Erasure Coding would require:

- 9 for swap, 9 for VM home space, 17 for vmdks = 35
- 35 components X 200 VMs = 7,000 components 2Mbps for every 1000 is 7 X 2Mbps = 14Mbps

These examples show that adding local protection increases component counts and witness bandwidth requirements.

Site disaster tolerance

This policy setting determines if data is mirrored between the two sites or which side it will be placed on if not mirrored.

None - Standard Cluster - This setting is used when not using a stretched cluster.

Dual site mirroring - This setting allows the loss of an entire data center site by mirroring the data to both sites. This is what most people think of with a stretched cluster.

vSAN

| Availability | Storage rules | Advanced Policy Rules | Tags |
|---------------------------|---------------|---|------|
| Site disaster tolerance ⓘ | | Dual site mirroring (stretched cluster) ▾ | |
| Failures to tolerate ⓘ | | 1 failure - RAID-1 (Mirroring) ▾ | |
| | | Consumed storage space for 100 GB VM disk would be 400 GB | |

The following two settings are often used for applications that handle their replication. An example would be setting one virtual machine to be pinned to one side and another to the second. Typical applications that can cluster themselves include Exchange DAG, SQL AlwaysOn, Oracle Dataguard, etc.

- **None - keep data on preferred.** This setting is helpful if you want to "pin" the data to the preferred site.
- **None - keep data on Non-preferred.** This setting pins data to the opposite site.

vSAN

| Availability | Storage rules | Advanced Policy Rules | Tags |
|---------------------------|---------------|---|------|
| Site disaster tolerance ⓘ | | None - keep data on Non-preferred (stretched cluster) ▾ | |
| Failures to tolerate ⓘ | | 1 failure - RAID-1 (Mirroring) ▾ | |
| | | Consumed storage space for 100 GB VM disk would be 200 GB | |

None - Stretched cluster. This option will keep the virtual machine components on a single side of the stretched cluster but will do so with a random distribution. The vSAN object manager will try to keep them on the same site, but beyond that, it will primarily look at the free capacity within a site and use that as its guiding logic to randomly place virtual machines across sites. This setting might make sense for a larger quantity of applications that lack the need for survivability of loss of either data center. Most people build stretched clusters explicitly to protect most workloads from data center failure. Still, this policy may make sense for a set of capacity-hungry but lower-priority virtual machines.

Failures to tolerate

This policy specifies the data protection that will happen within a site. Note that this raid protection level will be mirrored on both sites if a stretched cluster policy was previously selected. In this example, RAID-1 mirroring would create four copies of data (2 at each site).

Do note that the minimum number of hosts to reach the protection level (4 hosts for RAID 5, 6 hosts for RAID 6) must be maintained within each site.

vSAN File services support for vSAN Stretched cluster

File services can now be used in vSAN stretched clusters and vSAN 2-Node topologies, making it ideal for those edge locations also needing a file server.

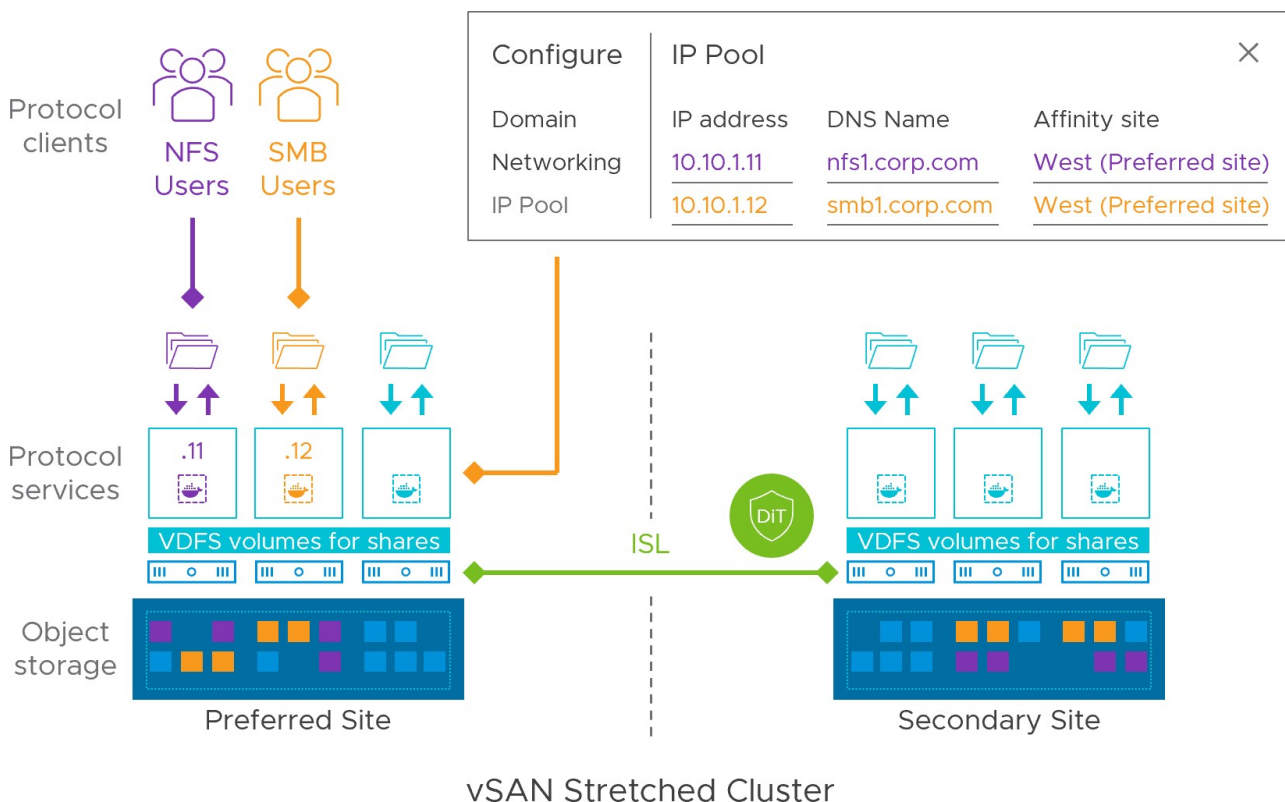
Even if the file share is protected across sites via site-level protection, the file server is a single entity responsible for the connection from the client system. Thus, one would want this to be on the same site as the file server connection to provide an optimal data path. This mechanism will maintain the collocation 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, Secondary, 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.

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.



Support Statements

vSphere Versions

Advanced feature support in vSAN Stretched Clusters requires a combination of vSAN version, On-Disk format version, architecture, and host count. The table below states the minimum requirements for each feature. For more detailed information, please refer to [KB 2148493](#).

| Feature | Minimum requirements |
|------------------------------------|--|
| Datastore Sharing | vSAN 8 Update 1 is required. This feature is OSA only |
| Express Storage Architecture (ESA) | vSAN 8 Update 1 is required. |

The latest on-disk format is generally recommended per version of vSAN.

vSphere Distributed Resource Scheduler (DRS)

For vSAN, vSphere DRS is very desirable. DRS will provide initial placement assistance, load balances the environment when there's an imbalance, and automatically migrate virtual machines to their correct site following VM/Host affinity rules. It can also help migrate virtual machines back to a node after recovery based on overall utilization. Otherwise, the administrator will have to carry out these tasks manually.

vSphere DRS is only available in vSphere Enterprise+ or higher editions.

vSphere Enterprise for ROBO provides a DRS-Lite functionality. While the full capabilities of proactive workload migration are not included in vSphere Enterprise for ROBO, it does include the ability to automatically migrate VMs off of hosts when placing hosts in maintenance mode, such as when performing upgrades.

vSphere Availability (HA)

As in any environment, vSphere HA is desirable for use with vSAN. HA will restart virtual machines when a host has failed. Additionally, when a vSAN node has failed, vSphere HA will restart virtual machines on an alternate host. When a vSAN node becomes isolated, vSAN will power off virtual machines but will not restart them. vSphere HA is used to restart these virtual machines on hosts that have not been isolated.

Some additional settings are required for vSphere HA to work correctly when used with vSAN. These settings are covered in detail later in this guide.

On-disk Formats

VMware recommends upgrading to the latest On-Disk format for improved performance, scalability, and feature capabilities.

vSAN Witness Host

Both physical ESXi hosts and vSAN Witness Appliances (nested ESXi) are supported as a Stretched Cluster Witness Host.

VMware provides a vSAN Witness Appliance for those customers who do not wish to use a physical host for this role. The vSAN Witness Appliance can run on an ESXi Free licensed host, a vSphere licensed (ESXi) host, a host residing in OVH (formerly vCloud Air), a vCloud Air Network (VCAN) partner, or any hosted ESXi installation.

Hybrid and All-Flash Support

VMware vSAN Stretched Clusters are supported on both Hybrid configurations (hosts with local storage comprised of magnetic disks for capacity and flash devices for cache) and All-Flash configurations (hosts with local storage made up of flash devices for capacity and flash devices for cache).

vSAN 8 using the Express Storage Architecture (ESA)

vSAN 8 using the Express Storage Architecture (ESA) supports all stretched cluster features and enhancements. Improved uptime for Stretched clusters is supported.

Requirements

Here is a list of requirements for implementing a vSAN Stretched Cluster.

VMware vCenter Server

A vSAN Stretched Cluster configuration can be created and managed by a single instance of VMware vCenter Server.

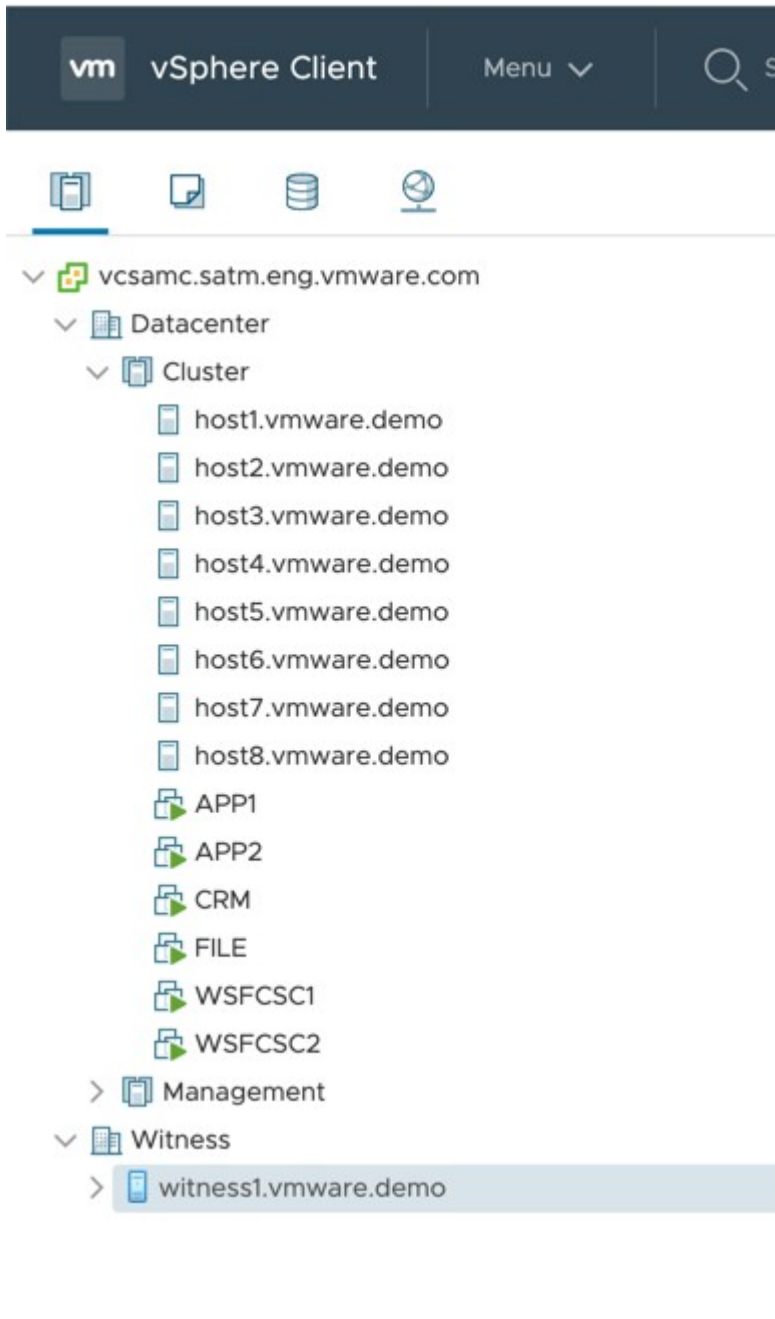
A Witness Host

In a vSAN Stretched Cluster, the witness components are only ever placed on the Witness host. A physical ESXi host or a special vSAN Witness Appliance provided by VMware can be the witness host.

If a vSAN Witness Appliance is used for the Witness host, it will not consume any of the customer's vSphere licenses. A physical ESXi host used as a witness host must be licensed accordingly, as this can still be used to provision virtual machines should a customer choose to do so.

It is essential that the witness host is not added to the vSAN cluster. The witness host is selected during the creation of a vSAN Stretched Cluster.

The witness appliance will have a unique identifier in the vSphere web client UI to assist with identifying that a host is a witness appliance (ESXi in a VM). It is shown as a "blue" host, as highlighted below:



Note: This is only visible when the appliance ESXi witness is deployed. If a physical host is used as the witness, then it does not change its appearance in the web client. A dedicated witness host is required for each Stretched Cluster.

Networking Requirements

Specific networking requirements must be adhered to when vSAN is deployed in a Stretched Cluster across multiple sites using Fault Domains.

Layer 2 and Layer 3 Support

All vSAN stretched cluster network traffic between the data sites and vSAN Witness Host is unicast.

When deciding on a network topology, it is crucial to consider that vSAN uses the same TCP stack as the Management VMkernel interface (typically vmk0). As a result, vSAN VMkernel interfaces use the same default gateway as the Management VMkernel interface. The default gateway for the VMkernel adapter can be overridden to provide a different gateway for the vSAN network. This feature simplifies routing configuration that previously required manual configuration of static routes.

- VMware recommends that vSAN communication between the Data Sites and the Witness Site is routed over Layer 3.
 - If using a traditional Stretched Cluster configuration, the Data Nodes will require a static route from the vSAN

VMkernel interfaces to the vSAN Witness Host VMkernel interface tagged for vSAN Traffic.

- When using Witness Traffic Separation, if a VMkernel interface other than the Management VMkernel interface, which is typically vmk0, is tagged with "witness" traffic, 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.
- Note: vSAN metadata traffic can use a dedicated interface rather than the Management VMkernel interface for isolation. If security concerns dictate, isolating Management traffic from workload traffic is supported. Choosing the Management VMkernel interface for vSAN metadata traffic is supported but should align with desired architecture & risk considerations.
- VMware supports vSAN communication between the Data Sites in either stretched Layer 2 or Layer 3, with the following considerations:
 - Stretched Layer 2 does not require static routing
 - Layer 3 requires static routing to communicate correctly between sites.

vSAN Witness Host Networking

The Management (vmk0) and WitnessPg (vmk1) VMkernel interfaces on the vSAN Witness Host must not be configured to use addresses on the same subnet. Doing so creates a Multi-Homing situation, referenced in KB 2010877.

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.

The use of Network Address Translation (NAT) is not supported with the vSAN Witness Host

The configuration of static routing mentioned above is not presented in the vSphere Client. Static routing can be configured using various methods, including:

- Using `esxcli` from the command line of each vSphere host or the vSphere CLI
- Using a PowerCLI script to configure static routes on one or more hosts
- Configuring static routing using Host Profiles

When deploying vSAN Stretched Clusters, each of these items is important to consider and will help in deciding which network topology to use with vSAN Stretched Clusters.

Supported Geographical Distances

For VMware vSAN Stretched Clusters, geographical distances are not a support concern. The critical requirement is the actual latency numbers between sites.

Bandwidth and Latency Requirements

Data Site to Data Site Network Latency

Data site-to-data site network refers to the communication between non-witness sites, such as sites that run virtual machines and hold virtual machine data. Latency or RTT (Round Trip Time) between sites hosting virtual machine objects should not be greater than 5 msec (< 2.5 msec one-way).

Data Site to Data Site Bandwidth

Bandwidth between sites hosting virtual machine objects will be workload dependent. For specific minimum requirements, see the vSAN Documentation for networking requirements.

Please refer to the Design Considerations section of this guide for further details on determining bandwidth requirements.

Data Site to Witness Network Latency

This refers to the communication between non-witness sites and the witness site.

In most vSAN Stretched cluster configurations, latency or RTT (Round Trip Time) between sites hosting VM objects and the witness nodes should not exceed 200msec (100msec one-way).

The latency to the witness is dependent on the number of objects in the cluster. VMware recommends that on vSAN Stretched Cluster configurations up to 10+10+1, a latency of less than or equal to 200 milliseconds is acceptable. However, a latency of less

than or equal to 100 milliseconds is preferred if possible. For configurations greater than 10+10+1, VMware requires a latency of less than or equal to 100 milliseconds.

Data Site to Witness Network Bandwidth

Bandwidth between sites hosting VM objects and the witness nodes depends on the number of objects residing on vSAN. It is essential to size data site-to-witness bandwidth appropriately for availability and growth. A standard rule is 2Mbps for every 1000 components on vSAN.

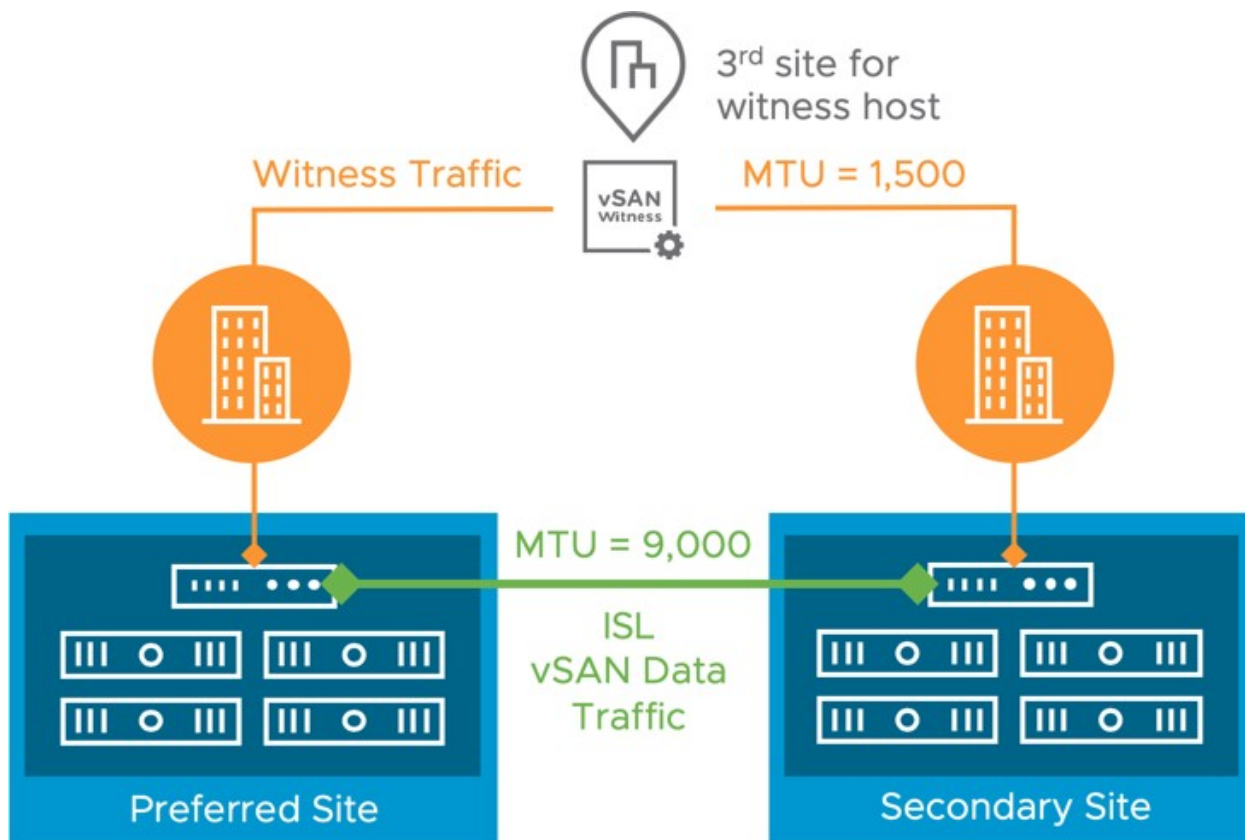
Please refer to the Design Considerations section of this guide for further details on determining bandwidth requirements.

Support for Mixed MTU for Witness Traffic separation

Traditional vSAN Clusters require the MTU size to be uniform across all vSAN VMkernel interfaces. vSAN Stretched Clusters had the exact requirement, due to the vSAN data network connecting to the vSAN Witness Host. When VMkernel ports communicate and have mixed MTU sizes, potential communication challenges include dropped packets, retransmits, etc.

That concept is changed when using Witness Traffic Separation. The “backend” vSAN data network doesn’t communicate with the vSAN Witness. Communication with the vSAN Witness is over a different VMkernel interface tagged for “witness” traffic.

Different MTU sizes on the vSAN data network and the network used to communicate with the vSAN Witness would work without issue.



VMware does not support mixed MTUs for interfaces that communicate directly with each other. It does support different MTUs for the different vSAN stretched cluster traffic types.

The vSAN Health Check recognizes Witness Traffic Separation deployments and allows for a different MTU for the vSAN data and vSAN witness networks.

This illustration shows the vSAN data network (VMkernel vmk2) with Jumbo Frames configured and vSAN Traffic enabled.

VMkernel adapters

Add Networking... Refresh Edit... Remove

| Device | Network Label | Switch | IP Address | TCP/IP Sta... | vMotion | P | F | Management | vS | vSAN | vSAN Witness |
|--------|--------------------|----------|----------------|---------------|----------|---|-----|------------|-------|----------|--------------|
| vmk0 | Management Network | vSwitch0 | 192.168.1.21 | Default | Disabled | D | Dis | Enabled | D ... | Disabled | Enabled |
| vmk1 | DSwitch-VMOTION | DSwitch | 192.168.151.21 | vMotion | Enabled | D | Dis | Disabled | D ... | Disabled | Disabled |
| vmk2 | DSwitch-VSAN | DSwitch | 192.168.152.21 | Default | Disabled | D | Dis | Disabled | D ... | Enabled | Disabled |

VMkernel network adapter: vmk2

All Properties IP Settings Policies

Port properties

Network label: DSwitch-VSAN
 TCP/IP stack: Default
 Enabled services: vSAN

NIC settings

MAC address: 00:50:56:67:9e:b7
 MTU: 9000

VMkernel adapters

Add Networking... Refresh Edit... Remove

| Device | Network Label | Switch | IP Address | TCP/IP Sta... | vMotion | P | F | Management | vS | vSAN | vSAN Witness |
|--------|--------------------|----------|----------------|---------------|----------|---|-----|------------|-------|----------|--------------|
| vmk0 | Management Network | vSwitch0 | 192.168.1.21 | Default | Disabled | D | Dis | Enabled | D ... | Disabled | Enabled |
| vmk1 | DSwitch-VMOTION | DSwitch | 192.168.151.21 | vMotion | Enabled | D | Dis | Disabled | D ... | Disabled | Disabled |
| vmk2 | DSwitch-VSAN | DSwitch | 192.168.152.21 | Default | Disabled | D | Dis | Disabled | D ... | Enabled | Disabled |

VMkernel network adapter: vmk0

All Properties IP Settings Policies

Port properties

Network label: Management Network
 VLAN ID: None (0)
 TCP/IP stack: Default
 Enabled services: Management, vSAN Witness

NIC settings

MAC address: 00:e0:81:c7:eb:86
 MTU: 1500

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

VMkernel adapters

Add Networking... Refresh Edit... Remove

| Device | Network Label | Switch | IP Address | TCP/I... | vM... | Prov... | FT L... | Man... | v... | v... | vSAN |
|--------|-----------------|---------------|-----------------|----------|---------|-----------|-----------|----------|--------|---------|----------|
| vmk0 | Management N... | vSwitch0 | 192.168.109.23 | Default | Disa... | Disabl... | Disabl... | Enabled | Dis... | Disa... | Enabled |
| vmk1 | witnessPg | witnessSwitch | 169.254.226.189 | Default | Disa... | Disabl... | Disabl... | Disabled | Dis... | Disa... | Disabled |

VMkernel network adapter: vmk0

All Properties IP Settings Policies

Port properties

Network label: Management Network
 VLAN ID: None (0)
 TCP/IP stack: Default
 Enabled services: Management, vSAN

NIC settings

MAC address: 00:50:56:b9:21:43
 MTU: 1500

The vSAN Witness has “vSAN Traffic” tagged, though it communicates with the “witness” tagged interfaces on the data nodes when using Witness Traffic Separation. vmk1 is not used in this installation, which is a valid configuration.

The vSAN Health Check shows that the hosts correctly communicate with the vSAN Witness, even though they have different MTU

Health (Last checked: Oct 8, 2018, 3:03:58 PM) RETEST

- > ✔ Cluster
- > ✔ Network
 - ✔ Hosts disconnected from VC
 - ✔ Hosts with connectivity issues
 - ✔ vSAN cluster partition
 - ✔ All hosts have a vSAN vmknic...
 - ✔ vSAN: Basic (unicast) connect...
 - ✔ vSAN: MTU check (ping with ...)
 - ✔ vMotion: Basic (unicast) conn...
 - ✔ vMotion: MTU check (ping wit...
 - > ✔ Stretched cluster
 - > ✔ Data
 - > ✔ Limits

vSAN: MTU check (ping with large packet size)

Only Failed Pings Ping Results Info ✕

[Silence Alert](#)

| From Host | To Host | To Device | Ping result |
|----------------------|----------------------|-----------|--------------------------------------|
| host1.demo.local | witness.demo.central | vmk0 | ✔ |
| host1.demo.local | host2.demo.local | vmk2 | ✔ |
| host2.demo.local | host1.demo.local | vmk2 | ✔ |
| host2.demo.local | witness.demo.central | vmk0 | ✔ |
| witness.demo.central | host1.demo.local | vmk0 | ✔ |
| witness.demo.central | host2.demo.local | vmk0 | ✔ |

6 items

settings.

Multiple vSAN Witness Hosts sharing the same VLAN

Multiple 2 Node vSAN deployments can send their witness traffic on the same shared VLAN. Each remote site does not require its own separate VLAN.

Configuration Minimums and Maximums

Virtual Machines Per Host

The maximum number of virtual machines per ESXi host is unaffected by the vSAN Stretched Cluster configuration. The maximum is the same as standard vSAN deployments.

VMware recommends that customers run their hosts at 50% of the maximum number of virtual machines supported in a standard vSAN cluster to accommodate a total site failure.

In the event of complete site failures, the virtual machines on the failed site can be restarted on the hosts in the surviving site.

Hosts Per Cluster

Minimum Host Count

The minimum number of hosts in a vSAN Stretched Cluster is two plus the vSAN Witness Host. Site 1 contains one physical ESXi host; Site 2 contains one physical ESXi host; Site 3 contains the vSAN Witness Host (virtual appliance or physical). The configuration is 1+1+1. This is commonly referred to as a 2 Node configuration. If both nodes in a 2 Node configuration are on the same site, the vSAN Witness Host can be on the same site or an alternate site. In configurations where the 2 Nodes are in alternate sites, the vSAN Witness Host must be in a third site.

Maximum Host Count

The maximum number of vSAN hosts is 40 plus the vSAN Witness host.

Symmetrical Configurations

vSAN supports symmetrical configurations where Site 1 contains up to 20 ESXi hosts, Site 2 contains the same number of ESXi hosts, and the vSAN Witness Host in a third site (20+20+1).

Asymmetrical Configurations

vSAN also supports asymmetrical configurations where some workloads may use the Site Affinity rule as part of a Storage Policy.

Example 1: Site 1 contains 20 ESXi hosts, Site 2 contains 10 ESXi hosts, and the Witness Host is in a third site.

In this use case, some workloads would only be available in Site 1 (Using a PFTT=0/SiteAffinity=Preferred) and others in Site 1 and Site 2.

Example 2: Site 1 contains 12 ESXi hosts, Site 2 contains 16 ESXi hosts, and the Witness Host is in a third site.

In this use case, some workloads would only be available in Site 2 (Using a PFTT=0/SiteAffinity=Non-Preferred) and others in Site 1 and Site 2.

Witness Host

There is a maximum of 1 Witness host per vSAN Stretched Cluster. The Witness host requirements are discussed in the Design Considerations section of this guide. VMware provides a fully supported vSAN Witness Appliance in Open Virtual Appliance (OVA) format. This is for customers who do not wish to dedicate a physical ESXi host as the witness. This OVA is essentially a pre-licensed ESXi host running in a virtual machine and can be deployed on a physical ESXi host at the third site.

vSAN Storage Policies

Primary Number of Failures To Tolerate (PFTT)

The FTT/PFTT policy setting has a maximum of 1 for objects. This is because Stretched Clusters are comprised of 3 Fault Domains.

Secondary Number of Failures To Tolerate (SFTT)

When used, the SFTT rule determines the Failure Tolerance Method for local protection in a Stretched Cluster.

Failure Tolerance Method (FTM)

Failure Tolerance Method rules provide object protection with RAID-1 (Mirroring) for Performance and RAID-5/6 (Erasure Coding) for Capacity.

In vSAN Stretched Clusters, Erasure Coding can be implemented using local protection, provided the host count and capacity are available. All-Flash vSAN is a requirement for supporting Erasure Coding.

Affinity

Affinity rules are used when the PFTT rule value is 0. This rule has two values, Preferred or Secondary. This determines which site an Affinity based vmdk would reside on.

Other Policy Rules

Other policy settings are not impacted by deploying vSAN in a Stretched Cluster configuration and can be used as per a non-stretched vSAN cluster.

Fault Domains

Fault domains play an essential role in vSAN Stretched Clusters. Similar to the Number Of Failures To Tolerate (FTT) policy setting discussed previously, the maximum number of fault domains in a vSAN Stretched Cluster is 3. The first fault domain is the “Preferred” data site, the second fault domain is the “Secondary” data site and the third fault domain is the Witness host site.

Design Considerations

The witness host must be capable of running the same version of ESXi as vSAN data nodes.

Witness Host Sizing

The vSAN Witness host can be a traditional physical ESXi host or the provided and packaged vSAN Witness Appliance (OVA). The purpose of the Witness host is to store witness components for virtual machine objects.

vSAN Witness Appliance (Virtual Machine)

Deploying the vSAN Witness Appliance provided by VMware is the recommended deployment choice for a vSAN Witness Host. When choosing this deployment option, there are some requirements to consider.

Licensing

A license is hardcoded in the vSAN Witness Appliance and is provided for free from VMware.

vSAN Witness Appliance Version

A vSAN Witness Appliance is provided with each release of vSAN. The underlying vSphere version is the same as the version running vSAN. Upon initial deployment of the vSAN Witness Appliance, it must be the same as the version of vSAN.

Example: A new vSAN 8.0 deployment requires an 8.0 version of the vSAN Witness Appliance.

When upgrading the vSAN Cluster, upgrade the vSAN Witness Appliance in the same fashion as upgrading vSphere.

vSAN Witness Appliance Size

When using a vSAN Witness Appliance, the size is dependent on the configurations and this is decided during the deployment process. vSAN Witness Appliance deployment options are hard coded upon deployment and there is typically no need to modify these.

Compute Requirements

The vSAN Witness Appliance, regardless of configuration, uses at least two vCPUs.

Memory Requirements

Memory requirements are dependent on the number of components.

Storage Requirements

Cache Device Size: Each vSAN Witness Appliance deployment option has a cache device size of 10GB. This is sufficient for each for a maximum of 64,000 components. In a typical vSAN deployment, the cache device must be a Flash/SSD device. Because the vSAN Witness Appliance has virtual disks, the 10GB cache device is configured as a virtual SSD. This device is not required to reside on a physical flash/SSD device. Traditional spinning drives are sufficient.

Capacity Device Sizing: First, consider that a capacity device can support up to 21,000 components. Also, a vSAN Stretched Cluster can support a maximum of 64,000 components. Each Witness Component is 16MB. As a result, the largest capacity device that can be used for storing Witness Components is approaching 350GB.

vSAN Witness Appliance Deployment Sizes & Requirements Summary

- Tiny - Supports up to 10 VMs/750 Witness Components
 - Compute - 2 vCPUs
 - Memory - 8GB vRAM
 - ESXi Boot Disk - 12GB Virtual HDD
 - Cache Device - 10GB Virtual SSD
 - Capacity Device - 15GB Virtual HDD
- Normal - Supports up to 500 VMs/21,000 Witness Components
 - Compute - 2 vCPUs
 - Memory - 16GB vRAM
 - ESXi Boot Disk - 12GB Virtual HDD

- Cache Device - 10GB Virtual SSD
- Capacity Device - 350GB Virtual HDD
- Large - Supports over 500 VMs/45,000 Witness Components
 - Compute: 2 vCPUs
 - Memory - 32 GB vRAM
 - ESXi Boot Disk - 12GB Virtual HDD
 - Cache Device - 10GB Virtual SSD
 - Capacity Devices - 3x350GB Virtual HDD
- Extra Large - Is for shared witness (2-node) only

Where can the vSAN Witness Appliance run?

It can be run in any of the following infrastructure configurations provided appropriate networking is in place:

- On a vSphere environment with any supported storage (vmfs datastore, NFS datastore, vSAN Cluster)
- Any vCloud Air Network partner-hosted solution
- On a vSphere Hypervisor (free) installation

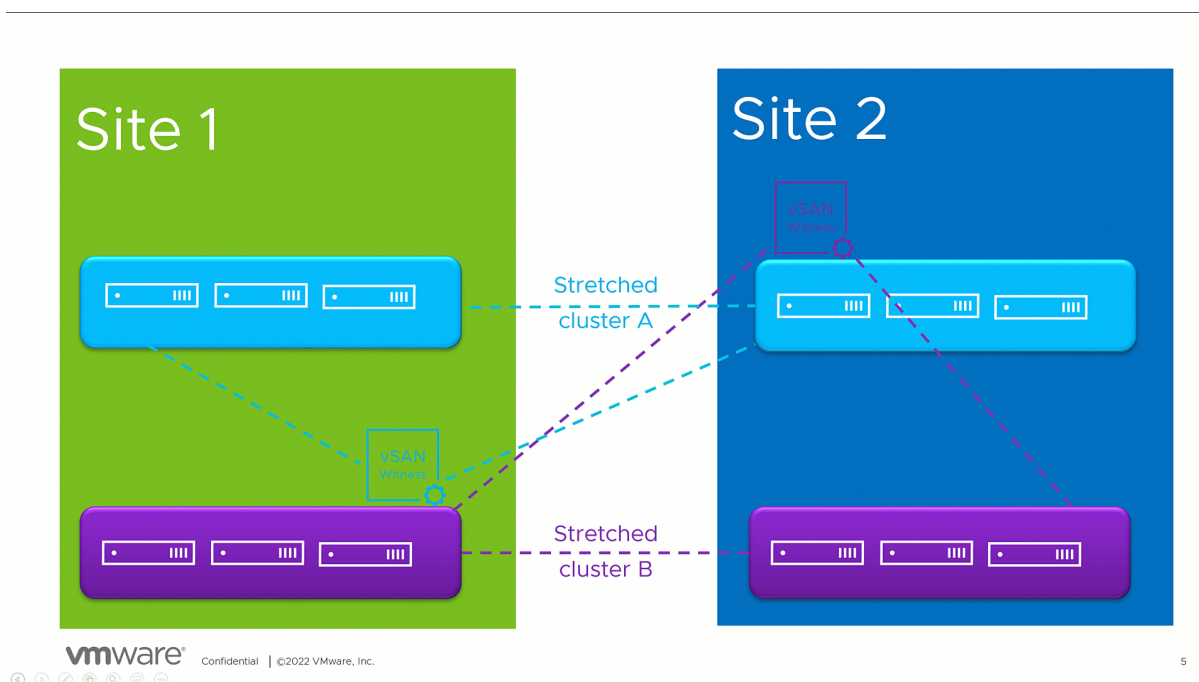
Support Statements specific to the placement of the vSAN Witness Appliance on a vSAN cluster:

- The vSAN Witness Appliance is supported running on a non-Stretched vSAN cluster.
- The vSAN Witness Appliance is supported on a Stretched Cluster vSAN for another vSAN Stretched Cluster only if four independent sites host two different stretched clusters (stretched cluster 'A' and stretched cluster 'B'). We can have the witness for stretched cluster 'A' deployed on stretched cluster 'B' and vice-versa. Below you can find more examples of what configurations are not supported and what is.

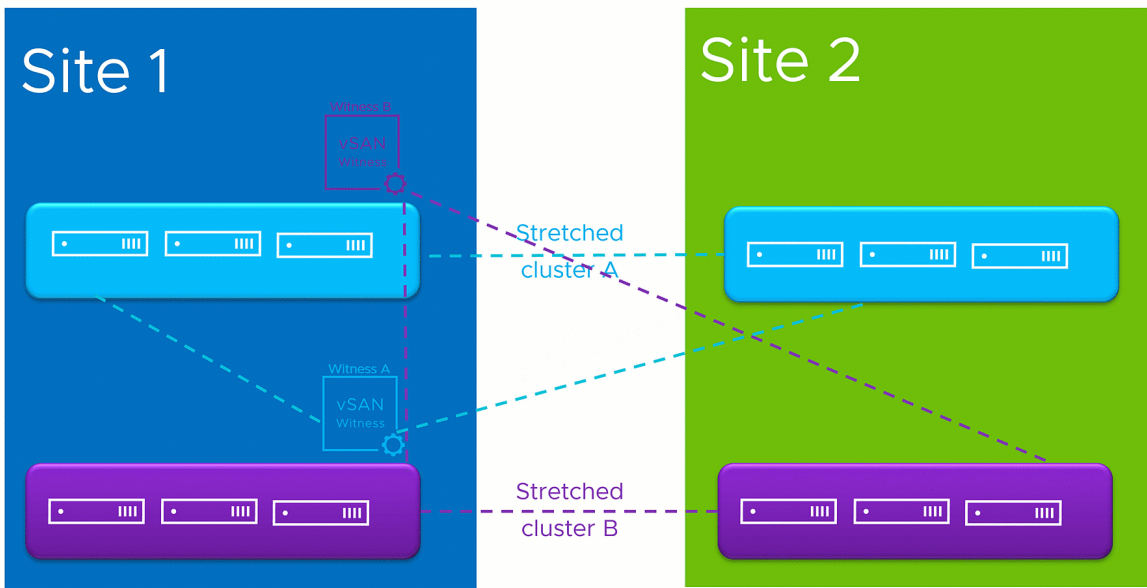
1. Cross-Site Witness Support is not supported on two sites hosting two stretched cluster implementations.

Setting up cross-witnesses for two stretched clusters on only two sites is not supported. The reason why we do not support this configuration is that each stretched cluster has a dependency on a site and might result in cascading failures. Let's examine the following configurations: two sites and two stretched clusters.

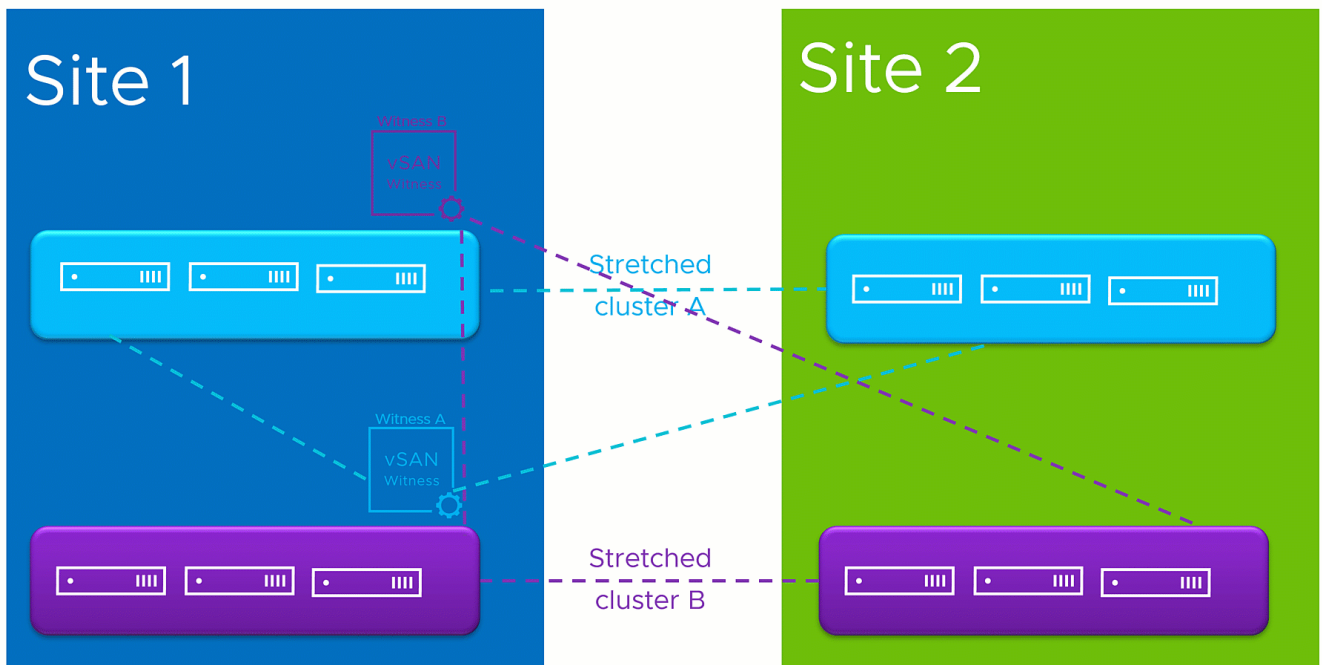
In this scenario, we have two sites. And two stretched clusters, A and B, across Site 1 and Site 2. Stretched cluster A (SCA) hosts Witness B (WB), and Stretched cluster B (SCB) hosts Witness A (WA). In this case, when one of the sites fails, this might result in cascading failures for both stretched cluster topologies. The VM residing on SCA will become unavailable due to losing one site and the witness site. While SCB will lose one site. This cross-witness support configuration will result in the same outcome if the alternate site fails.



Both witnesses can run on one site and lose only the site that doesn't host any witnesses. In this case, there will be a site failure for each stretched cluster, but no impact on the VM's running state.

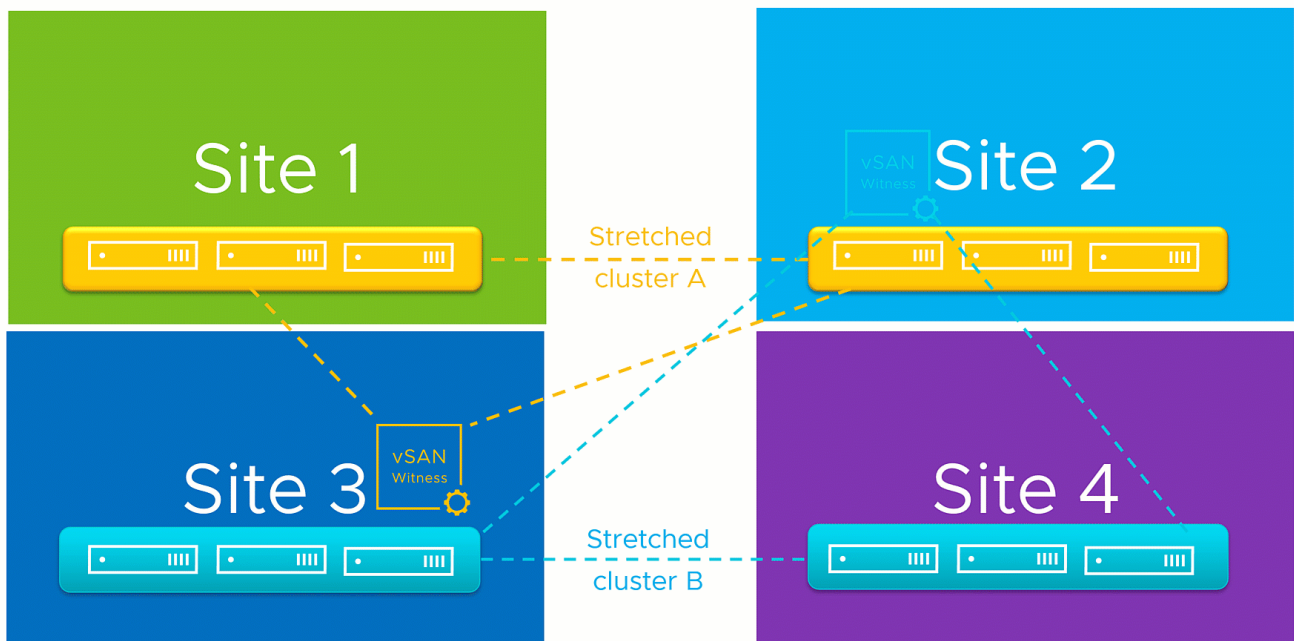


In this same configuration, losing Site 1 will impact the VMs' running state of both SCA and SCB since both of them will have more than one site failure.



2. Cross-Site Witness Support supported on four sites hosting two stretched cluster implementations.

Let's have a look at what the improvements are if we have four unique data sites. SCA is implemented across two sites, and stretched cluster B is implemented across two other sites. There are no dependencies between any of the sites other than that the witness appliance is hosted on the other stretched cluster, WA on SCB and WB on SCA. One site failure will not affect the object's quorum, and the VM will be up and running.



Physical Host as a vSAN Witness Host

If using a physical host as the vSAN Witness Host, there are some requirements to consider.

Licensing

When using a physical host as a vSAN Witness Host it must be licensed with a valid vSphere license. This does not require the same licensed edition as the vSAN Cluster it is supporting.

vSphere Build

If using a physical host as a vSAN Witness Host, it must be running the same build of vSphere as the Stretched Cluster that it is participating with.

Compute and Memory Requirements

The minimum specifications required for ESXi meet the minimum requirements for use as a vSAN Witness Host. Minimum requirements for vSphere are dependent on the build of vSphere, and can be found in the documentation section for each edition in VMware Documentation: <https://www.vmware.com/support/pubs/>

Storage Requirements

Storage requirements do not change for a physical host being used as a vSAN Witness Host compared to the vSAN Witness Appliance. An ESXi boot device, a cache device, and one or more capacity devices are still required.

Required

- 1st device - vSphere Boot Device - Normal vSphere Requirements
- 2nd device - vSAN Cache Device - No requirement for Flash/SSD, but it must be tagged as Flash/SSD in ESXi to be used as though it were a Flash/SSD device. This must be at least 10GB in size
- 3rd device - Can be up to 350GB and will support metadata for up to 21,000 components on a vSAN Cluster

Optional

- 4th device - Can be up to 350GB and will support metadata for up to 21,000 components on a vSAN Cluster
- 5th device - Can be up to 350GB and will support metadata for up to 21,000 components on a vSAN Cluster

Other workloads

If using a physical host as a vSAN Witness Host, it may run other workloads. Because the physical vSAN Witness Host is external to

the vSAN Cluster it contributes to; those workloads will not be part of the vSAN Cluster. The vSAN Disk Group, and the disks it includes, may not be used for those workloads.

*Important consideration: Multiple vSAN Witness Appliances can run on a single physical host. Using vSAN Witness Appliances is typically more cost-effective than dedicating physical hosts for to meet the vSAN Witness Host need.

Cluster Compute Resource Utilization

VMware recommends that customers run at 50% of resource consumption across the vSAN Stretched Cluster for full availability. In the event of a complete site failure, all virtual machines could be run on the surviving site.

VMware understands that some customers will want to run levels of resource utilization higher than 50%. While running at higher utilization in each site is possible, it is essential to understand that not all virtual machines will be restarted on the surviving site in the event of failure.

With the introduction of Per-Site Policies, capacity requirements depend on the policies used.

| Protection | FTT/PFTT | FTM | SFTT | Capacity Required in Preferred Site | Capacity Required in Secondary Site | Capacity Requirement |
|---|----------|----------------|------|-------------------------------------|-------------------------------------|----------------------|
| Across Sites Only | 1 | Mirroring | 0 | 100% | 100% | 200% |
| Across Sites with Local Mirroring (RAID1 Single Failure) | 1 | Mirroring | 1 | 200% | 200% | 400% |
| Across Sites with Local Mirroring (RAID1 Double Failure) | 1 | Mirroring | 2 | 300% | 300% | 600% |
| Across Sites with Local Mirroring (RAID1 Triple Failure) | 1 | Mirroring | 3 | 400% | 400% | 800% |
| Across Sites with Local Erasure Coding (RAID5/Single Failure) | 1 | Erasure Coding | 1 | 133% | 133% | 266% |
| Across Sites with Local Erasure Coding (RAID6/Double Failure) | 1 | Erasure Coding | 2 | 150% | 150% | 300% |
| Single Site with Mirroring (RAID1 Single Failure) | 0 | Mirroring | 1 | 200% | 0 | 200% |
| Single Site with Mirroring (RAID1 Double Failure) | 0 | Mirroring | 2 | 300% | 0 | 300% |
| Single Site with Mirroring (RAID1 Triple Failure) | 0 | Mirroring | 3 | 400% | 0 | 400% |
| Single Site with Erasure Coding (RAID5/Single Failure) | 0 | Erasure Coding | 1 | 133% | 0 | 133% |
| Single Site with Erasure Coding (RAID6/Single Failure) | 0 | Erasure Coding | 2 | 150% | 0 | 150% |

Network Design Considerations

Stretched Cluster Network Design Considerations

Sites

A vSAN Stretched Cluster requires three Fault Domains. Two fault domains are configured as Preferred and Non-Preferred in the vSphere Client, and the vSAN Witness Host resides in a third (implied) Fault Domain.

Configured Fault Domains - Contain vSAN Data nodes.

- Preferred Fault Domain - Specified to be the primary owner of vSAN objects. This is a crucial designation, specifically in cases of connectivity disruptions.
- Non-Preferred Fault Domain - The alternate Fault Domain
- These Fault Domains typically reside in geographically separated locations.

Witness Site - Contains vSAN Witness host.

- Maintains Witness Component data from Preferred/Non-Preferred Fault Domains when applicable*
- *When using "Site Affinity" Witness Components will not reside in the Witness site
- When using vSAN Stretched Clusters in a single datacenter, different rooms or different racks could be considered separate sites.

Connectivity and Network Types

| | Preferred Site | Secondary 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 |

Port Requirements

VMware vSAN requires these 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 |
| vSAN Unicast Agent (to Witness Host) | 12321 | UDP | vSAN Hosts and vSAN Witness Appliance |

TCPIP Stacks, Gateways, and Routing TCPIP Stacks

Currently, the vSAN traffic does not have a dedicated TCPIP stack. Custom TCPIP stacks are also not applicable for vSAN traffic.

Default Gateway on ESXi Hosts

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). It is a best practice to implement storage networking, in this case, vSAN networking, on an alternate VMkernel interface with alternate addressing.

vSAN networking uses the same TCPIP stack as the Management VMkernel interface, traffic defaults to using the same default gateway as the Management VMkernel interface. With the vSAN network isolated from the Management VMkernel interface, it is impossible to use the default gateway. Because of this, vSAN Data Nodes cannot communicate with the Witness Host by default.

One solution to this issue is to use static routes. This allows an administrator to define a new routing entry indicating which path should be followed to reach a particular network in the case of the vSAN network on a vSAN Stretched Cluster.

Static routes could be added as follows:

- Hosts on the Preferred Site have a static route added so that requests to reach the witness network on the Witness Site are routed out of the vSAN VMkernel interface.
- Hosts on the Secondary Site have a static route added so that requests to reach the witness network on the Witness Site are routed out of the vSAN VMkernel interface.
- The Witness Host on the Witness Site has a static route added so that requests to reach the Preferred Site and Secondary Site are routed out of the WitnessPg VMkernel interface.
- Using Layer 3 between the Preferred Site & Secondary Sites may require static routes to communicate across the inter-site link properly.

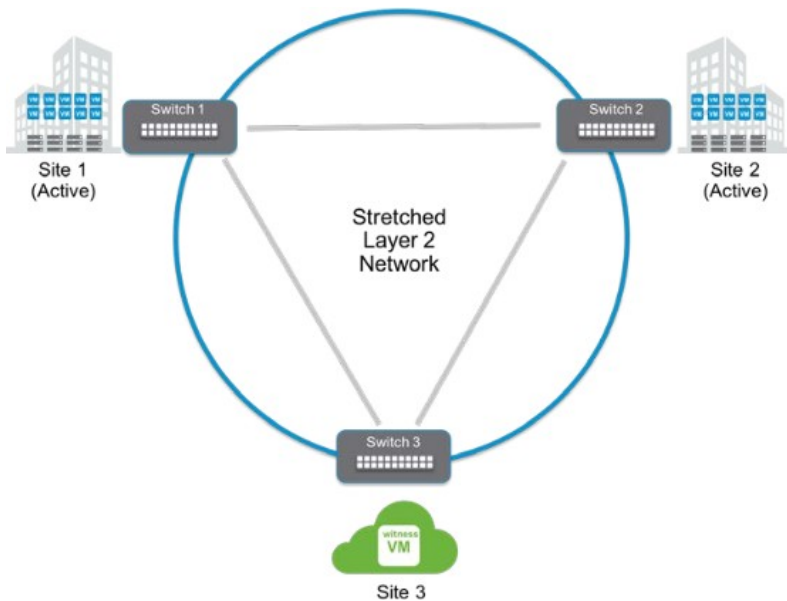
*Note, this may result in an alert (which may be disregarded provided connectivity is verified) that the vSAN network does have a matching subnet.

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.

Caution when implementing Static Routes: Using static routes requires administrator intervention. Any new ESXi hosts added to the cluster at either site one or site two need to manually add static routes before they can successfully communicate to the witness and the other data site. Any replacement of the witness host will also require the static routes to be updated to facilitate communication with the data sites.

Topology - L2 Design Versus L3 Design

Consider a design where the vSAN Stretched Cluster is configured in one large L2 design, where the Preferred Site (Site 1) and Secondary Site (Site 2) are where the virtual machines are deployed. The Witness site contains the Witness Host:

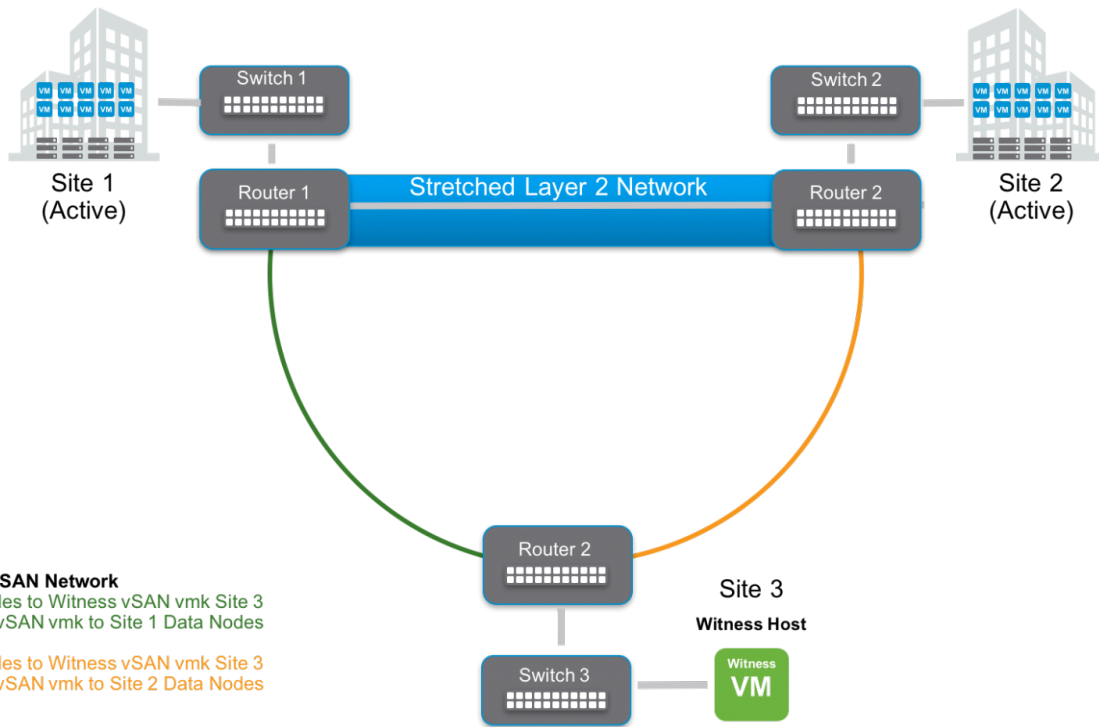


If the link between Switch 1 and Switch 2 is broken (the link between Site 1 and Site 2). Network traffic will now route from Site 1 to Site 2 via Site 3. Considering a much lower bandwidth requirement for connectivity to the Witness Host, customers would see a decrease in performance if network traffic is routed through a lower specification Site 3.

If there are situations where routing traffic between data sites through the witness site does not impact the latency of applications and bandwidth is acceptable, a stretched L2 configuration between sites is supported. However, in most cases, VMware feels that such a configuration is not feasible for most customers.

To avoid the situation outlined and to ensure that data traffic is not routed through the Witness Site, VMware recommends the following network topology:

- Between Site 1 and Site 2, implement either a stretched L2 (same subnet) or a L3 (routed) configuration.
- Implement an L3 (routed) configuration between data sites and the Witness site.
 - Ensure that Sites 1 and 2 can only connect to Site 3 directly, and not through the alternate site.
 - Static routing from data hosts (Site 1 & Site 2) to the Witness in Site 3 will be required.
 - Hosts in Site 1 should never traverse the inter-site link to reach Site 3.
 - Hosts in Site 2 should never traverse the inter-site link to reach Site 3.
 - Static routing will be required from the Witness host (Site 3) to the data hosts (Site 1 & Site 2)
 - The Witness should never route through Site 1, then across the inter-site link to reach Site 2.
 - The Witness should never route through Site 2, then across the inter-site link to reach Site 1.
- In the event of a failure on either of the data sites network, this configuration will also prevent any traffic from Site 1 being routed to Site 2 via Witness Site 3, and thus avoid any performance degradation.



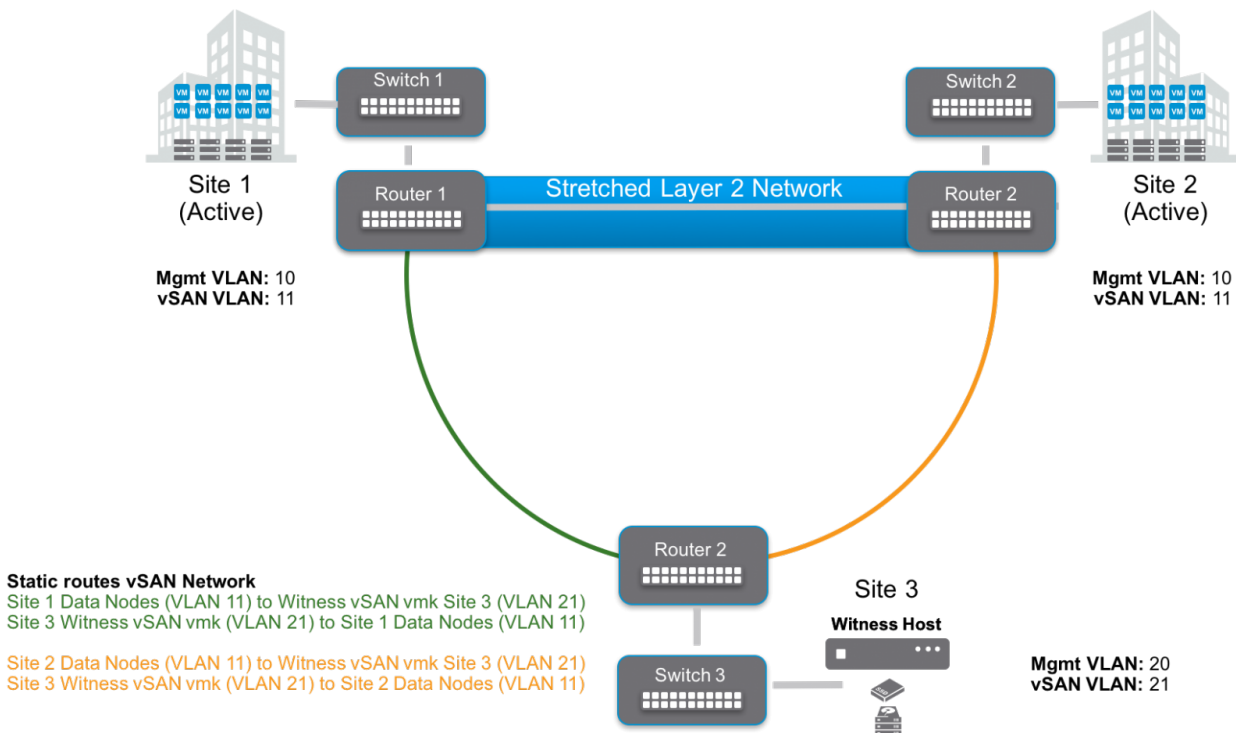
*If connectivity between the vSAN Network is configured to use L3:

- Each host in Site 1 will require a static route for the vSAN VMkernel interface to route across the inter-site link to each vSAN VMkernel interface for hosts in Site 2.
- Each host in Site 2 will require a static route for the vSAN VMkernel interface to route across the inter-site link to each vSAN VMkernel interface for hosts in Site 1.

Configuration of the Network from the Data Sites to the Witness

The next question is how to implement such a configuration, especially if the witness host is on a public cloud? How can the interfaces on the hosts in the data sites, which communicate to each other over the vSAN network, communicate to the witness host?

Option 1: Physical vSAN Witness Host connected over L3 & static routes



In this first configuration, the data sites are connected over a stretched L2 network. This is also true for the data sites' management network, vSAN network, vMotion network and virtual machine network. The physical network router in this network infrastructure does not automatically route traffic from the hosts in the data sites (Site 1 and Site 2) to the host in Site 3. For the vSAN Stretched Cluster to be successfully configured, all hosts in the cluster must communicate. How can a stretched cluster be deployed in this environment?

The solution is to use static routes configured on the ESXi hosts so that the vSAN traffic from Site 1 and Site 2 can reach the witness host in Site 3, and vice versa. While this is not a preferred configuration option, this setup can be very useful for proof-of-concept design where there may be issues with getting the required network changes implemented at a customer site.

In the case of the ESXi hosts on the data sites, a static route must be added to the vSAN VMkernel interface which will redirect traffic for the witness host on the witness site via a default gateway for that network. In the case of the witness host, the vSAN interface must have a static route added, redirecting vSAN traffic destined for the data sites' hosts. Adding static routes is achieved using the **esxcfg-route -a** command on the ESXi hosts. This must be repeated on all ESXi hosts in the stretched cluster.

For this to work, the network switches must be IP routing enabled between the vSAN network VLANs, in this example, VLANs 11 and 21. Once requests arrive for a remote host (either witness -> data or data -> witness), the switch will route the packet appropriately. This communication is essential for vSAN Stretched Cluster to work properly.

Note that we have not mentioned the ESXi management network here. The vCenter server will still be required to manage both the ESXi hosts at the data sites and the ESXi witness. In many cases, this is not an issue for customer. However, in the case of stretched clusters, it might be necessary to add a static route from the vCenter server to reach the management network of the witness ESXi host if it is not routable, and similarly a static route may need to be added to the ESXi witness management network to reach the vCenter server. This is because the vCenter server will route all traffic via the default gateway.

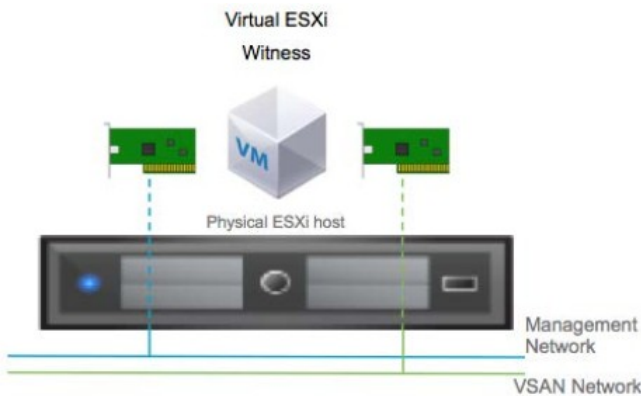
As long as there is direct connectivity from the witness host to vCenter (Not using NAT), there should be no additional concerns regarding the management network.

Also note that there is no need to configure a vMotion network or a VM network or add any static routes for these network in the context of a vSAN Stretched Cluster. This is because virtual machines will never be migrated or deployed to the vSAN Witness host. Its purpose is to maintain witness objects only, and does not require either of these networks for this task.

Option 2: Virtual vSAN Witness Host connected over L3 & static routes

Requirements: Since the virtual ESXi witness is a virtual machine that will be deployed on a physical ESXi host when deployed on-premises, the underlying physical ESXi host will need to have a minimum of one VM network preconfigured. This VM network will need to reach both the management network and the vSAN network shared by the ESXi hosts on the data sites. An alternative option that might be simpler to implement is to have two preconfigured VM networks on the underlying physical ESXi host, one for the management network and one for the vSAN network. When the virtual ESXi witness is deployed on this physical ESXi host, the

network will need to be attached/configured accordingly.

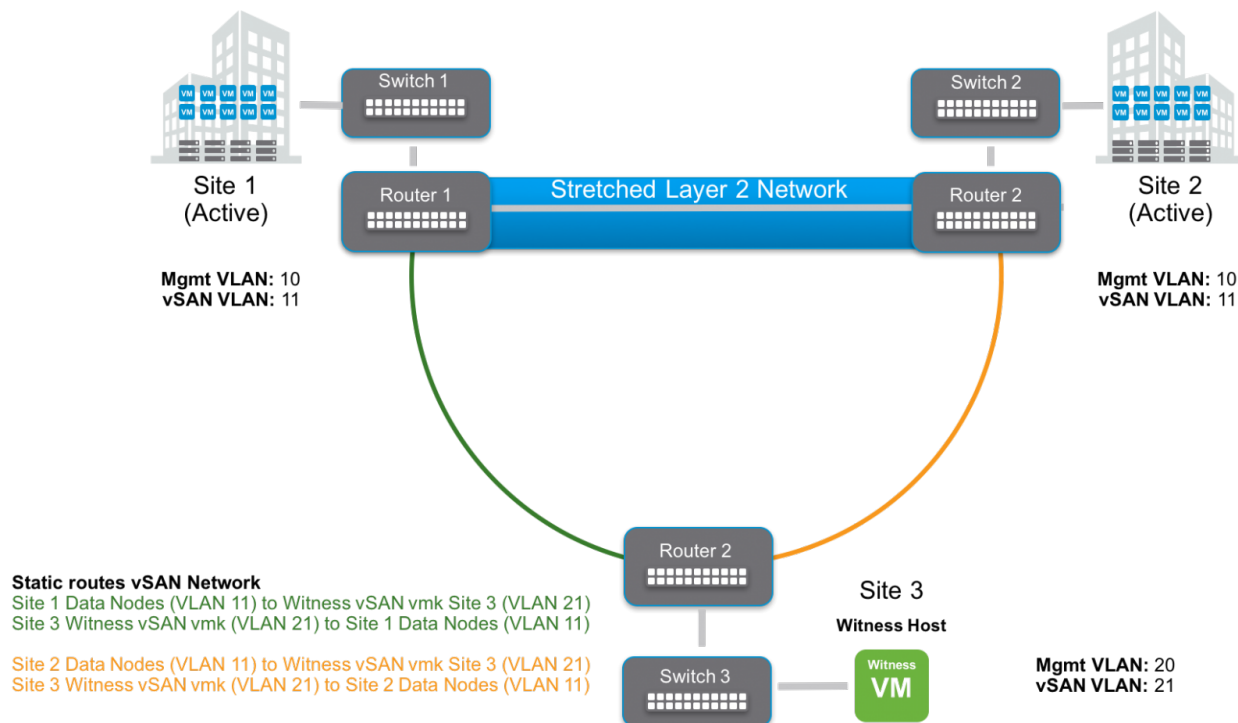


Once the vSAN Witness Appliance has been successfully deployed, the static route configuration must be configured.

As before, the data sites are connected over a stretched L2 network. This is also true for data sites' management network, vSAN network, vMotion network and virtual machine network. Once again, the physical network router in this environment does not automatically route traffic from the hosts in the Preferred and Secondary data sites to the host in the witness site. In order for the vSAN Stretched Cluster to be successfully configured, all hosts in the cluster require static routes added so that the vSAN traffic from the Preferred and Secondary sites is able to reach the Witness host in the witness site, and vice versa. As mentioned before, this is not a preferred configuration option, but this setup can be very useful for proof-of-concept design where there may be some issues with getting the required network changes implemented at a customer site.

Once again, the static routes are added using the `esxcfg-route-a` command on the ESXi hosts. This will have to be repeated on all ESXi hosts in the cluster, both on the data sites and on the witness host.

The switches should be configured to have IP routing enabled between the vSAN network VLANs on the data sites and the witness site, in this example VLANs 11 and 21. Once requests arrive for the remote host (either witness -> data or data -> witness), the switch will route the packet appropriately. With this setup, the vSAN Stretched Cluster will form.



Note that once again we have yet to mention the management network here. As mentioned before, vCenter must manage the remote ESXi witness and the hosts on the data sites. If necessary, a static route should be added to the vCenter server to reach the management network of the witness ESXi host. Similarly, a static route should be added to the ESXi witness to reach the vCenter server.

Also note that, as before, there is no need to configure a vMotion network or a VM network or add any static routes for these networks in the context of a vSAN Stretched Cluster. This is because virtual machines will never be migrated or deployed to the

vSAN witness. Its purpose is to maintain witness objects only and does not require either of these networks for this task.

Bandwidth Calculation

As stated in the requirements section, the bandwidth requirement between the two main sites depends on the workload, particularly the number of write operations per ESXi host. Other factors may also need to be factored in, such as read locality not in operation (where the virtual machine resides on one site but reads data from the other site) and rebuild traffic.

Requirements Between Data Sites

Reads are not included in the calculation as we assume read locality, which means there should be no inter-site read traffic. The required bandwidth between the two data sites (B) is equal to the Write bandwidth (Wb) * data multiplier (md) * resynchronization multiplier (mr):

$$B = Wb * md * mr$$

The data multiplier comprises overhead for vSAN metadata traffic and miscellaneous related operations. VMware recommends a data multiplier of 1.4. The resynchronization multiplier is included to account for resynchronizing events. It is recommended to allocate bandwidth capacity on top of the required bandwidth capacity for resynchronization events.

Making room for resynchronization traffic, an additional 25% is recommended.

- Data Site to Data Site Example 1

Take a hypothetical example of a 6-node vSAN Stretched Cluster (3+3+1) with the following:

A workload of 35,000 IOPS, 10,000 of those being write IOPS

A “typical” 4KB size write (This would require 40MB/s or 320Mbps bandwidth). Including the vSAN network requirements, the required bandwidth would be 560Mbps. $B = 320 \text{ Mbps} * 1.4 * 1.25 = 560 \text{ Mbps}$.

Data Site to Data Site Example 2

Take a 20-node vSAN Stretched Cluster (10+10+1) with a VDI (Virtual Desktop Infrastructure) with the following: A workload of 100,000 IOPS

With a typical 70%/30% distribution of writes to reads respectively, 70,000 of those are writes. A “typical” 4KB size write (This would require 280 MBps, or 2.24Gbps bandwidth)

Including the vSAN network requirements, the required bandwidth would be approximately 4Gbps.

$$B = 280 \text{ MBps} * 1.4 * 1.25 = 490 \text{ MBps or } 3.92\text{Gbps}$$

Using the above formula, a vSAN Stretched Cluster with a dedicated 10Gbps inter-site link, can accommodate approximately 170,000 4KB write IOPS

- Raw bandwidth (RB): $10 \text{ Gbps} = 10 * 1024^3 = 10,737,418,240 \text{ bps}$.
- Data Multiplier (DM): 1.4
- Resynch Multiplier (RM): 1.25
- Protocol Overhead Multiplier (POM): 1.1 (10% ethernet protocol overhead)
- I/O size (IOSz): 32,768 bits (4096 bytes)

- $\text{Max IOPS} = RB / DM / RM / POM / IOSz$
- $\text{Max IOPS} = 10,737,418,240 / 1.4 / 1.25 / 1.1 / 32,768 = 170,223 \text{ IOPS}$

Customers must evaluate their I/O requirements, but VMware feels that 10Gbps will meet most design requirements.

Above this configuration, customers would need to consider multiple 10Gb NICs teamed, or a 40Gb network.

VMware recommends a minimum of 10Gbps network connectivity between sites for optimal performance and possible future cluster expansion. For more specific requirements for vSAN 8 OSA and ESA, see the vSAN Networking Design Guide.

Requirements when Read Locality is not Available

The previous calculations are only for regular Stretched Cluster traffic with read locality. If a device fails, read operations must also traverse the inter-site network. This is because the mirrored copy of data is on the alternate site when using

NumberOfFailurestoTolerate=1.

The same equation for every 4K read IO of the objects in a degraded state would be added to the above calculations. The expected read IO would be used to calculate the additional bandwidth requirement.

In an example of a single failed disk, with objects from 5 VMs residing on the failed disk, with 10,000 (4KB) read IOPS, an additional 40 Mbps, would be required, in addition to the above Stretched Cluster requirements, to provide sufficient read IO bandwidth, during peak write IO, and resync operations.

Requirements Between Data Sites and the Witness Site

Witness bandwidth isn't calculated in the same way as the bandwidth between data sites because hosts designated as a witness do not maintain any VM data but only component metadata, so the requirements are much smaller.

Virtual Machines on vSAN are comprised of multiple objects, which can be split into multiple components depending on factors like policy and size. The number of components on vSAN directly impacts the bandwidth requirement between the data sites and the witness.

The required bandwidth between the Witness and each site is equal to $\sim 1138 \text{ B} \times \text{Number of Components} / 5\text{s}$

$1138 \text{ B} \times \text{NumComp} / 5 \text{ seconds}$

The 1138 B value comes from operations when the Preferred Site goes offline, and the Secondary Site takes ownership of all components.

When the primary site goes offline, the secondary site becomes the primary. The Witness sends updates to the new primary, followed by the new primary replying to the Witness as ownership is updated.

The 1138 B requirement for each component comes from a combination of a payload from the Witness to the backup agent, followed by metadata indicating that the Preferred Site has failed.

In the event of a Preferred Site failure, the link must be large enough to allow for the cluster ownership to change and ownership of all of the components within 5 seconds.

Witness to Site Examples

Workload 1

With a VM being comprised of:

- Three objects {VM namespace, vmdk (under 255GB), and vmSwap) Failure to Tolerate of 1 (FTT=1)
- Stripe Width of 1
- Approximately 166 VMs with the above configuration would require the Witness to contain 996 components.

To successfully satisfy the Witness bandwidth requirements for a total of 1,000 components on vSAN, the following calculation can be used:

Converting Bytes (B) to Bits (b), multiply by 8
 $B = 1138 \text{ B} * 8 * 1,000 / 5\text{s} = 1,820,800 \text{ Bits per second} = 1.82 \text{ Mbps}$

VMware recommends adding a 10% safety margin and round up.

$B + 10\% = 1.82 \text{ Mbps} + 182 \text{ Kbps} = 2.00 \text{ Mbps}$

With the 10% buffer included, a guideline can be stated that for every 1,000 components, 2 Mbps is appropriate.

Workload 2

With a VM being comprised of

- Three objects {VM namespace, vmdk (under 255GB), and vmSwap)
- Failure to Tolerate of 1 (FTT=1)
- Stripe Width of 2

Approximately 1,500 VMs with the above configuration would require 18,000 components to be stored on the Witness. To successfully satisfy the Witness bandwidth requirements for 18,000 components on vSAN, the resulting calculation is:

$B = 1138 \text{ B} * 8 * 18,000 / 5\text{s} = 32,774,400 \text{ Bits per second} = 32.78 \text{ Mbps}$
 $B + 10\% = 32.78 \text{ Mbps} + 3.28 \text{ Mbps} = 36.05 \text{ Mbps}$

Using the general equation of 2Mbps for every 1,000 components, $(\text{NumComp}/1000) \times 2\text{Mbps}$, it can be seen that 18,000

components does in fact require 36 Mbps.

The Role of vSAN Heartbeats

As mentioned previously, when vSAN is deployed in a Stretched Cluster configuration, the vSAN Primary Node is placed on the Preferred Site and the vSAN Backup Node is placed on the Non-Preferred Site. So long as there are nodes (ESXi hosts) available in the Preferred Site, then a primary node is always selected from one of the nodes on this site. Similarly, for the Non-Preferred Site, so long as there are nodes available on the Non-Preferred Site.

The vSAN Primary Node and the vSAN Backup Node send heartbeats every second. If communication is lost for 5 consecutive heartbeats (5 seconds) between the primary node and the Backup due to an issue with the Backup node, the primary node chooses a different ESXi host as a Backup on the remote site. This is repeated until all hosts on the remote site are checked. If a complete site fails, the primary node selects a Backup node from the Preferred Site.

A similar scenario arises when the primary node has a failure.

When a node rejoins an empty site after a complete site failure, either the primary node (in the case of the node joining the Preferred Site) or the Backup (where the node is joining the Non-Preferred Site) will migrate to that site.

If communication is lost for five consecutive heartbeats (5 seconds) between the primary node and the vSAN Witness Host, the vSAN Witness Host is deemed to have failed. If the vSAN Witness Host has permanently failed, a new vSAN Witness Host can be configured and added to the cluster.

Host number calculation

The storage policies applied to a stretched cluster would define the minimum number of hosts required per site. For example, a Site disaster tolerance set to “Dual site mirroring (stretched cluster)” and a Failures-to-tolerate set to “1 failure - RAID - 5 (Erasure coding)”, would result in 3 data blocks and one parity component. Thus we would need at least four hosts per site. The final configuration will be 4+4+ 1, 4 hosts per site, and one witness host.

From a capacity standpoint, if you have a 100 GB VM and set the Site disaster tolerance to “Dual site mirroring (stretched cluster)” and a Failures-to-tolerate equal to 1 failure - RAID - 1 (Mirroring), it means a RAID 1 is set in each site. In this case, a 100 GB VM would require 200 GB in each location. So, 200% required local capacity, and 400% for the total cluster. Using the below table, you can easily see the overhead. Note that RAID-5 and RAID-6 are only available when using all-flash.

| Description | Site disaster tolerance | Failures to tolerate | RAID | Hosts per site | Stretched Config | Single site capacity | Total cluster capacity |
|---|-------------------------|----------------------|--------|----------------|------------------|----------------------|------------------------|
| Standard Stretched across locations with local protection | 1 | 1 | RAID-1 | 3 | 3+3+1 | 200% of VM | 400% of VM |
| Standard Stretched across locations with local RAID-5 | 1 | 1 | RAID-5 | 4 | 4+4+1 | 133% of VM | 266% of VM |
| Standard Stretched across locations with local RAID-6 | 1 | 2 | RAID-6 | 6 | 6+6+1 | 150% of VM | 300% of VM |
| Standard Stretched across locations no local protection | 1 | 0 | RAID-1 | 1 | 1+1+1 | 100% of VM | 200% of VM |
| Not stretched, only local RAID-1 | 0 | 1 | RAID-1 | 3 | n/a | 200% of VM | n/a |
| Not stretched, only local RAID-5 | 0 | 1 | RAID-5 | 4 | n/a | 133% of VM | n/a |
| Not stretched, only local RAID-6 | 0 | 2 | RAID-6 | 6 | n/a | 150% of VM | n/a. |

Cluster Settings - vSphere HA

Certain vSphere HA behaviors have been modified, especially for vSAN. It checks the state of the virtual machines on a per-virtual-machine basis. vSphere HA can decide on whether a virtual machine should be failed over based on the number of components belonging to a virtual machine that can be accessed from a particular partition.

When vSphere HA is configured on a vSAN Stretched Cluster, VMware recommends the following:

| vSphere HA | Turn on |
|---|---|
| Host Monitoring | Enabled |
| Host Hardware Monitoring – VM Component Protection: “Protect against Storage Connectivity Loss” | Disabled (default) |
| Virtual Machine Monitoring | Customer Preference – Disabled by default |
| Admission Control | Set to 50% |
| Host Isolation Response | Power off and restart VMs |
| Datastore Heartbeats | “Use datastores only from the specified list”, but do not select any datastores from the list. This disables Datastore Heartbeats |
| Advanced Settings: | |
| das.usedefaultisolationaddress | False |
| das.isolationaddress0 | IP address on vSAN network on site 1 |
| das.isolationaddress1 | IP address on vSAN network on site 2 |
| das.ignoreInsufficientHbDatastore | True |

Always use an isolation address in the same network as vSAN. This ensures that the isolation is validated using the vSAN VMkernel interface. In a non-routable vSAN network, a switch virtual interface could be created on a physical switch in each site. This will give an isolation address IP on the vSAN segment that can be used for the das.isolationaddress entries.

Turn on vSphere HA

To turn on vSphere HA, select the cluster object in the vCenter inventory, Manage, then vSphere HA. From here, vSphere HA can be turned on and off via a checkbox.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses | Admission Control | Heartbeat Datastores | Advanced Options

You can configure how vSphere HA responds to the failure conditions on this cluster. The following failure conditions are supported: host, host isolation, VM component protection (datastore with PDL and APD), VM and application.

Enable Host Monitoring *i*

| | |
|-------------------------------|-----------------------------|
| > Host Failure Response | Restart VMs ▾ |
| > Response for Host Isolation | Power off and restart VMs ▾ |
| > Datastore with PDL | Disabled ▾ |
| > Datastore with APD | Disabled ▾ |
| > VM Monitoring | Disabled ▾ |

CANCEL OK

Host Monitoring

Host monitoring should be enabled on vSAN stretch cluster configurations. This feature uses network heartbeat 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.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses | Admission Control | Heartbeat Datastores | Advanced Options

You can configure how vSphere HA responds to the failure conditions on this cluster. The following failure conditions are supported: host, host isolation, VM component protection (datastore with PDL and APD), VM and application.

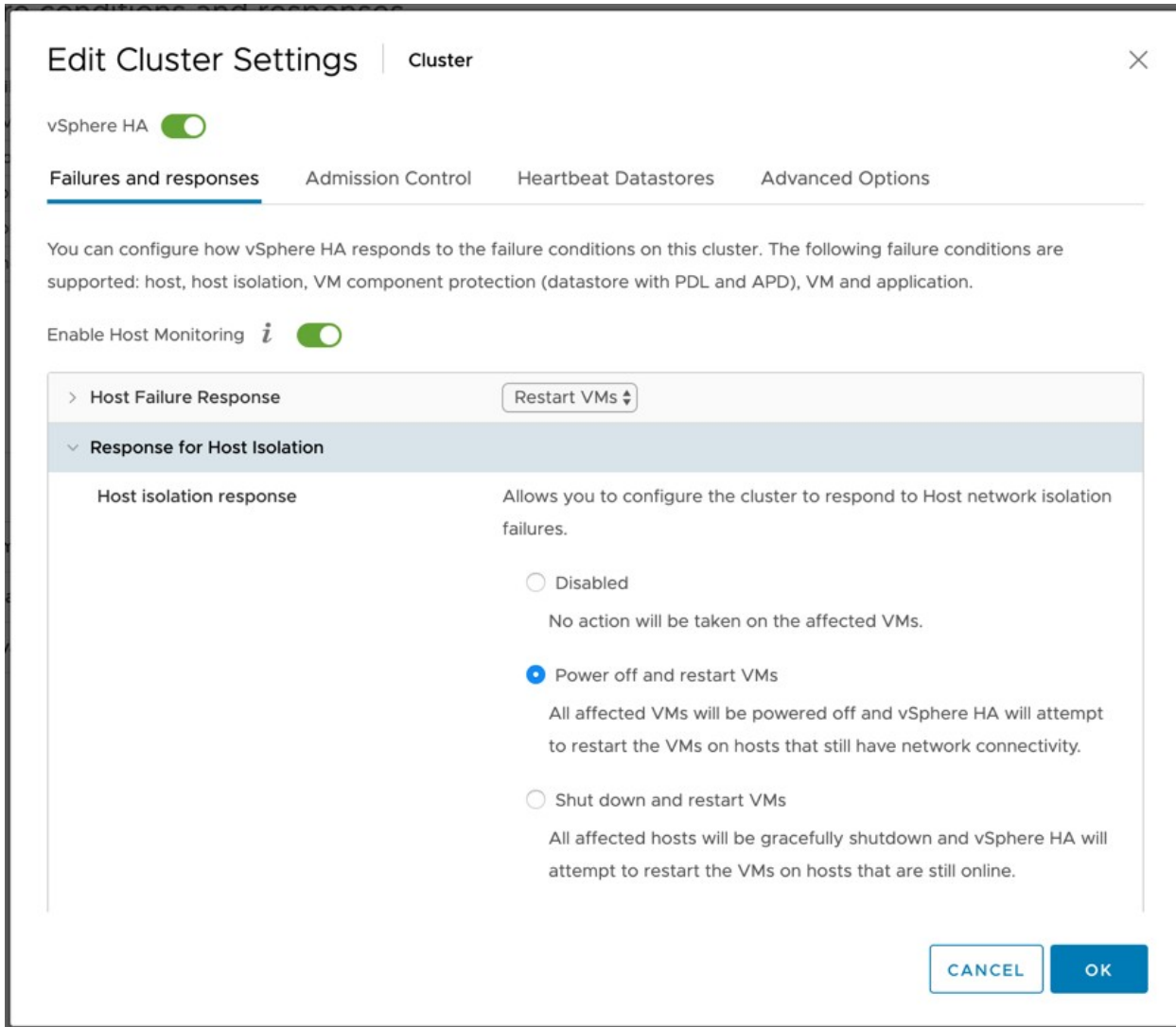
Enable Host Monitoring *i*

| | |
|-------------------------------|-----------------------------|
| > Host Failure Response | Restart VMs ▾ |
| > Response for Host Isolation | Power off and restart VMs ▾ |
| > Datastore with PDL | Disabled ▾ |
| > Datastore with APD | Disabled ▾ |
| > VM Monitoring | Disabled ▾ |

CANCEL OK

Virtual Machine Response for Host Isolation

This setting determines what happens to the virtual machines on an isolated host, i.e. a host that can no longer communicate to other nodes in the cluster nor reach the isolation response IP address. VMware recommends that the Response for Host Isolation is to Power off and restart VMs. This is because a clean shutdown will not be possible as on an isolated host, the access to the vSAN Datastore and the ability to write to disk is lost.



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 vSAN Stretched Cluster, and to ensure that vSphere HA can restart all workloads, VMware recommends configuring the admission control policy to 50 percent for both memory and CPU.

VMware recommends using the percentage-based policy as it offers the most flexibility and reduces operational overhead. For more details about admission control policies and the associated algorithms, we recommend the vSphere Availability Guide.

The following screenshot shows a vSphere HA cluster configured with admission control enabled using the percentage-based admission control policy set to 50%.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses | **Admission Control** | Heartbeat Datastores | Advanced Options

Admission control is a policy used by vSphere HA to ensure failover capacity within a cluster. Raising the number of potential host failures will increase the availability constraints and capacity reserved.

Host failures cluster tolerates _____
Maximum is one less than number of hosts in cluster.

Define host failover capacity by

Override calculated failover capacity.

Reserved failover CPU capacity: % CPU

Reserved failover Memory capacity: % Memory

Performance degradation VMs tolerate %

Percentage of performance degradation the VMs in the cluster are allowed to tolerate during a failure. 0% - Raises a warning if there is insufficient failover capacity to guarantee the same performance after VMs restart. 100% - Warning is disabled.

vSAN is not admission-control aware. There is no way to inform vSAN to set aside additional storage resources to accommodate fully compliant virtual machines running on a single site. This is an additional operational step for administrators to achieve such a configuration in the event of a failure.

Host Hardware Monitoring - VM Component Protection

vSphere 6.0 introduced a new enhancement to vSphere HA called VM Component Protection (VMCP) to allow for an automated fail-over of virtual machines residing on a datastore that has either an “All Paths Down” (APD) or a “Permanent Device Loss” (PDL) condition.

A permanent device loss condition (PDL) is a condition that is communicated by the storage controller to the ESXi host via a SCSI sense code. This condition indicates that a disk device has become unavailable and is likely permanently unavailable. When the storage controller can't communicate the status to the ESXi host, the condition is treated as an “All Paths Down” (APD) condition.

In traditional datastores, APD/PDL on a datastore affects all the virtual machines using that datastore. However, for vSAN, this may not be the case. An APD/PDL may only affect one or few VMs, but not all VMs on the vSAN datastore. Also, if an APD/PDL occurs on a subset of hosts, there is no guarantee that the remaining hosts will have access to all the virtual machine objects and be able to restart the virtual machine. Therefore, a partition may result in such a way that the virtual machine is not accessible on any partition.

The VM Component Protection (VMCP) way of handling a failover is to terminate the running virtual machine and restart it elsewhere in the cluster. VMCP/HA cannot determine the cluster-wide accessibility of a virtual machine on vSAN and thus cannot guarantee that the virtual machine can restart elsewhere after termination. For example, resources may be available to restart the virtual machine, but accessibility to the virtual machine by the remaining hosts in the cluster is not known to HA. This is not a problem for traditional datastores since we know host-datastore accessibility for the entire cluster, and by using that, we can determine if a virtual machine can be restarted on a host.

At the moment, vSphere HA can't understand the complete inaccessibility vs. partial inaccessibility on a per virtual machine basis on vSAN; hence the lack of VMCP support by HA for vSAN.

VMware recommends leaving VM Component Protection (VMCP) disabled.

Datastore for Heartbeating

vSphere HA provides an additional heartbeating mechanism for determining the state of hosts in the cluster. This is in addition to network heartbeating and is called "datastore heartbeating." In many vSAN environments, no additional datastores outside of vSAN are available. As such, VMware recommends disabling Heartbeat Datastores as the vSAN Datastore cannot be used for heartbeating. However, if additional datastores are available, then using heartbeat datastores is fully supported.

What do Heartbeat Datastores do, and when does it come into play? The heartbeat datastore is used by an isolated host to inform the rest of the cluster what its and the VMs' state is. When a host is isolated, and the isolation response is configured to "power off" or "shutdown", then the heartbeat datastore will be used to inform the rest of the cluster when VMs are powered off (or shutdown) as a result of the isolation. This allows the vSphere HA primary node to restart the impacted VMs immediately.

To disable datastore heartbeating, under vSphere HA settings, open the Datastore for Heartbeating section. Select "Use datastore from only the specified list" and ensure no datastore is selected if any exists. Datastore heartbeats are now disabled on the cluster. Note that this may give rise to a notification in the summary tab of the host, stating that the number of vSphere HA heartbeat datastore for this host is 0, which is less than required:2. This message may be removed by following KB Article 2004739, which details how to add the advanced setting `das.ignoreInsufficientHbDatastore = true`.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses | Admission Control | **Heartbeat Datastores** | Advanced Options

vSphere HA uses datastores to monitor hosts and virtual machines when the HA network has failed. vCenter Server selects 2 datastores for each host using the policy and datastore preferences specified below.

Heartbeat datastore selection policy:

- Automatically select datastores accessible from the hosts
- Use datastores only from the specified list
- Use datastores from the specified list and complement automatically if needed

Available heartbeat datastores

| Name | Datastore Cluster | Hosts Mounting Datastore ↓ |
|------|-------------------|----------------------------|
| | | |

CANCEL OK

Advanced Options

When vSphere HA is enabled on a vSAN Cluster, a heartbeat mechanism is used to validate the state of an ESXi host. Network heart beating is the primary mechanism for HA to validate host availability.

Suppose a host is not receiving any heartbeats. In that case, it uses a fail-safe mechanism to detect whether it is isolated from its HA primary node or completely isolated from the network. It does this by pinging the default gateway.

In vSAN environments, vSphere HA uses the vSAN traffic network for communication. This is different from traditional vSphere environments where the management network is used for vSphere HA communication. However, even in vSAN environments, vSphere HA continues using the management network's default gateway for isolation detection responses. This should be changed so that the isolation response IP address is on the vSAN network, as this allows HA to react to a vSAN network failure.

In addition to selecting an isolation response address on the vSAN network, additional isolation addresses can be specified manually to enhance the reliability of isolation validation.

Network Isolation Response and Multiple Isolation Response Addresses

In a vSAN Stretched Cluster, one of the isolation addresses should reside in the site 1 data center and the other should reside in the site 2 data center. This would enable vSphere HA to validate host isolation even in the case of a partitioned scenario (network failure between sites).

VMware recommends enabling host isolation response and specifying isolation response addresses that are on the vSAN network rather than the management network.

The vSphere HA advanced setting `das.usedefaultisolationaddress` should be set to `false`.

VMware recommends specifying two additional isolation response addresses, and each of these addresses should be site-specific. In other words, select an isolation response IP address from the Preferred Site and another isolation response IP address from the Non-Preferred Site.

The vSphere HA advanced setting used for setting the first isolation response IP address is `das.isolationaddress0` and it should be set to an IP address on the vSAN network which resides on the one site.

The vSphere HA advanced setting used for adding a second isolation response IP address is `das.isolationaddress1` and this should be an IP address on the vSAN network that resides on the alternate site.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses | Admission Control | Heartbeat Datastores | **Advanced Options**

You can set advanced options that affect the behavior of your vSphere HA cluster.

+ Add X Delete

| Option | Value |
|--------------------------------|----------------|
| das.isolationaddress0 | 192.168.152.30 |
| das.isolationaddress1 | 192.168.153.30 |
| das.usedefaultisolationaddress | false |

3 items

CANCEL OK

Cluster Settings - DRS

vSphere DRS is used in many environments to distribute load within a cluster. vSphere DRS offers many other features which can be very helpful in vSAN Stretched Cluster environments.

If administrators wish to enable DRS on vSAN Stretched Cluster, there is a requirement to have a vSphere Enterprise Plus license edition or higher. The vSphere Enterprise for ROBO license provides DRS features when putting hosts into or taking them out of maintenance mode.

VM/Host Groups & Rules and vSphere DRS

Creating VM to Host affinity rules mapping VM to Host groups is recommended. These specify which virtual machines and hosts reside in the Preferred Site and which reside in the Non-Preferred Site. Using Host/VM groups and rules, it becomes easy for administrators to manage which virtual machines should run on which site, and balance workloads between sites. In the next section, Host/VM groups and rules are discussed.

When used with VM to Host Affinity groups/rules, vSphere DRS addresses: Administrators can easily balance workloads between sites.

When virtual machines are powered on, they will only be powered-on on hosts that conform to the VM/Host groups and rules settings.

VM/Host groups are discussed more thoroughly in the next section.

Complete Site Failure/Restoration & vSphere DRS

When a complete site failure occurs, vSphere HA will restart all virtual machines on the remaining site.

VM/Host Rules: Should be set to "Should run on hosts in group" for workloads that can be run on the alternate site in the event of a failure. For workloads that should only ever run on a single site, such as in cases of asymmetrical Stretched Clusters or when using Site Affinity Storage Policies, a "Must run on hosts in group" rule should be used. More information can be found in the Per-Site Policy Considerations section.

Partially Automated or Fully Automated DRS

Customers can decide whether to place DRS in partially or fully automated mode. 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 on 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.

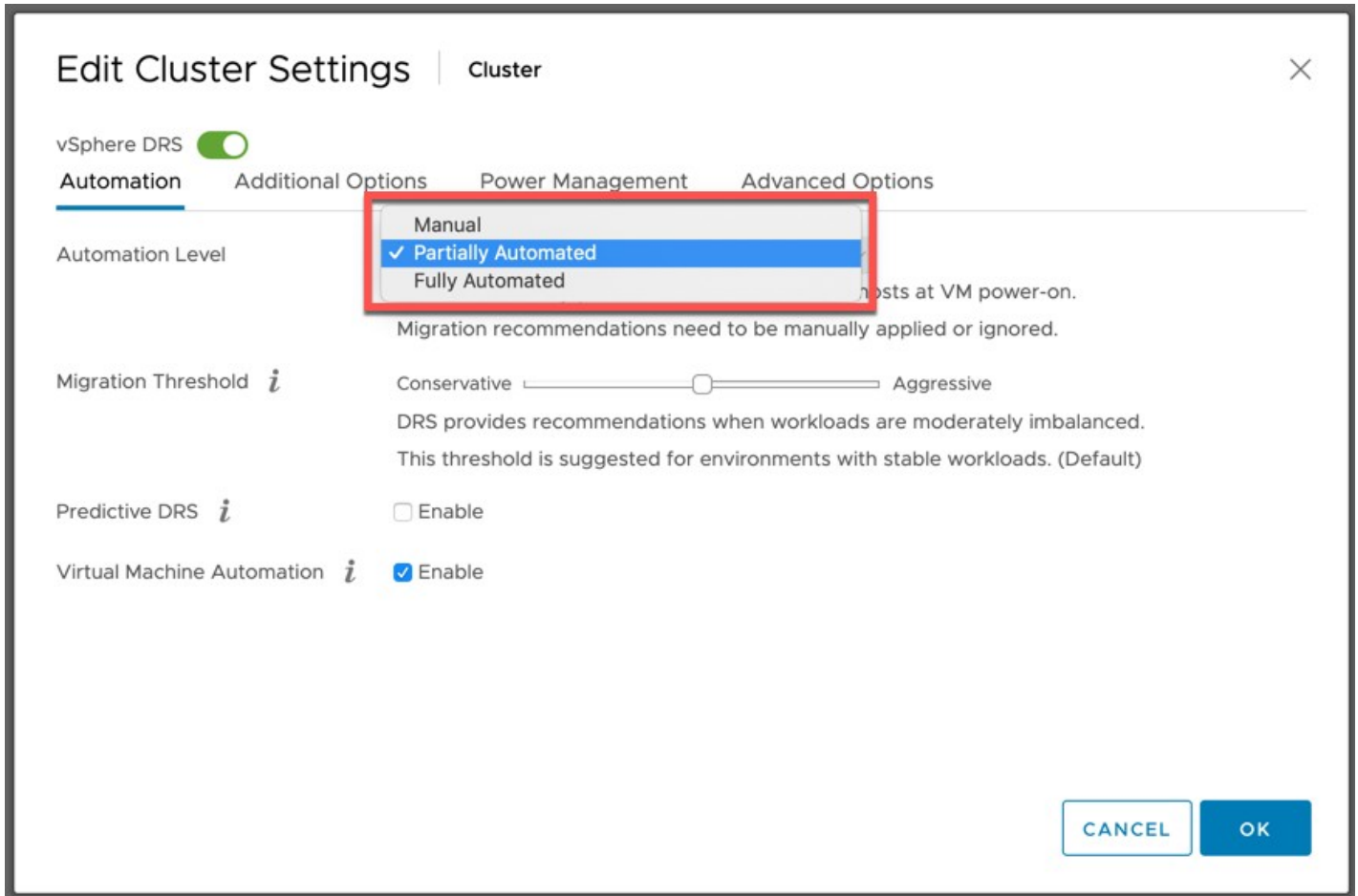
One significant consideration with the fully automated mode is a site failure. Consider a situation where a site has failed, and all virtual machines are now running on a single site. All virtual machines on the running site have read locality with the running site and are caching their data on the running site. Perhaps the outage has been a couple of hours or even a day. The issue at the failed site has been addressed (e.g. power, network, etc.). When the hosts on the recovered rejoin the vSAN cluster, there has to be a resync of all components from the running site to the recovered site. This may take some time. However, at the same time, DRS is informed that the hosts are back in the cluster. If in fully automated mode, the affinity rules are checked, and obviously, many are not compliant. Therefore DRS begins to move virtual machines back to the recovered site, but the components may not yet be active (i.e. still synchronizing). Therefore virtual machines could end up on the recovered site, but since there is no local copy of the data, I/O from these virtual machines will have to traverse the link between sites to the active data copy. This is undesirable due to latency/performance issues. Therefore, for this reason, VMware recommends that DRS is placed in partially automated mode if there is an outage. Customers will continue to be informed about DRS recommendations when the hosts on the recovered site are online. Still, they can now wait until vSAN has fully resynced the virtual machine components. DRS can then be changed back to fully automated mode, allowing virtual machine migrations to conform to the VM/Host affinity rules.

DRS awareness of vSAN stretched clusters in vSAN 7 Update 2

DRS awareness of vSAN stretched clusters is built into vSAN 7 Update 2. There is no need to make changes in configuration or operating processes. It works with all affinity rules. The recommendation with previous versions of vSAN was to set DRS Automation to Manual or Partially Automated. With vSAN 7 Update 2, this can be set to Fully Automated to get the most benefits from DRS with vSAN stretched clusters.

vSphere's DRS, tightly integrated with vSAN Update 2, introduces a fully automated read locality solution for recovering from failures on a vSAN stretched cluster. The read locality information indicates the hosts the VM has full access to, and DRS uses this information when placing a VM on a host on vSAN Stretched Clusters. DRS migrates the VMs back to the primary site once the vSAN resynchronization is completed and the VM's data components have achieved full read locality. This DRS to be placed in fully automatic mode in case of site failures.

In case of partial site failures, if a VM loses read locality due to loss of data components equal to or greater than its Failures to Tolerate, vSphere DRS will identify the VMs that consume a very high read bandwidth and try to rebalance them to the secondary site. This way, we ensure that VMs with read-heavy workloads are not impacted during partial site failures. Once the primary site is back online and the data components have resynchronized, the VM is moved back to the site it is affinity to.



VM/Host Groups & Rules

VMware recommends enabling vSphere DRS to allow for the creation of Host-VM affinity rules.

VMware also recommends creating VM/Host groups & affinity rules, as well as using vSphere DRS to perform initial placement of VMs and to avoid unnecessary vMotion of VMs between sites.

Because a vSAN Stretched Cluster is still a single cluster, DRS is unaware of the fact that it is made up of different sites and it may decide to move virtual machines between them. The use of VM/Host Groups will allow administrators to “pin” virtual machines to sites, preventing unnecessary vMotions/migrations.

vSAN Stretched Clusters use Read Locality to ensure that reads only occur in the site the virtual machine resides on. In Hybrid vSAN Stretched Cluster configurations, the read cache is only warm on the site the VM resides on. Should the VM be migrated to the alternate site, the read case will have to be warmed. Ensuring VMs do not freely move between sites will overcome the need for cache to be warmed unnecessarily.

Note that vSAN Stretched Cluster has its own notion of a Preferred site. This is set up during the configuration and refers to which site takes over in the event of a split-brain. It has no bearing on virtual machine placement. It is used for the case where there is a partition between the Preferred and Non-Preferred Site while the vSAN Witness Host can talk to both sites. More detailed failure scenarios are discussed later in this document.

Host Groups





When configuring DRS with a vSAN Stretched Cluster, VMware recommends creating Host Affinity Groups and VM Affinity Groups.

Hosts in each site should be grouped to a site-centric Host Group. When paired with hosts in site-centric groups, VM's assigned to VM groups can easily be configured to determine where VM's are allowed to run.

Create VM/Host Group | Cluster ✕

| | |
|-------|---------------------------|
| Name: | Hosts-SiteA |
| Type: | Host Group ⬇ |

+ Add... ✕ Remove

| Members | |
|---|-------------------|
|  | host1.vmware.demo |
|  | host2.vmware.demo |
|  | host3.vmware.demo |
|  | host4.vmware.demo |

CANCEL OK

VM Groups

VM groups should also be created depending on where VMs are desired to be run.

Create VM/Host Group | Cluster

Name: VM-SiteA

Type: VM Group

+ Add... - Remove

Members

- APP1
- CRM
- FILE

CANCEL OK

VM Groups should include VMs that have similar placement requirements.

The example above shows a VM Group that includes several VMs. The naming of the VM Group suggests that these VMs will likely run on Site A.

A VM/Host Rule must be created for this to be the case. VM Groups do not natively perform any function other than grouping VMs.

Remembering that VMs must be assigned to a VM Group after deployment is essential. This can be accomplished in the vSphere Client or via API calls or PowerCLI script.

VM/Host Rules

When deploying virtual machines on a vSAN Stretched Cluster, in most cases, we wish the virtual machine to reside on the set of hosts in the selected host group. However, in the event of an entire site failure, we wish the virtual machines to be restarted on the surviving site.

VMware recommends implementing “should respect rules” in the VM/Host Rules configuration section to achieve this. vSphere HA may violate these rules in the case of a full-site outage. If “must rules” were implemented, vSphere HA does not violate the rule set, and this could potentially lead to service outages. vSphere HA will not restart the virtual machines, as they will not have the required affinity to start on the hosts in the other site. Thus, the recommendation to implement “should rules” will allow vSphere HA to restart the virtual machines in the other site.

Create VM/Host Rule

Cluster
✕

| | | |
|------|---------------------------|--|
| Name | Rule-SiteA-Should | <input checked="" type="checkbox"/> Enable rule. |
| Type | Virtual Machines to Hosts | |

Description:

Virtual machines that are members of the Cluster VM Group VMs-SiteA should run on host group Hosts-SiteA.

VM Group:

VMs-SiteA

Should run on hosts in group

Host Group:

Hosts-SiteA

CANCEL

OK

The vSphere HA Rule Settings are found in the VM/Host Rules section. This allows administrators to decide which virtual machines (that are part of a VM Group) are allowed to run on which hosts (that are part of a Host Group). It also allows an administrator to decide on how strictly “VM to Host affinity rules” are enforced.

As stated above, the VM to Host affinity rules should be set to “should respect” to allow the virtual machines on one site to be started on the hosts on the other site in case of a complete site failure. The “should rules” are implemented by clicking the Edit button in the vSphere HA Rule Settings at the bottom of the VM/Host Rules view, and setting VM to Host affinity rules to “vSphere HA should respect rules during failover”. By default vSphere HA will respect these rules when possible.

vSphere DRS communicates these rules to vSphere HA, and these are stored in a “compatibility list” governing allowed startup behavior. Note once again that with a full site failure, vSphere HA will be able to restart the virtual machines on hosts that violate the rules. Availability takes preference in this scenario.

Per-Site Policy Rule Considerations

With the introduction of Per-Site Policy Rules, VM/Hosts Group Rules are more important than ever to consider.

Misconfiguration

It is entirely possible to have a VM Storage Policy using the Affinity rule placing data in the one site, with a VM/Host Group Rule placing the VM in the alternate site.

Here is an illustration of such a configuration:

The screenshot shows the vSphere Cluster configuration page for 'Configure'. The left sidebar lists various configuration options, with 'VM/Host Rules' selected. The main area displays a table of VM/Host Rules:

| Name | Type | Enabled | Conflicts | Defined By |
|-------------------|------------------|---------|-----------|------------|
| Rule-SiteA-Should | Run VMs on Hosts | Yes | 0 | User |
| Rule-SiteB-Should | Run VMs on Hosts | Yes | 0 | User |

Below the table, the 'VM/Host Rule Details' section shows 'Virtual Machines that are members of the VM Group should run on hosts that are members of the Host Group.' It lists 'VMs-SiteA Group Members' (FILE, APP1, CRM) and 'Hosts-SiteA Group Members' (host2.vmware.demo, host1.vmware.demo, host4.vmware.demo, host3.vmware.demo). A red box highlights 'APP1' in the VMs-SiteA Group Members list, and a red arrow points from the 'Rule-SiteA-Should' rule to it. Another red arrow points from the text 'VM is running in the Preferred Site based on the VM/Host Rule assignment' to the 'APP1' entry.

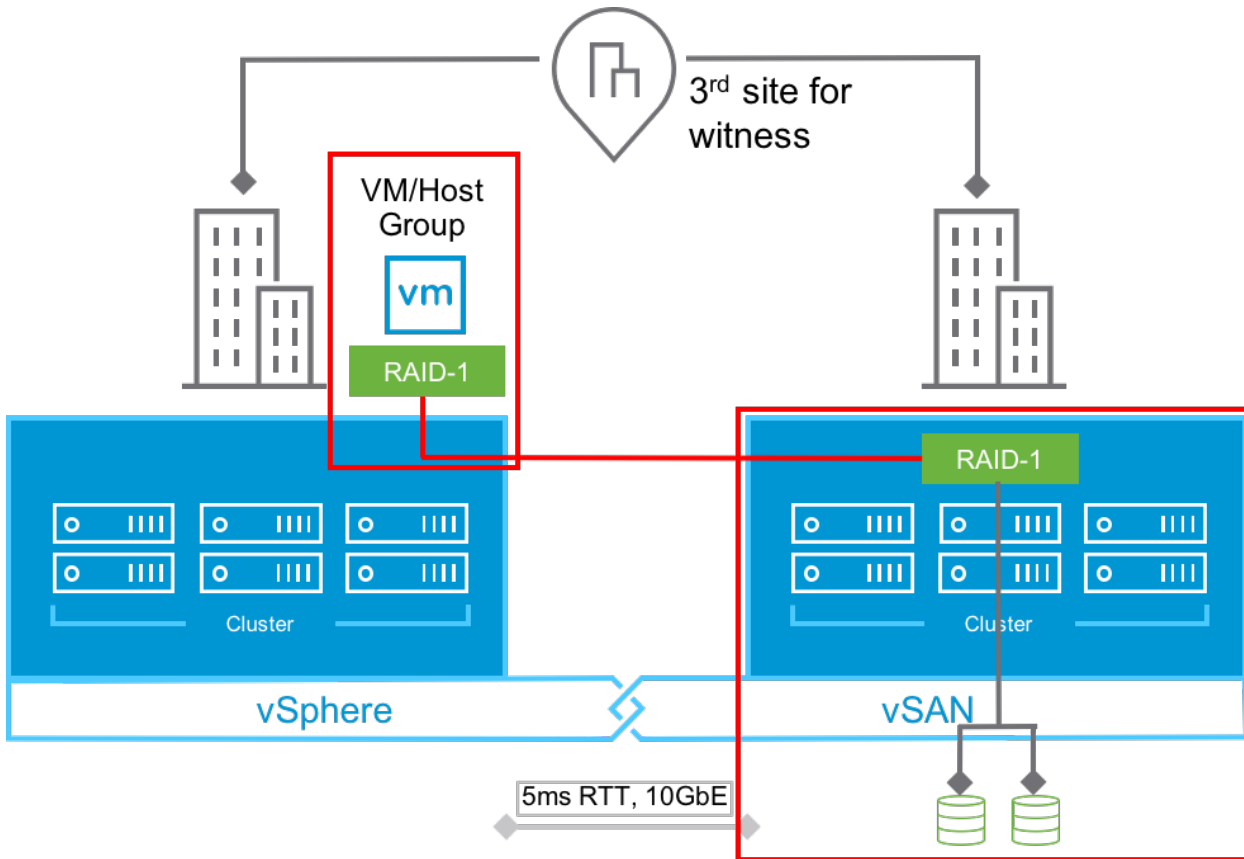
The screenshot shows the vSphere Monitor page for VM 'APP1'. The left sidebar lists various monitoring options, with 'Physical disk placement' selected. The main area displays 'Virtual Object Components' with a table showing component placement across hosts and fault domains:

| Type | Component State | Host | Fault Domain |
|--------------------------------------|-----------------|-------------------|--------------|
| Hard disk 1 (RAID 1) | | | |
| Component | Active | host8.vmware.demo | SiteB |
| Component | Active | host7.vmware.demo | SiteB |
| Witness | Active | host6.vmware.demo | SiteB |
| VM home (RAID 1) | | | |
| Component | Active | host7.vmware.demo | SiteB |
| Component | Active | host8.vmware.demo | SiteB |
| Witness | Active | host6.vmware.demo | SiteB |
| Virtual machine swap object (RAID 1) | | | |
| Witness | Active | host7.vmware.demo | SiteB |
| Component | Active | host6.vmware.demo | SiteB |
| Component | Active | host8.vmware.demo | SiteB |

A red box highlights the 'Fault Domain' column, which shows 'SiteB' for all components. A red arrow points from the text 'Data in Non-Preferred Site' to this column.

On a local network, this may not be a significant issue. In a Stretched Cluster configuration, with sites spread across large

geographical distances, this is considered a misconfiguration. This is because reads and writes must traverse the inter-site link when the VM does not run on the same site.



Reads & Writes must both traverse the inter-site link

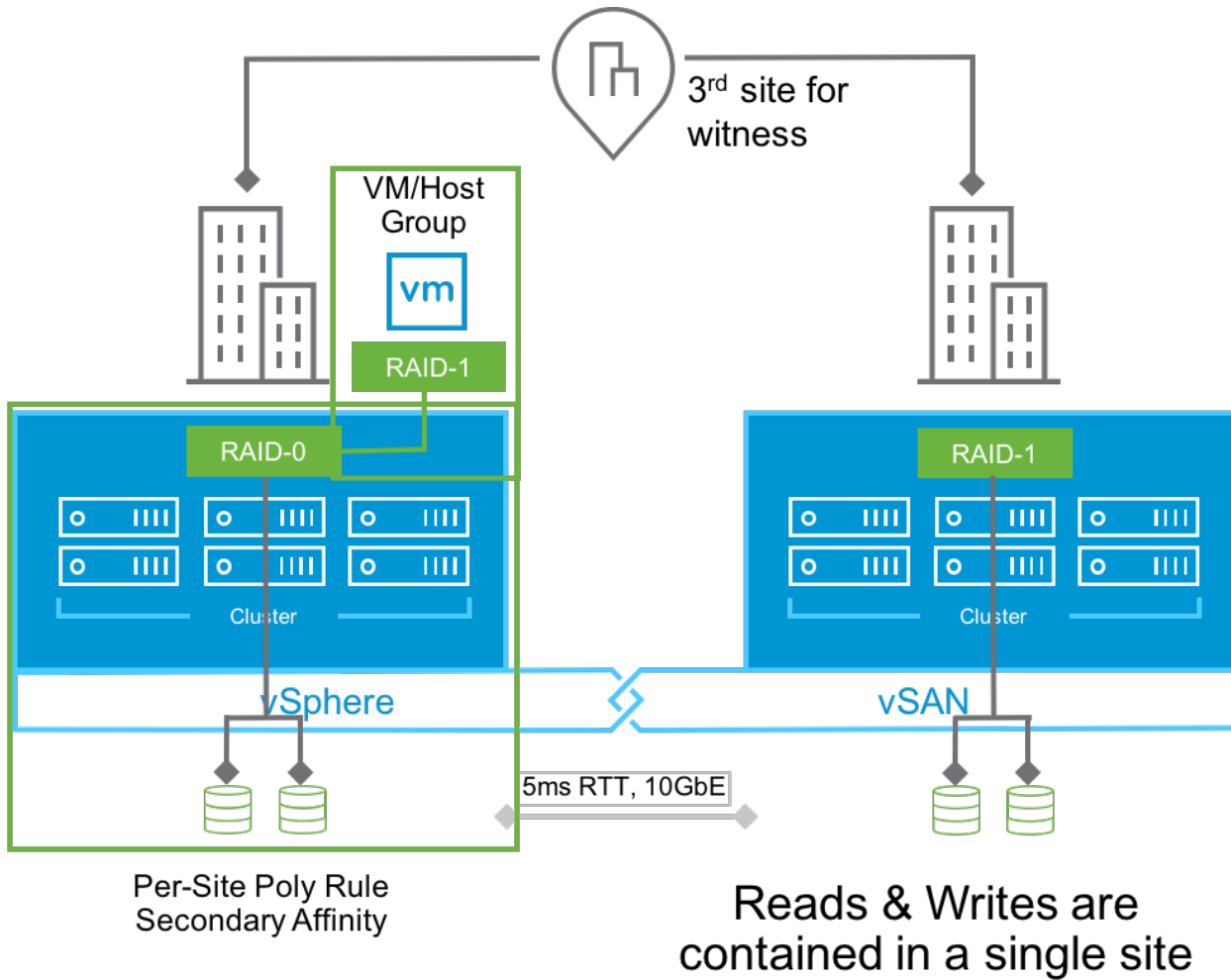
Per-Site Poly Rule
Secondary Affinity

Additional unnecessary bandwidth is consumed to operate the VM running on a site opposite where the data is stored. This bandwidth should stay within the same site to ensure lower bandwidth utilization.

When the alternate site is disconnected, the VM will no longer have access to its vmdk and will essentially become a zombie VM.

Proper Configuration

A proper configuration includes VM/Host Group Rules that align with the Affinity Rules assigned to VMs by their corresponding VM Storage Policies.



Setting proper VM/Host Group Rules and VM Storage Policy Affinity Rules are beneficial for several reasons

- Bandwidth is not unnecessarily sent across the inter-site link
- Lower inter-site bandwidth utilization
- In the situation where the alternate site is disconnected, the VM will continue to have access to its vmdk.

Summary

It is important to ensure proper rules are in place to maintain a properly utilized configuration. VMware recommends VM/Host Group rules along with Affinity rules for VM Storage Policies for VMs that are only being stored on one of the two data sites.

Installation

Installing a vSAN Stretched Cluster is almost identical to how Fault Domains were implemented in earlier vSAN versions, with a couple of additional steps. This part of the guide will walk the reader through a stretched cluster configuration.

Before You Start

Before installing a vSAN Stretched Cluster, several essential features are specific to stretch cluster environments that are important to understand before installing a vSAN Stretched Cluster.

What is a Preferred Site?

The Preferred Site is the site that vSAN wishes to remain running when there is a failure, and the sites can no longer communicate. One might say that the Preferred Site is expected to have the most reliability.

Since virtual machines can run on any of the two sites if network connectivity is lost between site 1 and site 2, but both still have connectivity to the Witness, the Preferred Site is the one that survives and its components remains active. At the same time, the storage on the Non-Preferred Site is marked as down, and components on that site are marked as absent.

What is Read Locality?

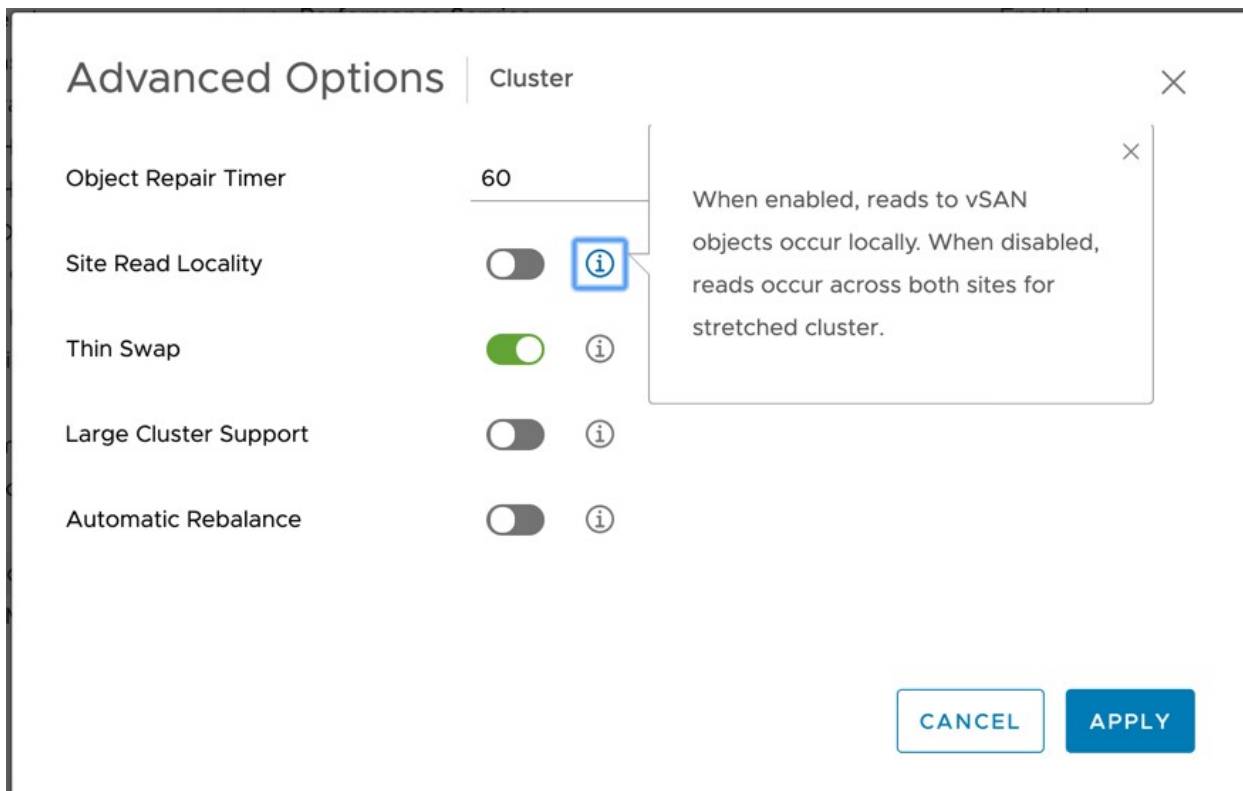
Since virtual machines deployed on a vSAN Stretched Cluster will have compute on one site but a copy of the data on both sites, vSAN will use a read locality algorithm to read 100% from the data copy on the local site, i.e. same site where the virtual machine resides. This is not the regular vSAN algorithm, which reads across all components of the vSAN object.

This algorithm for vSAN Stretched Clusters reduces the latency incurred on read operations.

If latency is less than 5ms and there is enough bandwidth between the sites, read locality could be disabled. However, please note that disabling read locality means that the read algorithm reverts to the round-robin mechanism, and for vSAN Stretched Clusters, 50% of the read requests will be sent to the remote site. This is a significant consideration for sizing the network bandwidth.

For more details, please refer to the network bandwidth sizing between the two main sites.

Read locality can be disabled/re-enabled in the vSphere Client.



The advanced parameter `VSAN.DOMOwnerForceWarmCache` can also be set on each host in a cluster from the command line:

```
esxcfg-advcfg -s 1 /VSAN/DOMOwnerForceWarmCache
```

Read locality is enabled by default when vSAN Stretched Cluster is configured – it should only be disabled under the guidance of VMware’s Global Support Services organization, and only when extremely low latency is available across all sites.

Witness Host Must not be part of the vSAN Cluster

When configuring your vSAN Stretched Cluster, only data hosts must be in the cluster object in vCenter. The vSAN Witness Host must remain outside of the cluster, and must not be added to the cluster at any point.

The illustration below shows several vSAN Witness Hosts. Witness3.demo.local is will not function properly because it has been added to a Cluster.

The screenshot displays the vSphere Client interface for a vSAN Witness-Datacenter. The left-hand navigation pane shows a tree view of the vSAN environment. Under the 'Witness-Datacenter' folder, there are two sub-folders: 'Central' and 'Local'. The 'Central' folder contains 'witness.demo.central' with a green checkmark. The 'Local' folder contains three hosts: 'witness1.demo.local', 'witness2.demo.local', and 'witness4.demo.local', all with green checkmarks. Below the 'Local' folder is a 'Cluster' folder containing 'witness3.demo.local' with a red prohibition sign, indicating it is incorrectly placed in a cluster.

The right-hand pane shows the 'Witness-Datacenter' summary. It includes a 'Summary' tab with a table of statistics:

| | |
|-------------------|---|
| Hosts: | 5 |
| Virtual Machines: | 0 |
| Clusters: | 1 |
| Networks: | 1 |
| Datastores: | 0 |

Below the summary is a 'Custom Attributes' table with two columns: 'Attribute' and 'Value'.

The other vSAN Witness Hosts will all function properly because they have been added to folders.

vSAN Witness Hosts can reside in the root of a Datacenter object or a folder but may not reside in a vSphere Cluster.

vSAN Health Check Plugin & Stretched Clusters

vSAN Health Checks for Stretched Cluster Configurations

Select the Cluster object in the vCenter inventory, click on Monitor > vSAN > Health. Ensure the Stretched Cluster Health Checks pass when the cluster is fully configured.

Note: The Stretched Cluster checks will not be visible until the Stretched Cluster configuration is completed.

Cluster | ACTIONS

Summary Monitor Configure Permissions Hosts VMs Datastores Networks Updates

Health (Last checked: 10/22/2019, 4:21:33 PM) RETEST

- Cluster
- Network
- Stretched cluster
 - Unexpected number of fault domains
 - Witness host within vCenter cluster
 - Invalid preferred fault domain on witness host
 - Unsupported host version
 - Unicast agent configuration inconsistent
 - Unicast agent not configured
 - Witness host not found
 - Invalid unicast agent
 - Preferred fault domain unset
 - Witness host fault domain misconfigured
 - No disk claimed on witness host
 - Site latency health

Details and setup of the vSAN Witness Appliance

vSAN Stretched Cluster supports the use of a vSAN Witness Appliance. This section will cover details about the vSAN Witness appliance and instructions for how to set it up.

vSAN Witness Appliance Details

VMware vSAN Stretched Cluster supports using a vSAN Witness Appliance as the Witness host. This is available as an OVA (Open Virtual Appliance) from VMware. However this vSAN Witness Appliance needs to reside on a physical ESXi host, which requires some special networking configuration.

Networking

The vSAN Witness Appliance contains two 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.

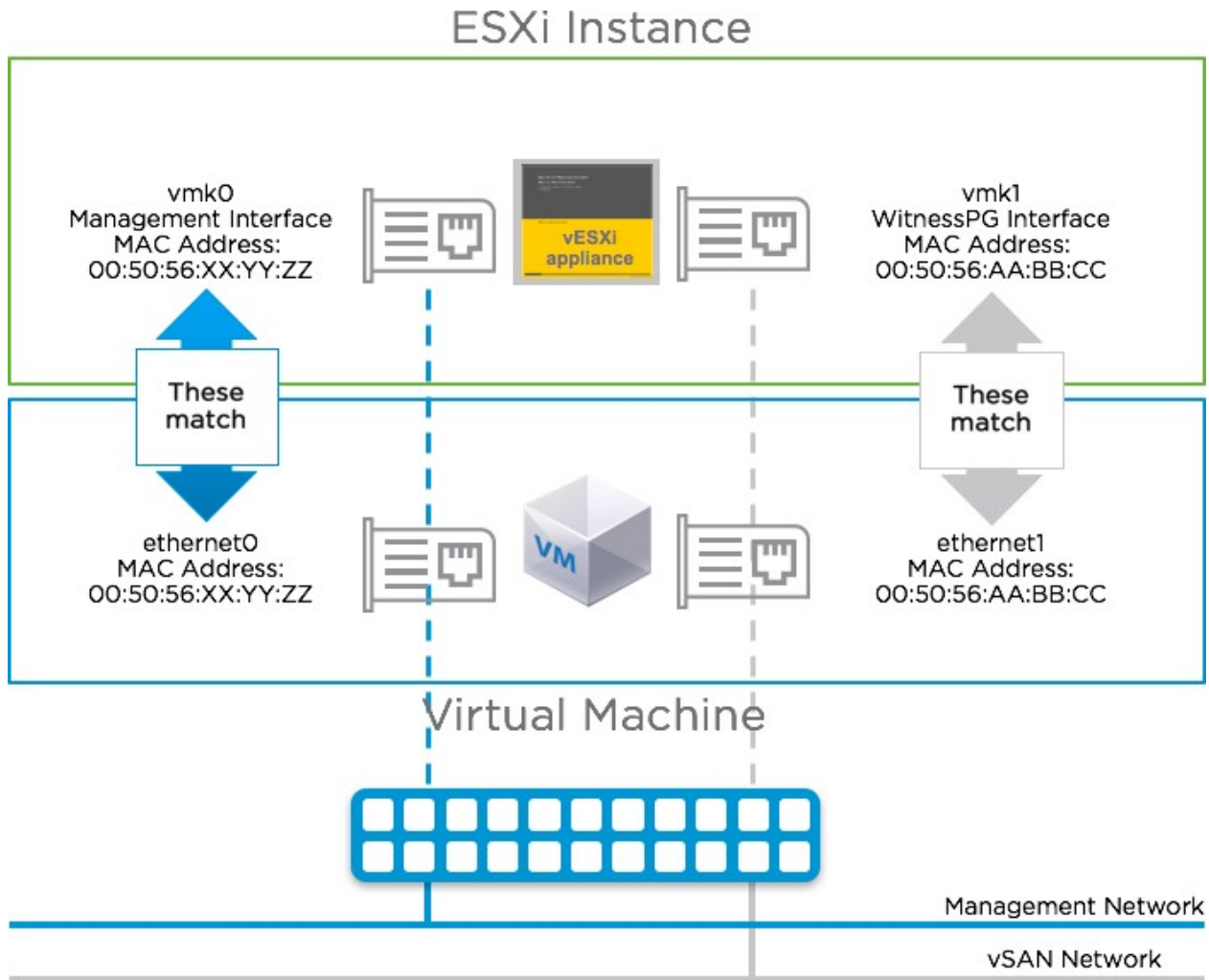
The Management VMkernel interface could be tagged to include vSAN Network traffic as well as Management traffic. In this case, vmk0 would require connectivity to vCenter Server and the vSAN Network.

A Note About Promiscuous Mode

In many nested ESXi environments, a promiscuous mode is recommended 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. VMware does not support Nested ESXi deployments other than the vSAN Witness Appliance.

The vSAN Witness Appliance is a nested ESXi installation tailored for use with vSAN Stretched Clusters.

The MAC addresses of the VMkernel interfaces vmk0 & vmk1 are configured to match the MAC addresses of the vSAN Witness Appliance host's NICs, vmnic0 and vmnic1. Because of this, packets destined for either the Management VMkernel interface (vmk0) or the WitnessPG VMkernel interface are not dropped.

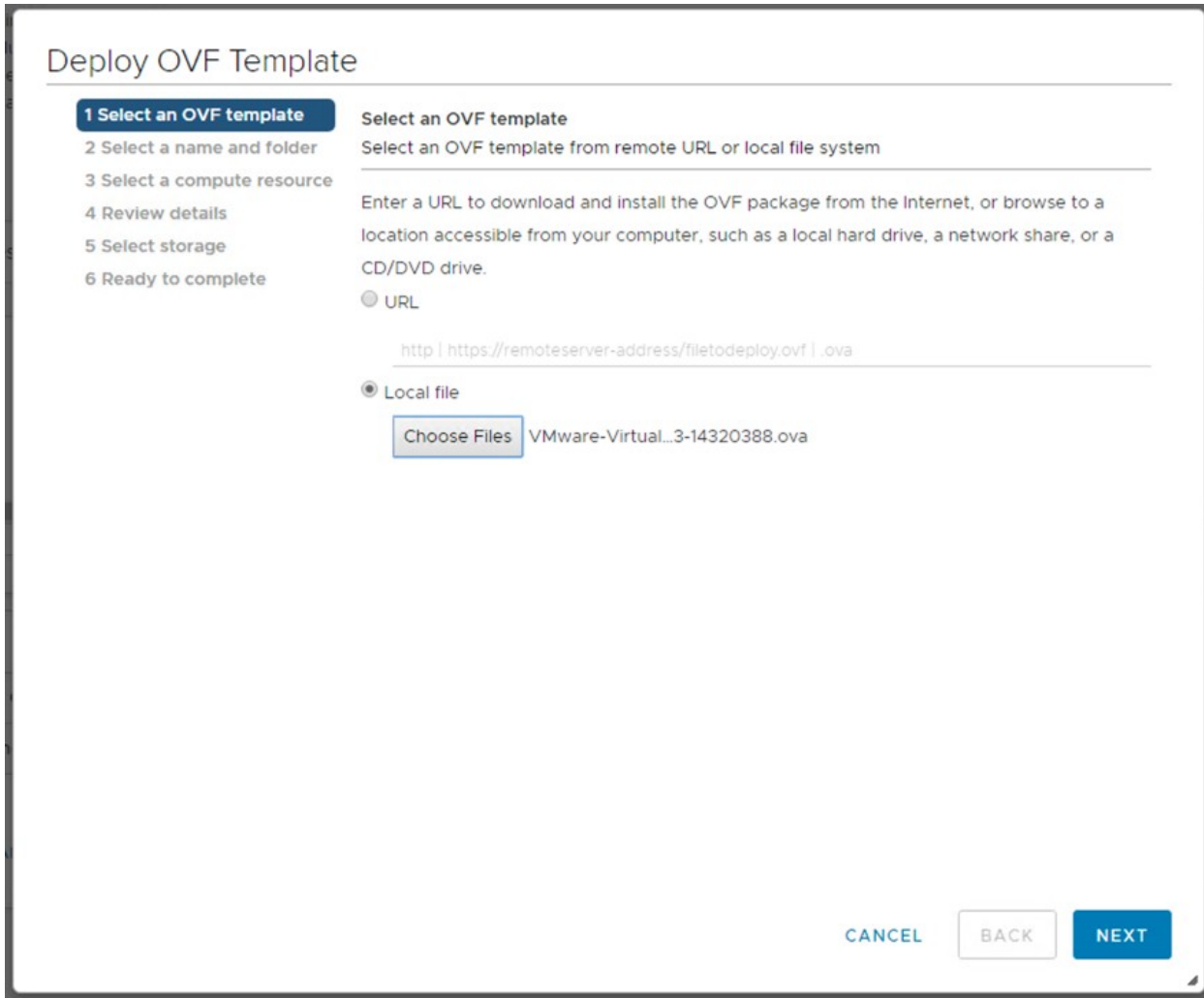


Because of this, promiscuous mode is not required when using a vSAN Witness Appliance.

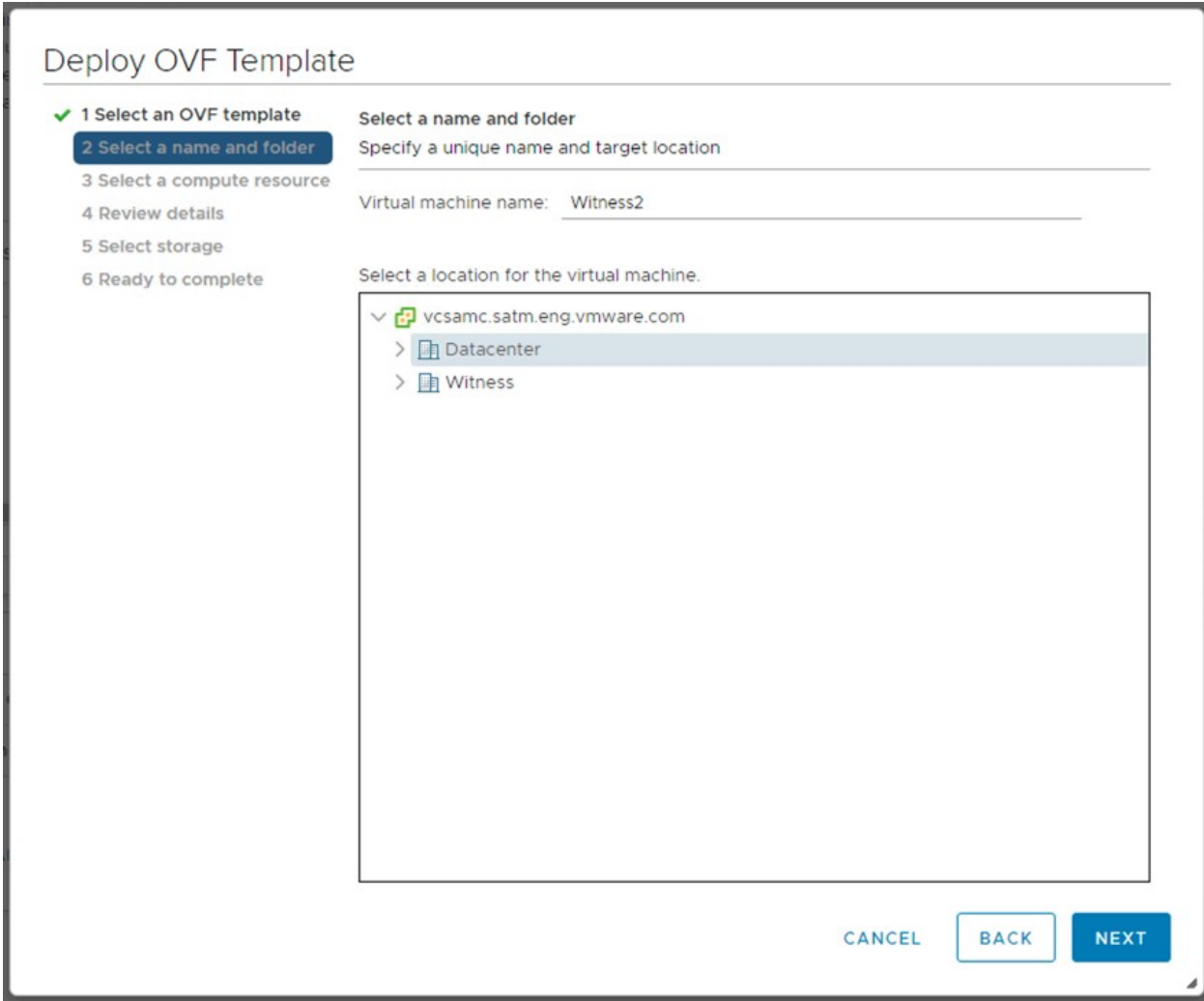
Setup Step 1: Deploy the vSAN Witness Appliance

The vSAN Witness Appliance must be deployed on a different infrastructure than the Stretched Cluster. This step will cover deploying the vSAN Witness Appliance to a different cluster.

The first step is to download and deploy the vSAN Witness Appliance or to deploy it directly via a URL, as shown below. In this example, it has been downloaded:



Select a Datacenter for the vSAN Witness Appliance to be deployed to and provide a name (Witness2 or something similar).



Select a cluster for the vSAN Witness Appliance to reside on.

Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- 3 Select a compute resource**
- 4 Review details
- 5 Select storage
- 6 Ready to complete

Select a compute resource
Select the destination compute resource for this operation

- ∨ Datacenter
 - > Cluster
 - > Management

Compatibility

✓ Compatibility checks succeeded.

CANCEL BACK NEXT

Review the details of the deployment and press next to proceed.

Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- 4 Review details**
- 5 License agreements
- 6 Configuration
- 7 Select storage
- 8 Select networks
- 9 Customize template
- 10 Ready to complete

Review details

Verify the template details.

⚠ The OVF package contains advanced configuration options, which might pose a security risk. Review the advanced configuration options below. Click next to accept the advanced configuration options.

| | |
|---------------------|---|
| Publisher | VMware\, Inc. (Trusted certificate) |
| Product | VMware vSAN Witness Appliance |
| Version | 6.7 |
| Vendor | VMware, Inc. |
| Description | VMware vSAN Witness Appliance |
| Download size | 455.4 MB |
| Size on disk | Unknown (thin provisioned) |
| | 1.4 TB (thick provisioned) |
| Extra configuration | svga.maxWidth = 720 svga.maxHeight = 480 |

CANCEL

BACK

NEXT

The license must be accepted to proceed.

Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- ✓ 4 Review details
- 5 License agreements**
- 6 Configuration
- 7 Select storage
- 8 Select networks
- 9 Customize template
- 10 Ready to complete

License agreements

The end-user license agreement must be accepted.

Read and accept the terms for the license agreement.

VMWARE END USER LICENSE AGREEMENT

PLEASE NOTE THAT THE TERMS OF THIS END USER LICENSE AGREEMENT SHALL GOVERN YOUR USE OF THE SOFTWARE, REGARDLESS OF ANY TERMS THAT MAY APPEAR DURING THE INSTALLATION OF THE SOFTWARE.

IMPORTANT-READ CAREFULLY: BY DOWNLOADING, INSTALLING, OR USING THE SOFTWARE, YOU (THE INDIVIDUAL OR LEGAL ENTITY) AGREE TO BE BOUND BY THE TERMS OF THIS END USER LICENSE AGREEMENT ("EULA"). IF YOU DO NOT AGREE TO THE TERMS OF THIS EULA, YOU MUST NOT DOWNLOAD, INSTALL, OR USE THE SOFTWARE, AND YOU MUST DELETE OR RETURN THE UNUSED SOFTWARE TO THE VENDOR FROM WHICH YOU ACQUIRED IT WITHIN THIRTY (30) DAYS AND REQUEST A REFUND OF THE LICENSE FEE, IF ANY, THAT YOU PAID

I accept all license agreements.

CANCEL BACK NEXT

At this point, a decision must be made regarding the expected size of the Stretched Cluster configuration. There are three options offered. If you expect the number of VMs deployed on the vSAN Stretched Cluster to be ten or fewer, select the Tiny configuration. If you expect to deploy more than 10 VMs, but less than 500 VMs, then the Normal (default option) should be chosen. For more than 500 VMs, choose the Large option. On selecting a particular configuration, the resources consumed by the appliance and displayed in the wizard (CPU, Memory, and Disk):

Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- ✓ 4 Review details
- ✓ 5 License agreements
- 6 Configuration**
- 7 Select storage
- 8 Select networks
- 9 Customize template
- 10 Ready to complete

Configuration
Select a deployment configuration

| | Description |
|---|---|
| <input type="radio"/> Tiny (10 VMs or fewer) | |
| <input checked="" type="radio"/> Medium (up to 500 VMs) | Configuration for Medium vSAN Deployments of up to 500 VMs * 2 vCPUs * 16GB vRAM * 1x 12GB ESXi Boot Disk * 1x 350GB Magnetic Disk * 1x 10GB Solid-State Disk * Maximum of 21K Witness Components |
| <input type="radio"/> Large (more than 500 VMs) | |

3 Items

CANCEL BACK NEXT

Select a datastore for the vSAN Witness Appliance. This will be one of the datastores available to the underlying physical host. Consider when the vSAN Witness Appliance is deployed as “thin”, as thin VMs may grow over time, so ensure there is enough capacity on the selected datastore. Remember that the vSAN Witness Appliance is not supported on vSAN Stretched Cluster datastores.

Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- ✓ 4 Review details
- ✓ 5 License agreements
- ✓ 6 Configuration
- 7 Select storage**
- 8 Select networks
- 9 Customize template
- 10 Ready to complete

Select storage
Select the storage for the configuration and disk files

Encrypt this virtual machine (Requires Key Management Server)

Select virtual disk format: As defined in the VM storage policy ▾

VM Storage Policy: vSAN Default Storage Policy ▾

| Name | Capacity | Provisioned | Free |
|---------------------------------------|----------|-------------|---------|
| ▲ Storage Compatibility: Compatible | | | |
| vsanDatastore | 6.55 TB | 5.34 TB | 5.23 TB |
| ▲ Storage Compatibility: Incompatible | | | |
| logs | 4 GB | 192.23 MB | 3.81 GB |

Compatibility

✓ Compatibility checks succeeded.

CANCEL
BACK
NEXT

Select a network for the Management Network and for the Witness Network.

Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- ✓ 4 Review details
- ✓ 5 License agreements
- ✓ 6 Configuration
- ✓ 7 Select storage
- 8 Select networks**
- 9 Customize template
- 10 Ready to complete

Select networks
Select a destination network for each source network.

| Source Network | Destination Network |
|--------------------|---------------------|
| Witness Network | DSwitch-Router |
| Management Network | VM Network |

2 items

IP Allocation Settings

IP allocation: Static - Manual

IP protocol: IPv4

CANCEL BACK NEXT

Give a root password for the vSAN Witness Appliance:

Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- ✓ 4 Review details
- ✓ 5 License agreements
- ✓ 6 Configuration
- ✓ 7 Select storage
- ✓ 8 Select networks
- 9 Customize template
- 10 Ready to complete

Customize template
Customize the deployment properties of this software solution.

✓ All properties have valid values ✕

| Uncategorized | 1 settings |
|---------------|--|
| Root password | Set password for root account. A valid password must be at least 7 characters long and must contain a mix of upper and lower case letters, digits, and other characters. You can use a 7 character long password with characters from at least 3 of these 4 classes. An upper case letter that begins the password and a digit that ends it do not count towards the number of character classes used. Password <input style="width: 100px;" type="password" value="....."/> Confirm <input style="width: 100px;" type="password" value="....."/> Password |

CANCEL BACK NEXT

At this point, the vSAN Witness Appliance is ready to be deployed. It will need to be powered on manually via the vSphere web client UI later:

Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- ✓ 4 Review details
- ✓ 5 License agreements
- ✓ 6 Configuration
- ✓ 7 Select storage
- ✓ 8 Select networks
- ✓ 9 Customize template
- 10 Ready to complete**

Ready to complete
Click Finish to start creation.

| | |
|------------------------|--|
| Provisioning type | Deploy from template |
| Name | Witness2 |
| Template name | VMware-VirtualSAN-Witness-6.7.0.update03-14320388 |
| Download size | 455.4 MB |
| Size on disk | 1.4 TB |
| Folder | Datacenter |
| Resource | SCD |
| Storage mapping | 1 |
| All disks | Policy: vSAN Default Storage Policy; Datastore: vsanDatastore; Format: As defined in the VM storage policy |
| Network mapping | 2 |
| Witness Network | DSwitch-Nested |
| Management Network | VM Network |
| IP allocation settings | |
| IP protocol | IPV4 |
| IP allocation | Static - Manual |

CANCEL
BACK
FINISH

Once the vSAN Witness Appliance is deployed and powered on, select it in the vSphere web client UI and begin the next steps in the configuration process.

Setup Step 2: vSAN Witness Appliance Management

Once the vSAN Witness Appliance has been deployed, select it in the vSphere web client UI, open the console.

The console of the vSAN Witness Appliance should be access to add the correct networking information, such as IP address and DNS, for the management network.

On launching the console, unless you have a DHCP server on the management network, the landing page of the DCUI will look something similar to the following:

```

VMware ESXi 6.7.0 (UMKernel Release Build 14320388)

VMware, Inc. VMware Virtual Platform

2 x Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz
16 GiB Memory

To manage this host go to:
http://169.254.49.53/ (Waiting for DHCP...)
http://[fe80::250:56ff:fea1:6e8f]/ (STATIC)

<F2> Customize System/View Logs                <F12> Shut Down/Restart

```

Use the <F2> key to customize the system. The root login and password will need to be provided at this point. This is the root password that was added during the OVA deployment earlier.

Select the Network Adapters view. There will be two network adapters, each corresponding to the network adapters on the virtual machine. You should note that the MAC address of the network adapters from the DCUI view match the MAC address of the network adapters from the virtual machine view. Because of this match, there is no need to use promiscuous mode on the network, as discussed earlier.

Select vmnic0, and if you wish to view further information, select the key <D> to see more details.

```

Configure Management Network      Network Adapters

Network Adapters

Select the adapters for this host's default management network
connection. Use two or more adapters for fault-tolerance and
load-balancing.

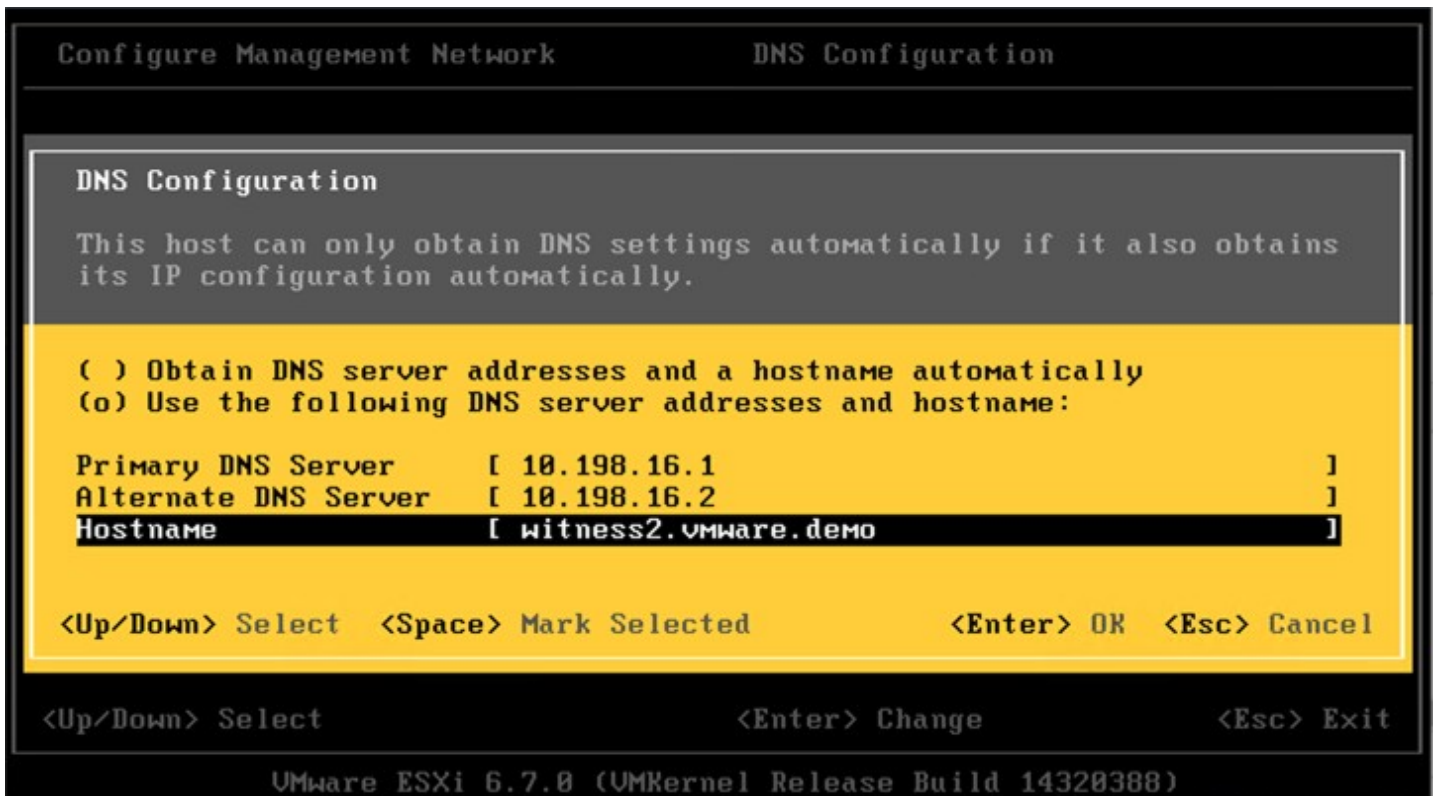
Device Name  Hardware Label (MAC Address)  Status
[X] vmnic0   Ethernet0 (...0:56:a1:6e:8f)  Disconnected (...)
[ ] vmnic1   Ethernet1 (...0:56:a1:f4:f5)  Disconnected (...)

<D> View Details  <Space> Toggle Selected      <Enter> OK  <Esc> Cancel

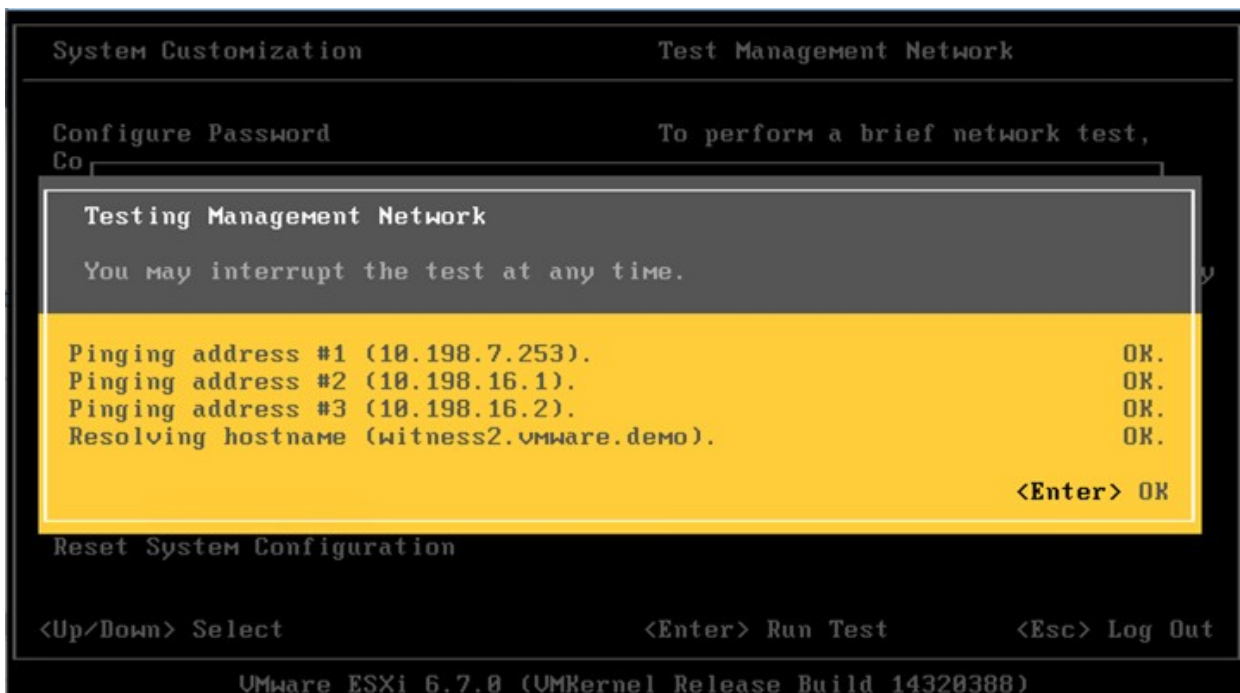
VMware ESXi 6.7.0 (UMKernel Release Build 14320388)

```

The next step is to configure DNS. A primary DNS server should be added and an optional alternate DNS server can also be added. The FQDN, fully qualified domain name, of the host should also be added at this point.



One final recommendation is to do a test of the management network. One can also try adding the IP address of the vCenter server at this point just to make sure that it is also reachable.



When all the tests have passed and the FQDN is resolvable, administrators can move onto the next step of the configuration, which is adding the vSAN Witness Appliance ESXi instance to the vCenter server.

Setup Step 3: Add Witness to vCenter Server

There is no difference in adding the vSAN Witness Appliance ESXi instance to the vCenter server compared to adding physical ESXi hosts. However, there are some exciting items to highlight during the process. The first step is to provide the name of the Witness. In this example, the vCenter server manages multiple data centers, so we are adding the host to the Witness data center.

Add Host

- 1 Name and location
- 2 Connection settings
- 3 Host summary
- 4 Assign license
- 5 Lockdown mode
- 6 VM location
- 7 Ready to complete

Name and location
Enter the name or IP address of the host to add to vCenter Server.

| | |
|--------------------------|----------------------|
| Host name or IP address: | witness2.vmware.demo |
| Location: | Witness |

CANCEL BACK NEXT

Provide appropriate credentials. In this example, the root user and password.

Add Host

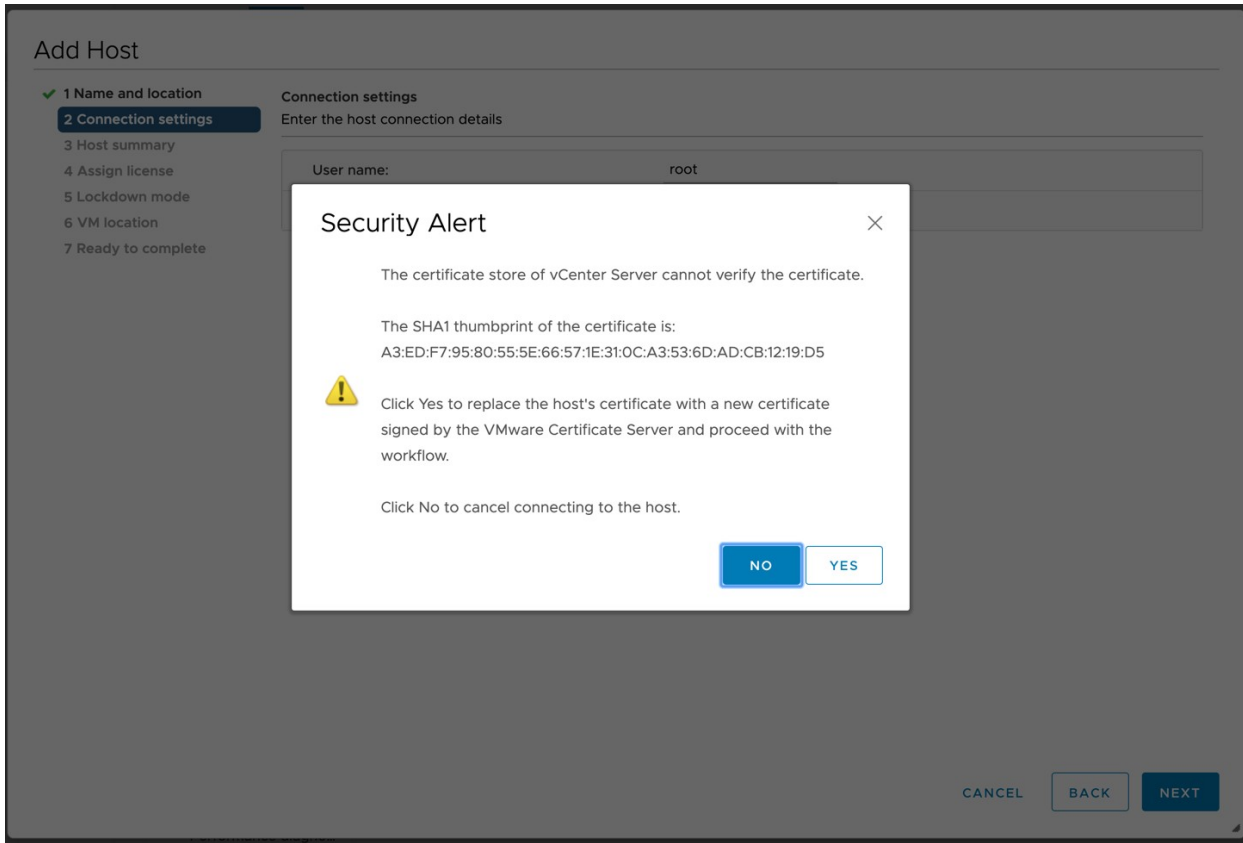
- ✓ 1 Name and location
- 2 Connection settings
- 3 Host summary
- 4 Assign license
- 5 Lockdown mode
- 6 VM location
- 7 Ready to complete

Connection settings
Enter the host connection details

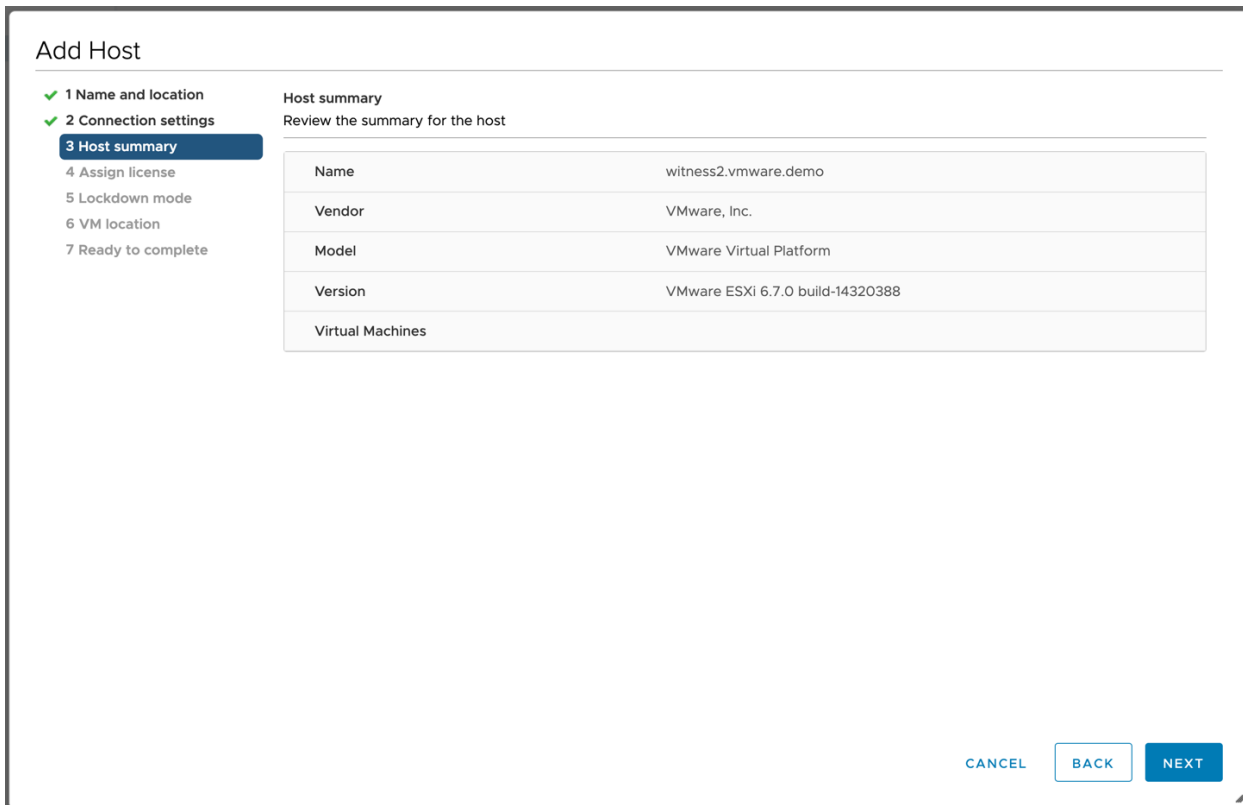
| | |
|------------|-------|
| User name: | root |
| Password: | ***** |

CANCEL BACK NEXT

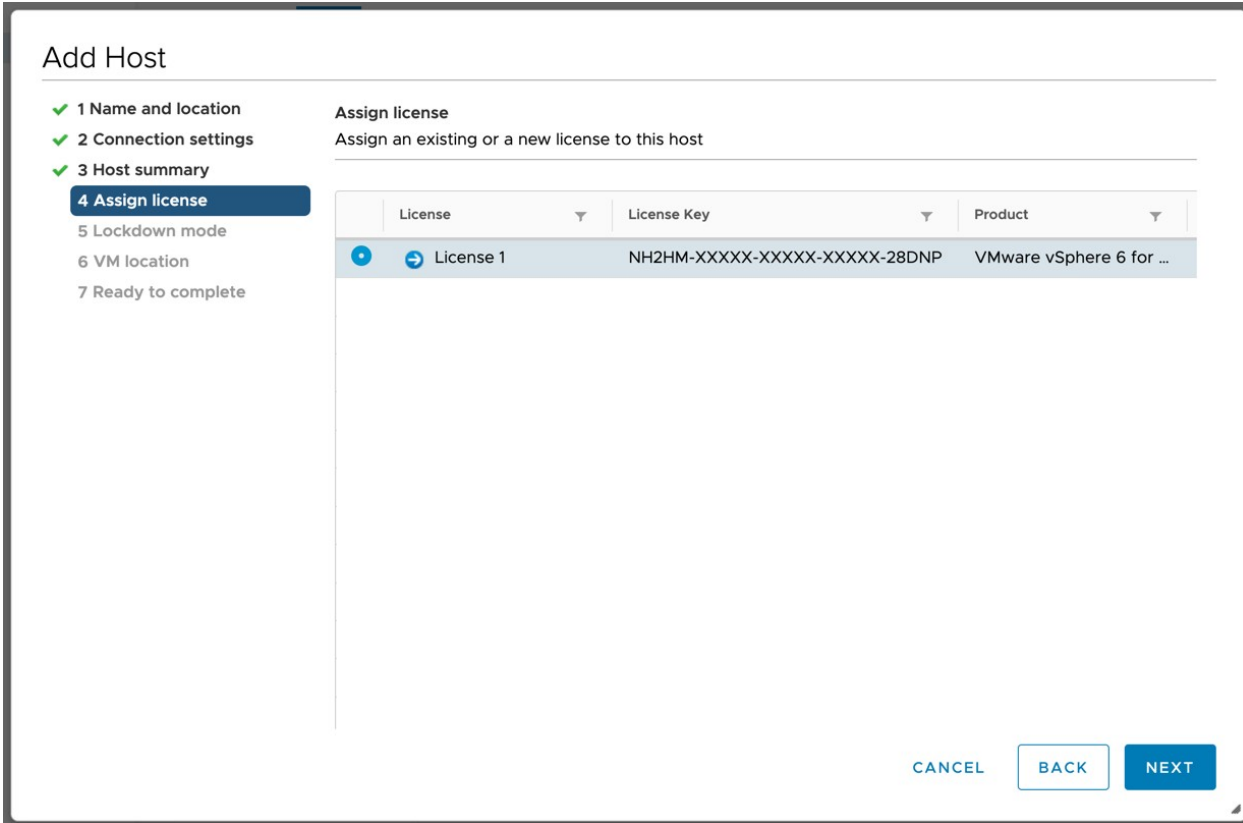
Acknowledge the certificate warning:



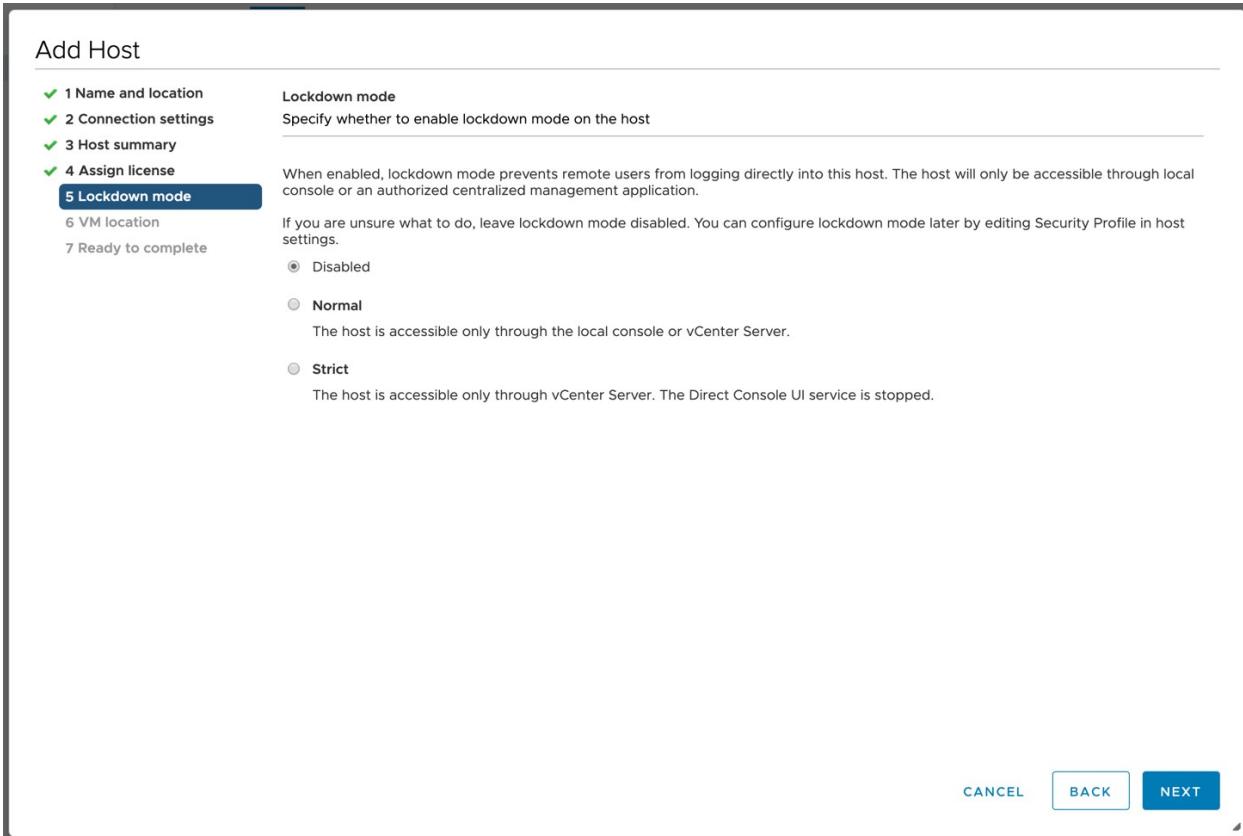
There should be no virtual machines on the vSAN Witness Appliance. Note: It can never run VMs in a vSAN Stretched Cluster configuration. Note also the mode: VMware Virtual Platform. Note also that build numbers may differ to the one shown here.



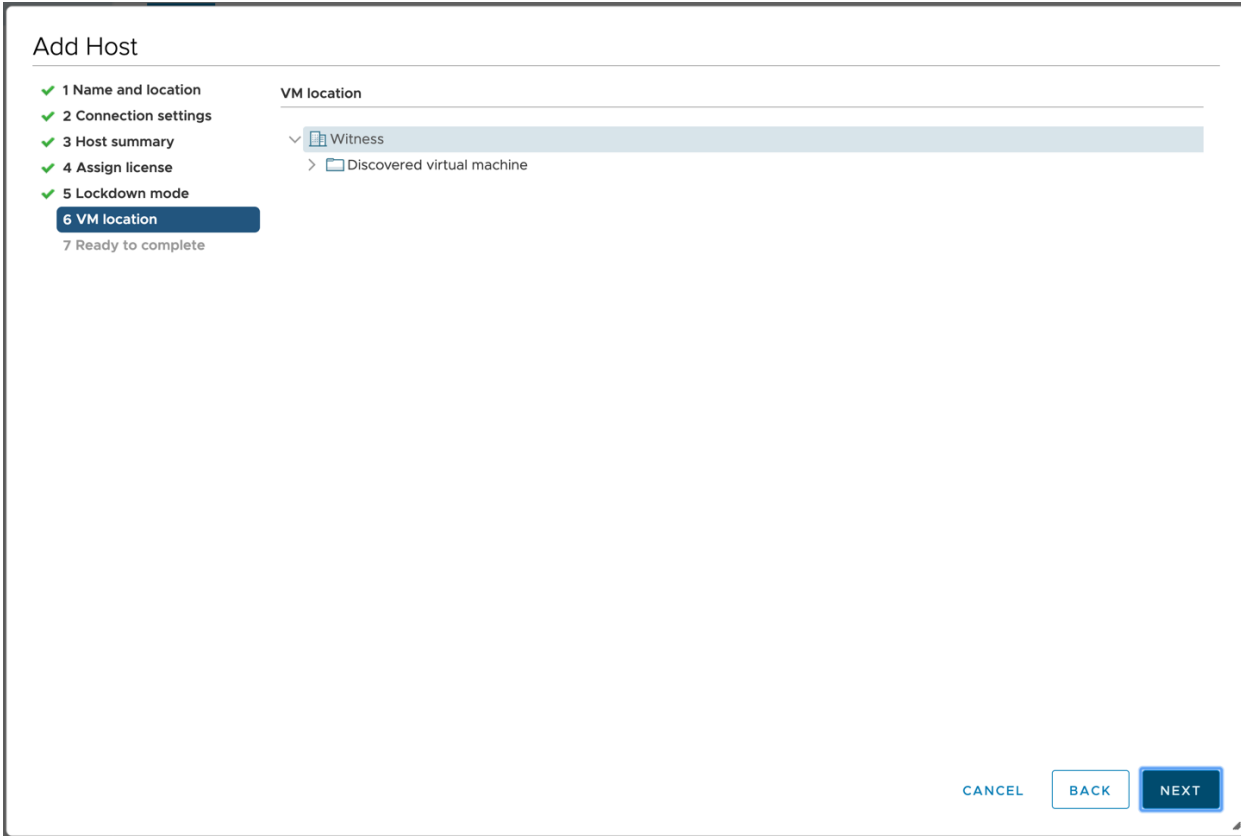
The vSAN Witness Appliance also comes with its own license. You do not need to consume vSphere licenses for the witness appliance:



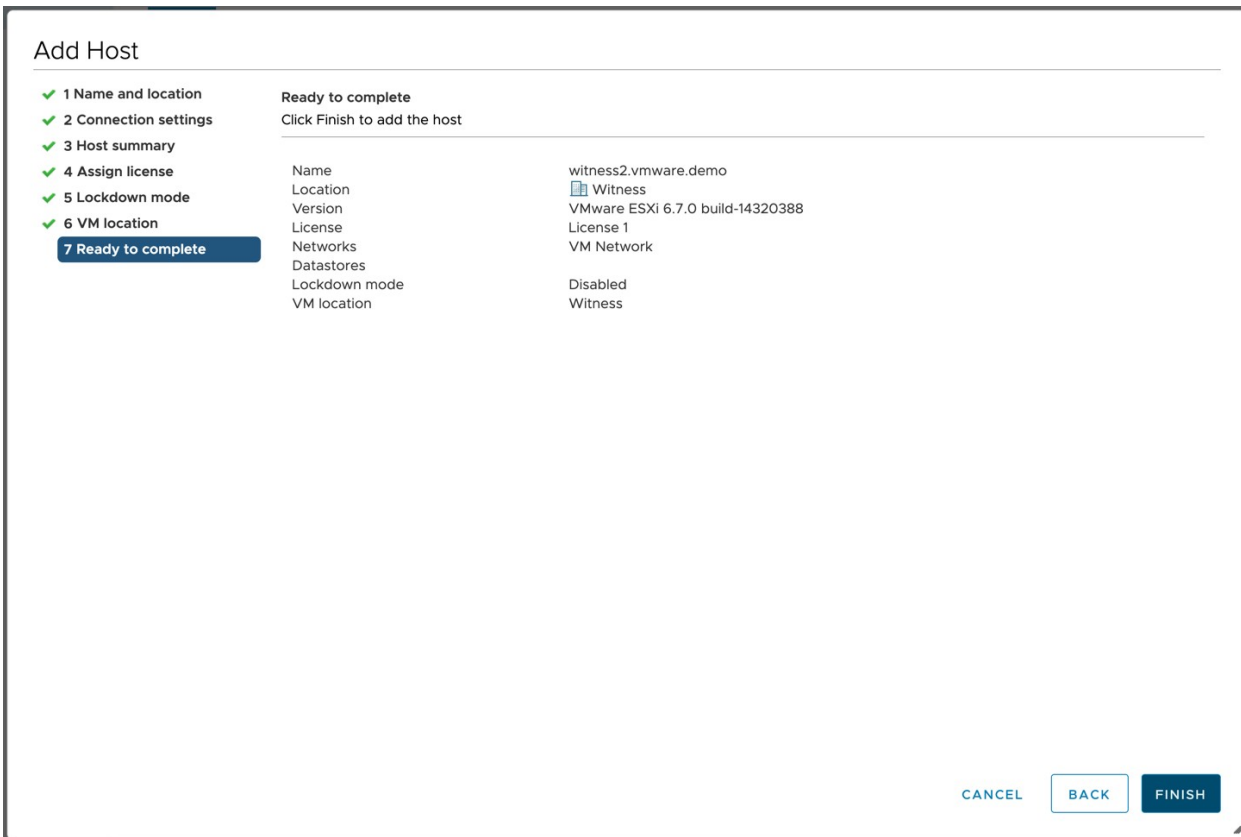
Lockdown mode is disabled by default. Depending on the policies in use at a customer’s site, the administrator may choose a different mode.



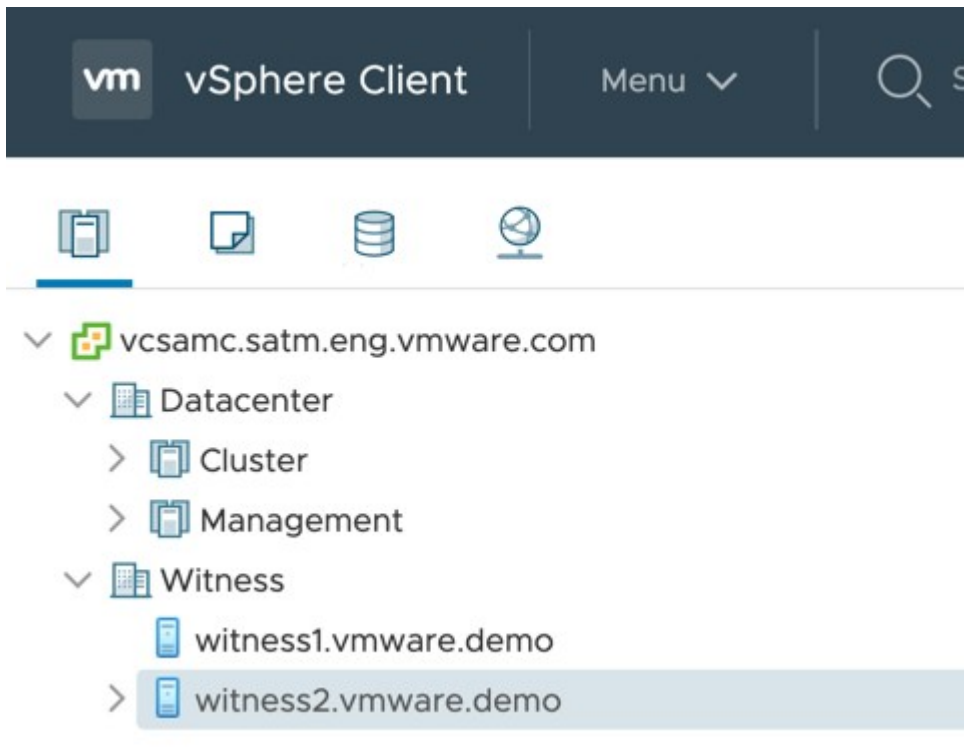
Choose the datacenter that the vSAN Witness Host is being added to for VM Location.



Click Finish when ready to complete the addition of the Witness to the vCenter server:



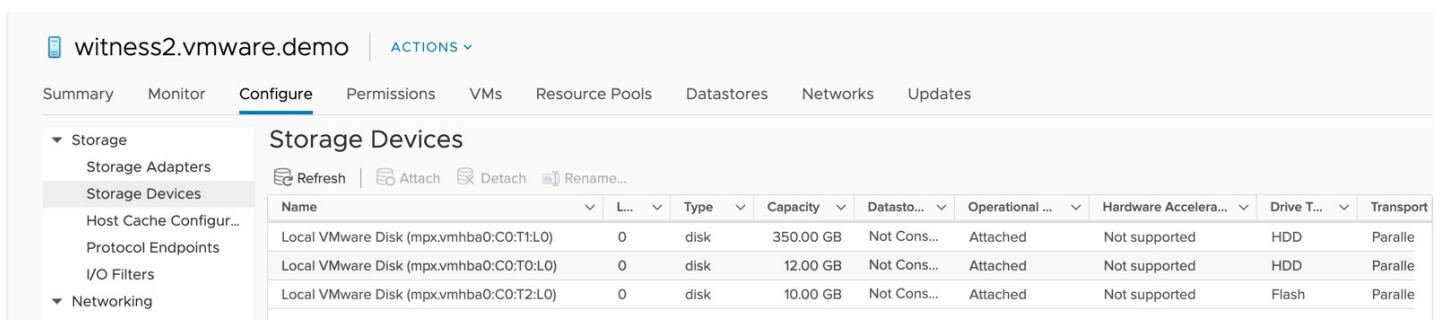
One final item of note is the appearance of the vSAN Witness Appliance ESXi instance in the vCenter inventory. It has a light blue shading, to differentiate it from standard ESXi hosts. It might be a little difficult to see in the screenshot below, but it should be visible in your infrastructure.



One final recommendation is to verify that the settings of the vSAN Witness Appliance match the Tiny, Normal or Large configuration selected during deployment.

Deployments should have:

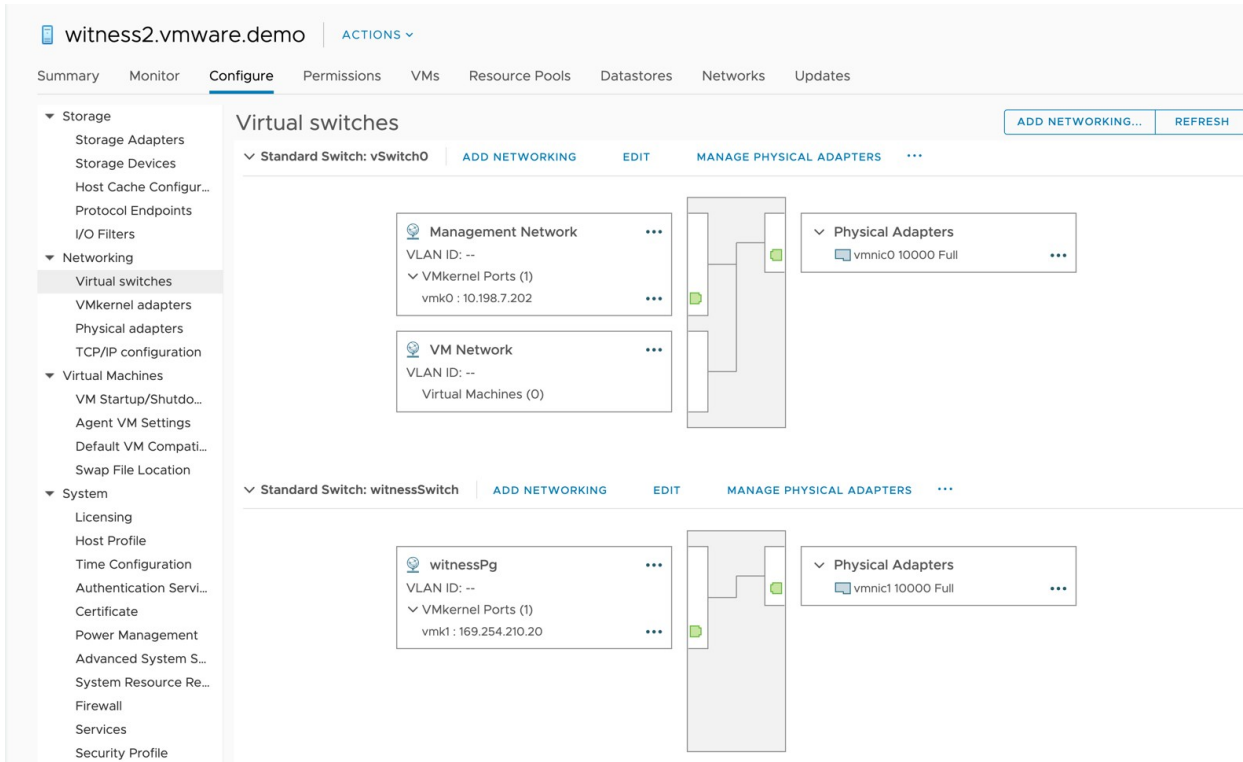
- Boot Device
 - 12GB HDD
- Cache Device to be configured later
 - 10GB Flash Drive
- Capacity Device(s) to be configured later
 - Tiny: A 15GB HDD
 - Normal: A 350GB HDD
 - Large: 3x 350GB HDD



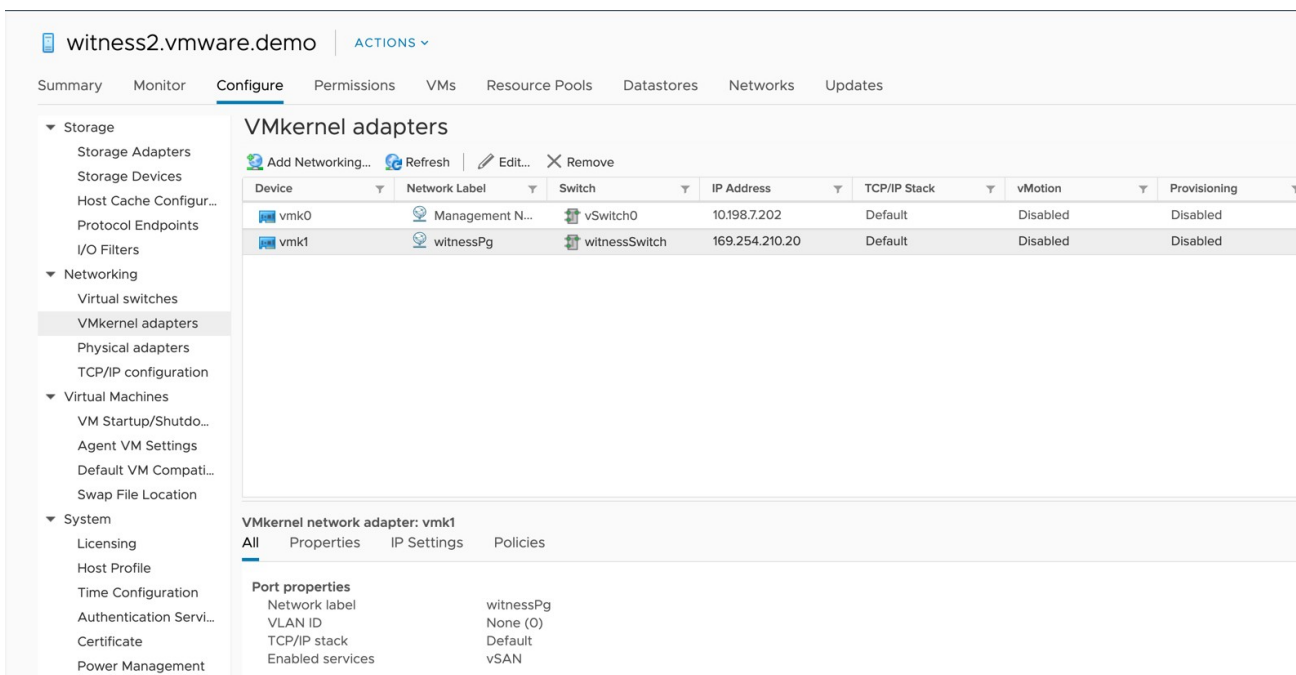
Once confirmed, you can proceed to configuring the vSAN network for the vSAN Witness Appliance.

Setup Step 4: Config vSAN Witness Host Networking

The next step is to configure the vSAN network correctly on the vSAN Witness Appliance. When the Witness is selected, navigate to Configure > Networking > Virtual switches as shown below.



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 depicted below.



*** Engineering note:** A few things to consider when configuring vSAN traffic on the vSAN Witness Appliance.

The default configuration has vmk0 configured for Management Traffic and vmk1 configured for vSAN Traffic. The vmk1 interface cannot be configured with an IP address on the same range as vmk0. This is because

Management traffic and vSAN traffic use the default TCP/IP stack. If both vmk0 and vmk1 are configured on the same range, a multihoming condition will occur and vSAN traffic will flow from vmk0 rather than vmk1. Health Check reporting will fail because vmk0 does not have vSAN enabled. The multihoming issue is detailed in KB 2010877 <https://kb.vmware.com/kb/2010877>.

Configure the network address.

Select the witnessPg and edit the properties by selecting the pencil icon.

The screenshot shows the vSphere Client interface for the host 'witness2.vmware.demo'. The 'Configure' tab is active, and the 'VMkernel adapters' section is expanded. A table lists the VMkernel adapters:

| Device | Network Label | Switch | IP Address | TCP/IP Stack | vMotion | Provisioning |
|--------|-----------------|---------------|----------------|--------------|----------|--------------|
| vmk0 | Management N... | vSwitch0 | 10.198.7.202 | Default | Disabled | Disabled |
| vmk1 | witnessPg | witnessSwitch | 169.254.210.20 | Default | Disabled | Disabled |

If vSAN is not an enabled service, select the witnessPg port group (vmk1), and then select the option to edit it. Tag the VMkernel port for vSAN traffic, as shown below:

The screenshot shows the 'vmk1 - Edit Settings' dialog box. The 'Port properties' tab is selected. The 'VMkernel port settings' section shows 'TCP/IP stack' set to 'Default' and 'MTU' set to '1500'. The 'Available services' section lists several services, with 'vSAN' checked and highlighted by a red box.

Ensure the MTU is set to an appropriate value. This value can be different when using Witness Traffic Separation, as long as it matches the Witness Tagged VMkernel interface MTU value.

vmk1 - Edit Settings

Port properties

IPv4 settings

IPv6 settings

VMkernel port settings

TCP/IP stack Default

MTU 1500

Available services

Enabled services

- vMotion
- Provisioning
- Fault Tolerance logging
- Management
- vSphere Replication
- vSphere Replication NFC
- vSAN

CANCEL OK

In the IPV4 settings, a default IP address has been allocated. Modify it for the vSAN traffic network.

vmk1 - Edit Settings

Port properties

IPv4 settings

IPv6 settings

No IPv4 settings

Obtain IPv4 settings automatically

Use static IPv4 settings

IPv4 address 101.98.5.60

Subnet mask 255.255.255.0

Default gateway Override default gateway for this adapter

10.198.7.253

DNS server addresses 10.198.6.8
10.198.16.1

CANCEL OK

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.

Setup Step 5: Validate Networking

The final step before a vSAN Stretched Cluster can be configured, is 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.

Default Gateways and Static Routes

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 TCP/IP stack to use a different path to reach a particular network. Now we can tell the TCP/IP 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.

Note once again that the vSAN network is a stretched L2 broadcast domain between the data sites as per VMware recommendations, but L3 is required to reach the vSAN network of the witness appliance. Therefore, 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.

```

esxi01-sitea.rainpole.com
-----
default      0.0.0.0      172.40.0.1   vmk0        MANUAL
172.3.0.0    255.255.255.0  0.0.0.0     vmk1        MANUAL
172.40.0.0   255.255.255.0  0.0.0.0     vmk0        MANUAL
[root@esxi01-sitea:~] esxcfg-route -n
Neighbor      MAC Address   Interface    Expiry      Type
172.40.0.1    f8:b1:56:6a:86:54  vmk0        1m54s      Unknown
172.3.0.14    00:50:56:6f:87:ca  vmk1        9m40s      Unknown
172.3.0.13    00:50:56:61:62:3b  vmk1        9m16s      Unknown
172.3.0.12    00:50:56:61:ad:ac  vmk1        40s        Unknown
[root@esxi01-sitea:~] vmkping -I vmk1 147.80.0.15
PING 147.80.0.15 (147.80.0.15): 56 data bytes

--- 147.80.0.15 ping statistics ---
3 packets transmitted, 0 packets received, 100% packet loss
[root@esxi01-sitea:~] esxcli network ip route ipv4 add -n 147.80.0.0/24 -g 172.3.0.1
[root@esxi01-sitea:~] esxcli network ip route ipv4 list
Network      Netmask      Gateway      Interface    Source
-----
default      0.0.0.0      172.40.0.1   vmk0        MANUAL
147.80.0.0    255.255.255.0  172.3.0.1   vmk1        MANUAL
172.3.0.0    255.255.255.0  0.0.0.0     vmk1        MANUAL
172.40.0.0    255.255.255.0  0.0.0.0     vmk0        MANUAL
[root@esxi01-sitea:~]

```

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**.

```

-----
default      0.0.0.0      192.60.0.1   vmk0        MANUAL
172.3.0.0    255.255.255.0 0.0.0.0     vmk1        MANUAL
192.60.0.0   255.255.255.0 0.0.0.0     vmk0        MANUAL
[root@esxi02-siteb:~] esxcfg-route -n
Neighbor      MAC Address  Interface    Expiry      Type
192.60.0.1    f8:b1:56:6a:86:55 vmk0         16m23s     Unknown
172.3.0.13    00:50:56:61:62:3b vmk1         47s        Unknown
172.3.0.12    00:50:56:61:ad:ac vmk1         13m25s     Unknown
172.3.0.11    8c:60:4f:36:bc:3c vmk1         14m13s     Unknown
[root@esxi02-siteb:~] vmkping -I vmk1 147.80.0.15
PING 147.80.0.15 (147.80.0.15): 56 data bytes

--- 147.80.0.15 ping statistics ---
3 packets transmitted, 0 packets received, 100% packet loss
[root@esxi02-siteb:~] esxcli network ip route ipv4 add -n 147.80.0.0/24 -g 172.3.0.253
[root@esxi02-siteb:~] esxcli network ip route ipv4 list
Network      Netmask      Gateway      Interface    Source
-----
default      0.0.0.0      192.60.0.1   vmk0        MANUAL
147.80.0.0    255.255.255.0 172.3.0.1   vmk1        MANUAL
172.3.0.0    255.255.255.0 0.0.0.0     vmk1        MANUAL
192.60.0.0    255.255.255.0 0.0.0.0     vmk0        MANUAL
[root@esxi02-siteb:~]

```

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.

```
witness-01.rainpole.com
[root@witness-01:~] esxcli network ip route ipv4 list
Network      Netmask      Gateway      Interface    Source
-----
default      0.0.0.0      147.70.0.1   vmk0         MANUAL
147.70.0.0   255.255.255.0 0.0.0.0     vmk0         MANUAL
147.80.0.0   255.255.255.0 0.0.0.0     vmk1         MANUAL
[root@witness-01:~] esxcli network ip route ipv4 list
Neighbor      MAC Address   Interface    Expiry      Type
-----
147.70.0.1    f8:b1:56:6a:86:2a vmk0         15m03s      Unknown
[root@witness-01:~] vmkping -I vmk1 172.3.0.11
PING 172.3.0.11 (172.3.0.11): 56 data bytes

--- 172.3.0.11 ping statistics ---
3 packets transmitted, 0 packets received, 100% packet loss
[root@witness-01:~] esxcli network ip route ipv4 add -n 172.3.0.11/32 -g 147.80.0.1
[root@witness-01:~] esxcli network ip route ipv4 add -n 172.3.0.12/32 -g 147.80.0.1
[root@witness-01:~] esxcli network ip route ipv4 list
Network      Netmask      Gateway      Interface    Source
-----
default      0.0.0.0      147.70.0.1   vmk0         MANUAL
147.70.0.0   255.255.255.0 0.0.0.0     vmk0         MANUAL
147.80.0.0   255.255.255.0 0.0.0.0     vmk1         MANUAL
172.3.0.11   255.255.255.0 147.80.0.1   vmk1         MANUAL
172.3.0.12   255.255.255.0 147.80.0.1   vmk1         MANUAL
[root@witness-01:~] vmkping -I vmk1 172.3.0.13
PING 172.3.0.13 (172.3.0.13): 56 data bytes

--- 172.3.0.13 ping statistics ---
3 packets transmitted, 0 packets received, 100% packet loss
[root@witness-01:~] esxcli network ip route ipv4 add -n 172.3.0.13/32 -g 147.80.0.253
[root@witness-01:~] esxcli network ip route ipv4 add -n 172.3.0.14/32 -g 147.80.0.253
[root@witness-01:~] esxcli network ip route ipv4 list
Network      Netmask      Gateway      Interface    Source
-----
default      0.0.0.0      147.70.0.1   vmk0         MANUAL
147.70.0.0   255.255.255.0 0.0.0.0     vmk0         MANUAL
147.80.0.0   255.255.255.0 0.0.0.0     vmk1         MANUAL
172.3.0.11   255.255.255.0 147.80.0.1   vmk1         MANUAL
172.3.0.12   255.255.255.0 147.80.0.1   vmk1         MANUAL
172.3.0.13   255.255.255.0 147.80.0.253 vmk1         MANUAL
172.3.0.14   255.255.255.0 147.80.0.253 vmk1         MANUAL
[root@witness-01:~]
```

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
```

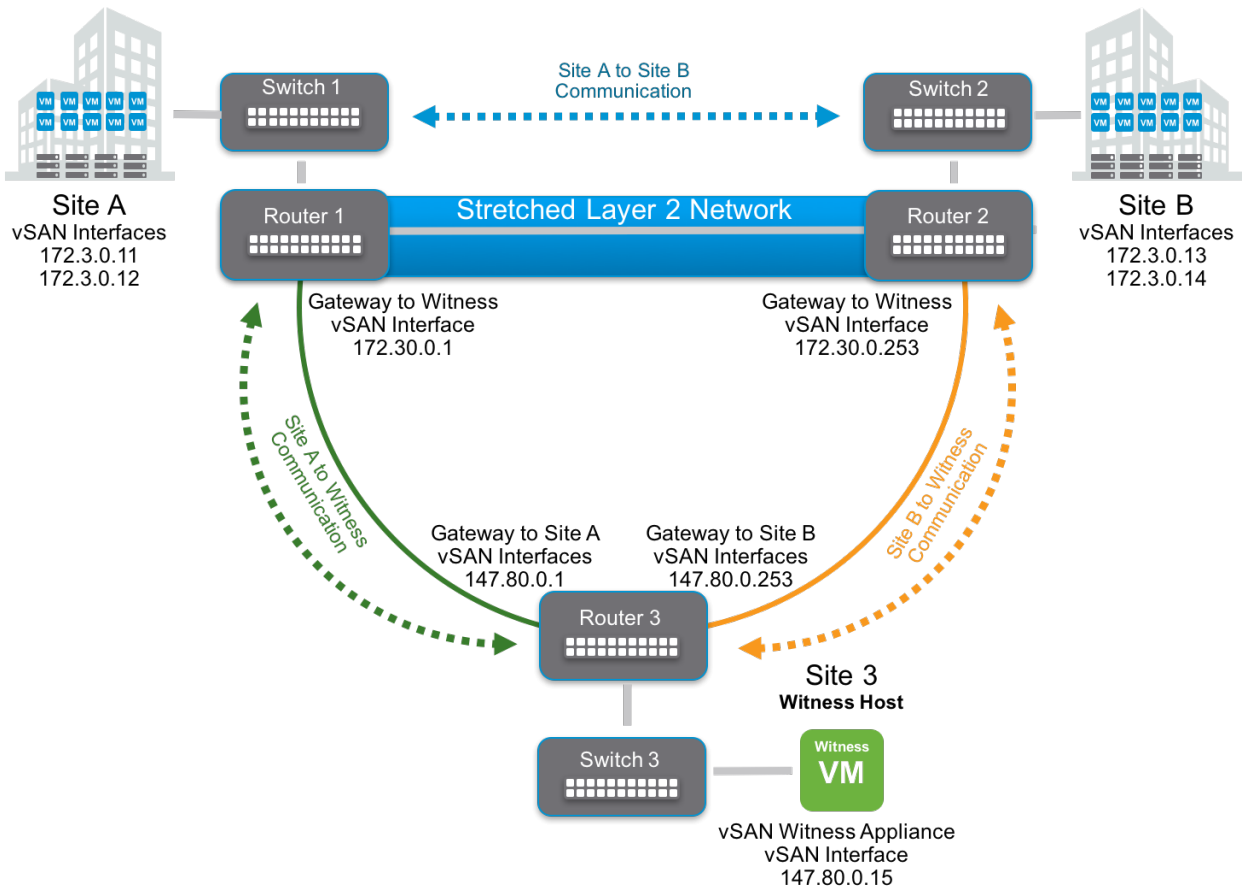
To do this individually for each host in Site B, the commands would be:

```
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

The following illustration shows the data flow between each of the data sites and the vSAN Witness Appliance in Site 3.



| Host | VMkernel | IP | Site | Static Route to Witness | Static Route to Site A | Static Route to Site B | Fault Domain |
|------|----------|-------------|------|-------------------------|------------------------|------------------------|-------------------------|
| | vmk1 | 172.3.0.11 | A | 172.3.0.1 | NA | NA | Preferred |
| | vmk1 | 172.3.0.12 | A | 172.3.0.1 | NA | NA | Preferred |
| | vmk1 | 172.3.0.13 | B | 172.3.0.253 | NA | NA | Secondary |
| | vmk1 | 172.3.0.14 | B | 172.3.0.253 | NA | NA | Secondary |
| | vmk1 | 147.80.0.15 | 3 | NA | 147.80.0.1 | 147.80.0.253 | witness-01.rainpole.com |

Next Steps

Once routing is in place, ping from each host's appropriate interface using `vmkping -I vmkX <destination IP>` to check connectivity. When connectivity has been verified between the vSAN data node VMkernel interfaces and the vSAN Witness Appliance vSAN VMkernel interface, the Stretched Cluster can be configured.

Configuring a vSAN Stretched Cluster

There are several methods to configure a vSAN Stretched Cluster.

A new cluster can be stretched or an existing cluster can be converted to a Stretched Cluster.

Creating a New vSAN Stretched Cluster

Creating a vSAN stretched cluster from a group of hosts that do not already have vSAN configured is relatively simple. A new vSAN cluster wizard makes the process very easy.

Create a Cluster using the Cluster QuickStart

The following steps should be followed to install a new vSAN Stretched Cluster. This example is a 3+3+1 deployment, meaning three ESXi hosts at Site A, three ESXi hosts at the Site B and 1 vSAN Witness Host.

In 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.

To set up vSAN and configure stretch cluster navigate to the Manage > vSAN > Cluster QuickStart to begin the vSAN wizard.

Cluster basics

Name the cluster, select the vSphere DRS, vSphere HA, and vSAN Services.

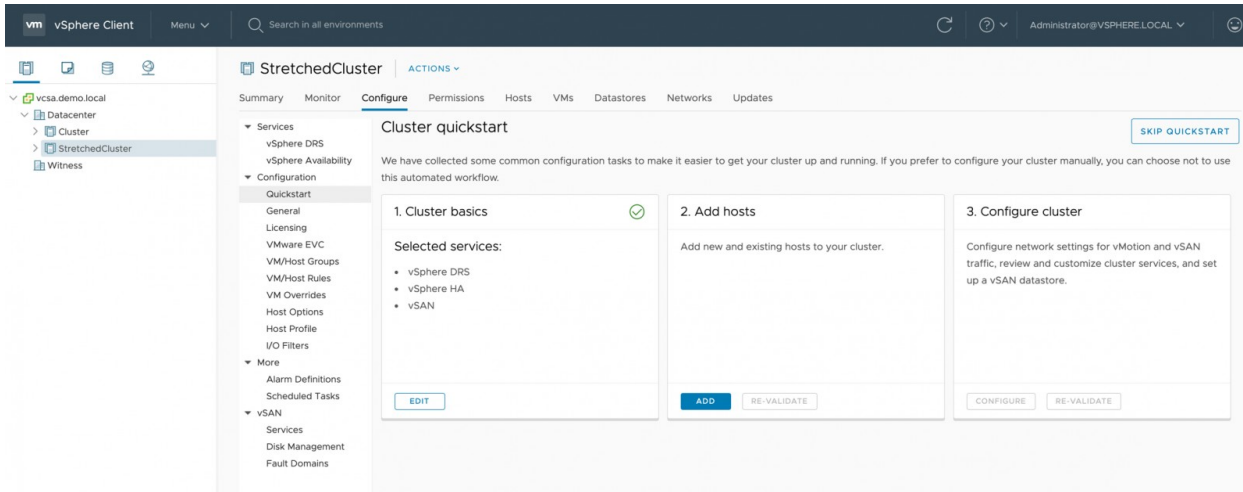
New Cluster
Datacenter
✕

| | |
|-------------|-------------------------------------|
| Name | StretchedCluster |
| Location | Datacenter |
| vSphere DRS | <input checked="" type="checkbox"/> |
| vSphere HA | <input checked="" type="checkbox"/> |
| vSAN | <input checked="" type="checkbox"/> |

These services will have default settings - these can be changed later in the Cluster Quickstart workflow.

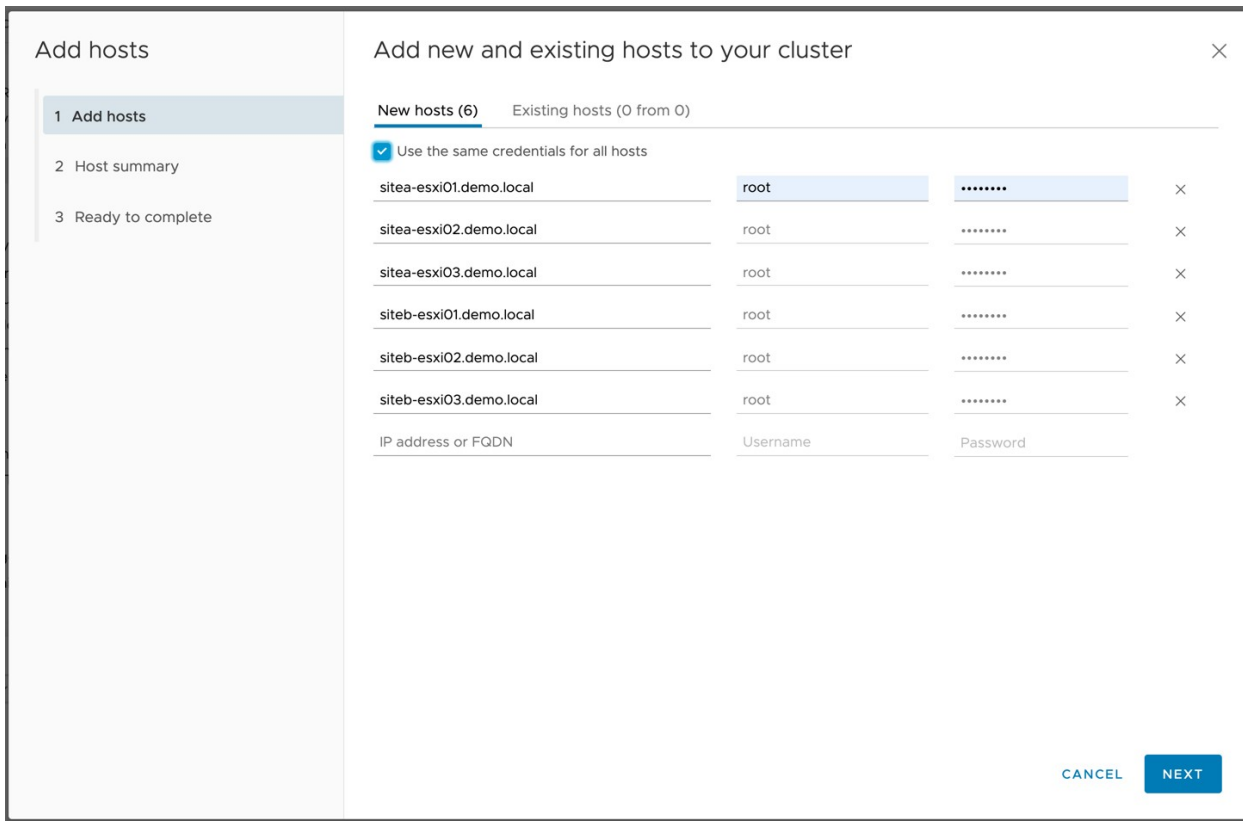
CANCEL
OK

1. Add Hosts



With the Cluster basics configured, hosts must be added to the cluster

An Add hosts dialog box adds new hosts or existing hosts in vCenter.



Security Alert

The certificates on 6 hosts could not be verified. The SHA1 thumbprints of the certificates are listed below. To continue connecting, manually verify these certificates and accept the thumbprints below.

| <input checked="" type="checkbox"/> | Hostname / IP Address | SHA1 Thumbprint |
|-------------------------------------|-------------------------|---|
| <input checked="" type="checkbox"/> | sitea-esxi01.demo.local | 00:53:35:4F:13:B1:DB:AF:93:53:A5:6C:B3:6D:8F:00:52:11:CF:60 |
| <input checked="" type="checkbox"/> | sitea-esxi02.demo.local | 23:19:54:67:F3:91:B7:2F:C5:9B:0F:84:15:02:E0:56:03:BE:FD:D2 |
| <input checked="" type="checkbox"/> | sitea-esxi03.demo.local | 98:AA:2F:B2:AF:E4:9D:BE:EF:B3:3B:8D:0B:F9:1A:BF:B5:78:2E:10 |
| <input checked="" type="checkbox"/> | siteb-esxi01.demo.local | 6E:23:AC:54:8C:28:1D:A2:0D:84:0E:EE:B0:CC:1A:CD:4B:8B:1B:88 |
| <input checked="" type="checkbox"/> | siteb-esxi02.demo.local | 19:98:CD:2F:3E:FD:07:8C:BD:95:83:05:6E:C0:42:40:01:06:1E:CD |
| <input checked="" type="checkbox"/> | siteb-esxi03.demo.local | C4:D7:FE:6A:4D:34:F6:83:62:B0:94:1C:8C:AF:86:DD:03:A0:8B:EA |

6

If hosts have not been added to vCenter, the certificates must be verified.

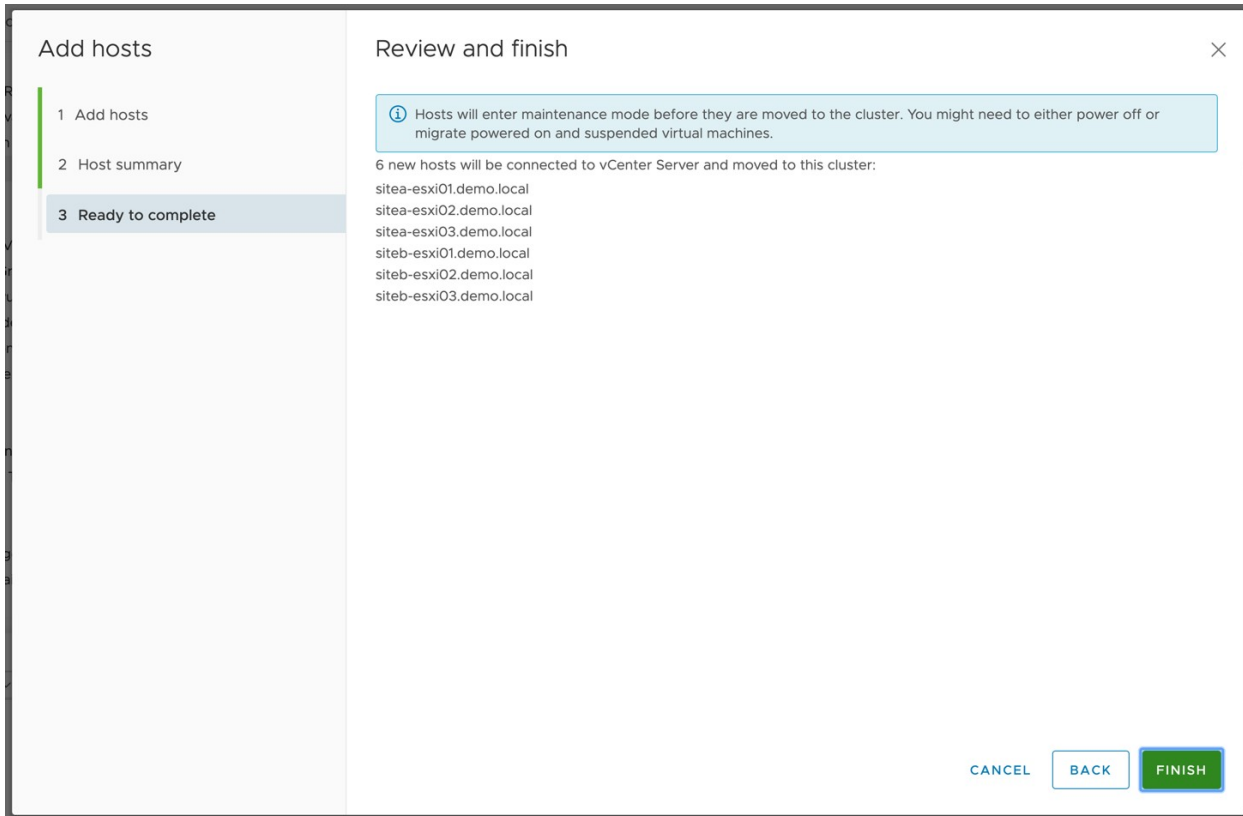
Add hosts

- Add hosts
- Host summary**
- Ready to complete

Host summary

| Hostname / IP Address | ESX Version | Model |
|---------------------------|-------------|------------------------|
| > sitea-esxi01.demo.local | 6.7.0 | VMware, Inc. VMware7,1 |
| > sitea-esxi02.demo.local | 6.7.0 | VMware, Inc. VMware7,1 |
| > sitea-esxi03.demo.local | 6.7.0 | VMware, Inc. VMware7,1 |
| > siteb-esxi01.demo.local | 6.7.0 | VMware, Inc. VMware7,1 |
| > siteb-esxi02.demo.local | 6.7.0 | VMware, Inc. VMware7,1 |
| > siteb-esxi03.demo.local | 6.7.0 | VMware, Inc. VMware7,1 |

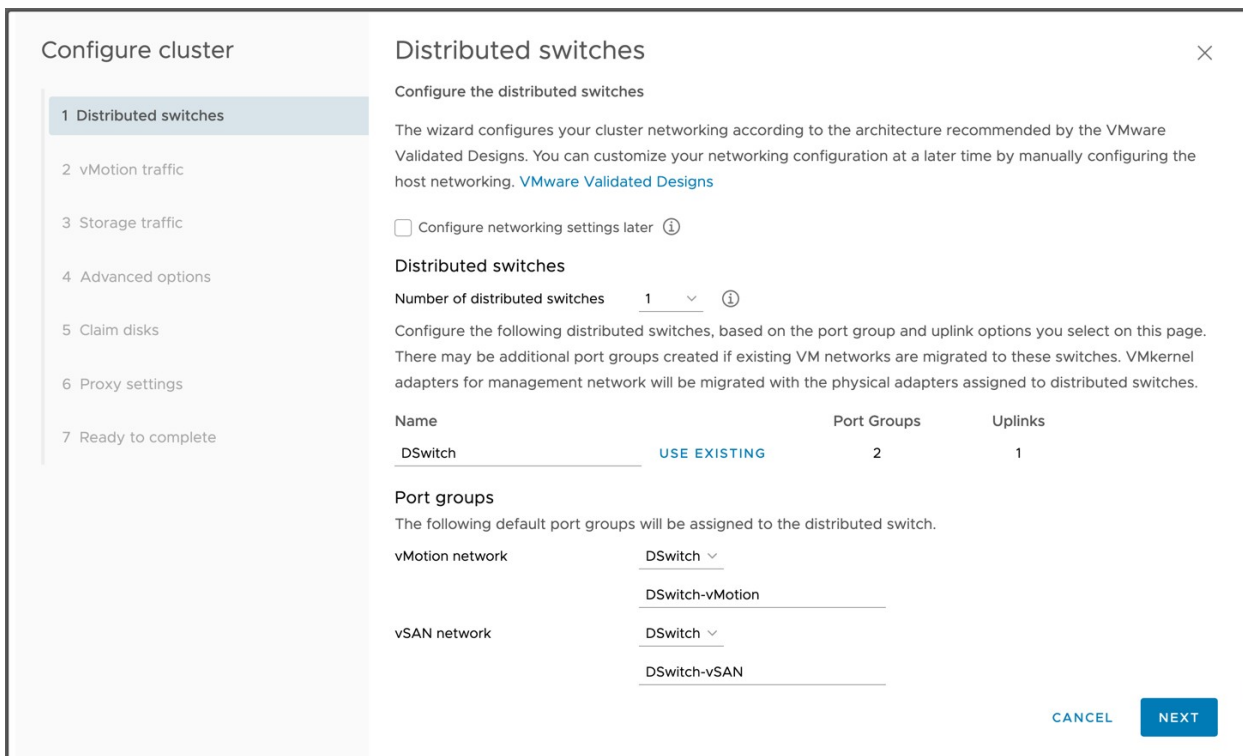
A summary of added hosts is displayed.



2. Configure Cluster

While services have been configured and hosts have been added, the hosts have not been configured for use by the cluster.

Selecting Configure will launch the Configure cluster wizard. 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 Switch(es) 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.



vMotion traffic is configured next.

Network-specific information such as VLAN tagging, protocol, and network addresses for the vMotion network. Note: Use the AUTOFILL option to automatically populate network addressing sequentially.

Configure cluster
vMotion traffic ✕

- 1 Distributed switches
- 2 vMotion traffic
- 3 Storage traffic
- 4 Advanced options
- 5 Claim disks
- 6 Proxy settings
- 7 Ready to complete

Specify the IP addresses for the vMotion traffic

Distributed switch: DSwitch

Distributed port group name: DSwitch-vMotion

Use VLAN: 30

Protocol: IPv4

IPv4 Configuration

IP type: Static IPs

Each host is configured automatically based on the input below. Empty default gateway might result in a segmented network.

| | | | | |
|--------------------------|-----------------|---------------|-----------------|----------|
| sitea-esxi01.demo.local | 192.168.150.101 | 255.255.255.0 | Default gateway | AUTOFILL |
| sitea-esxi02.demo.loc... | 192.168.150.102 | 255.255.255.0 | Default gateway | |
| sitea-esxi03.demo.loc... | 192.168.150.103 | 255.255.255.0 | Default gateway | |
| siteb-esxi01.demo.local | 192.168.150.104 | 255.255.255.0 | Default gateway | |
| siteb-esxi02.demo.lo... | 192.168.150.105 | 255.255.255.0 | Default gateway | |
| siteb-esxi03.demo.lo... | 192.168.150.106 | 255.255.255.0 | Default gateway | |

CANCEL BACK NEXT

vSAN traffic is configured next.

Network-specific information such as VLAN tagging, protocol, and network addresses for the vSAN network. Note: Use the AUTOFILL option to populate network addressing sequentially automatically.

Configure cluster
Storage traffic ✕

- 1 Distributed switches
- 2 vMotion traffic
- 3 Storage traffic
- 4 Advanced options
- 5 Claim disks
- 6 Proxy settings
- 7 Ready to complete

Specify the IP addresses for the vSAN traffic

Distributed switch: DSwitch

Distributed port group name: DSwitch-vSAN

Use VLAN: 40

Protocol: IPv4

IPv4 Configuration

IP type: Static IPs

Each host is configured automatically based on the input below.

| | | | | |
|--------------------------|-----------------|---------------|--|----------|
| sitea-esxi01.demo.local | 192.168.151.101 | 255.255.255.0 | | AUTOFILL |
| sitea-esxi02.demo.loc... | 192.168.151.102 | 255.255.255.0 | | |
| sitea-esxi03.demo.loc... | 192.168.151.103 | 255.255.255.0 | | |
| siteb-esxi01.demo.local | 192.168.151.104 | 255.255.255.0 | | |
| siteb-esxi02.demo.lo... | 192.168.151.105 | 255.255.255.0 | | |
| siteb-esxi03.demo.lo... | 192.168.151.106 | 255.255.255.0 | | |

CANCEL BACK NEXT

Advanced options are where the basics for HA, DRS, vSAN, NTP, EVC, and more can be set.

Note: Advanced HA & DRS settings must be configured afterward for a proper configuration.

Be certain to set:

- HA with
 - Host Failure Monitoring
 - Admission Control
 - Host failures to tolerate to the number of hosts in a site
- DRS
 - Fully Automated
- vSAN Deployment Type
 - Stretched cluster

Configure cluster

- 1 Distributed switches
- 2 vMotion traffic
- 3 Storage traffic
- 4 Advanced options
- 5 Claim disks
- 6 Proxy settings
- 7 Configure fault domains
- 8 Select witness host
- 9 Claim disks for witness host
- 10 Ready to complete

Advanced options ×

Customize the cluster settings.

▼ vSphere HA

| | |
|----------------------------|--|
| Host Failure Monitoring | <input checked="" type="checkbox"/> |
| Virtual Machine Monitoring | <input type="checkbox"/> |
| Admission Control | <input checked="" type="checkbox"/> |
| Host failure to tolerate | 3 ▼ <small>The maximum is one less than number of hosts in cluster.</small> |

▼ vSphere DRS

| | |
|---------------------|-----------------------|
| Automation level | Partially Automated ▼ |
| Migration threshold | 3 ▼ ⓘ |

▼ vSAN Options

| | |
|-------------------------------|--|
| Deployment type | Stretched cluster ▼ |
| Encryption | <input type="checkbox"/> |
| KMS Cluster | AKM ▼ |
| Deduplication and compression | <input type="checkbox"/> <small>Requires all-flash configuration</small> |
| Fault domains | <input type="checkbox"/> |

CANCEL
BACK
NEXT

Disks can then be claimed for vSAN.

vmware®
by Broadcom

© VMware LLC.

Document | 85

Configure cluster

- 1 Distributed switches
- 2 vMotion traffic
- 3 Storage traffic
- 4 Advanced options
- 5 Claim disks
- 6 Proxy settings
- 7 Configure fault domains
- 8 Select witness host
- 9 Claim disks for witness host
- 10 Ready to complete

Claim disks

Select disks to contribute to the vSAN datastore.

Claim disks on hosts for cache and capacity.
Non-empty disks will be deleted.

Claimed capacity 2.34 TB
Claimed cache 300.00 GB
Unclaimed storage 0.00 B

CLAIM DISK FOR DISK DRIVE TYPE Group by: Disk model/size

| Disk Model/Serial Number | Claim For | Drive Type | Disk Distribution/Host |
|-------------------------------|--------------|------------|-------------------------|
| VMware Virtual disk, 12 x ... | Capacity tie | Flash | 2 disks on 6 hosts |
| Local VMware Disk (naa....) | Capacity tie | Flash | sitea-esxi02.demo.local |
| Local VMware Disk (naa....) | Capacity tie | Flash | sitea-esxi02.demo.local |
| Local VMware Disk (naa....) | Capacity tie | Flash | sitea-esxi01.demo.local |
| Local VMware Disk (naa....) | Capacity tie | Flash | sitea-esxi01.demo.local |
| Local VMware Disk (naa....) | Capacity tie | Flash | sitea-esxi03.demo.local |
| Local VMware Disk (naa....) | Capacity tie | Flash | sitea-esxi03.demo.local |
| Local VMware Disk (naa....) | Capacity tie | Flash | siteb-esxi02.demo.local |
| Local VMware Disk (naa....) | Capacity tie | Flash | siteb-esxi02.demo.local |
| Local VMware Disk (naa....) | Capacity tie | Flash | siteb-esxi01.demo.local |

20 items

✔ Configuration correct.

CANCEL
BACK
NEXT

Proxy settings can be configured to allow vCenter configurations that do not have direct access to the Internet to communicate with VMware for the Online Health check and anonymized telemetry data.

Configure cluster

- 1 Distributed switches
- 2 vMotion traffic
- 3 Storage traffic
- 4 Advanced options
- 5 Claim disks
- 6 Proxy settings
- 7 Configure fault domains
- 8 Select witness host
- 9 Claim disks for witness host
- 10 Ready to complete

Proxy settings

Configure proxy to establish connection to send data for CEIP. vSAN Support Insight requires to allow outbound traffic to <https://vcsa.vmware.com:443/ph/api/> and <http://www.vmware.com:80/>

Configure the Proxy server if your system uses one

Host name: *

Port: *

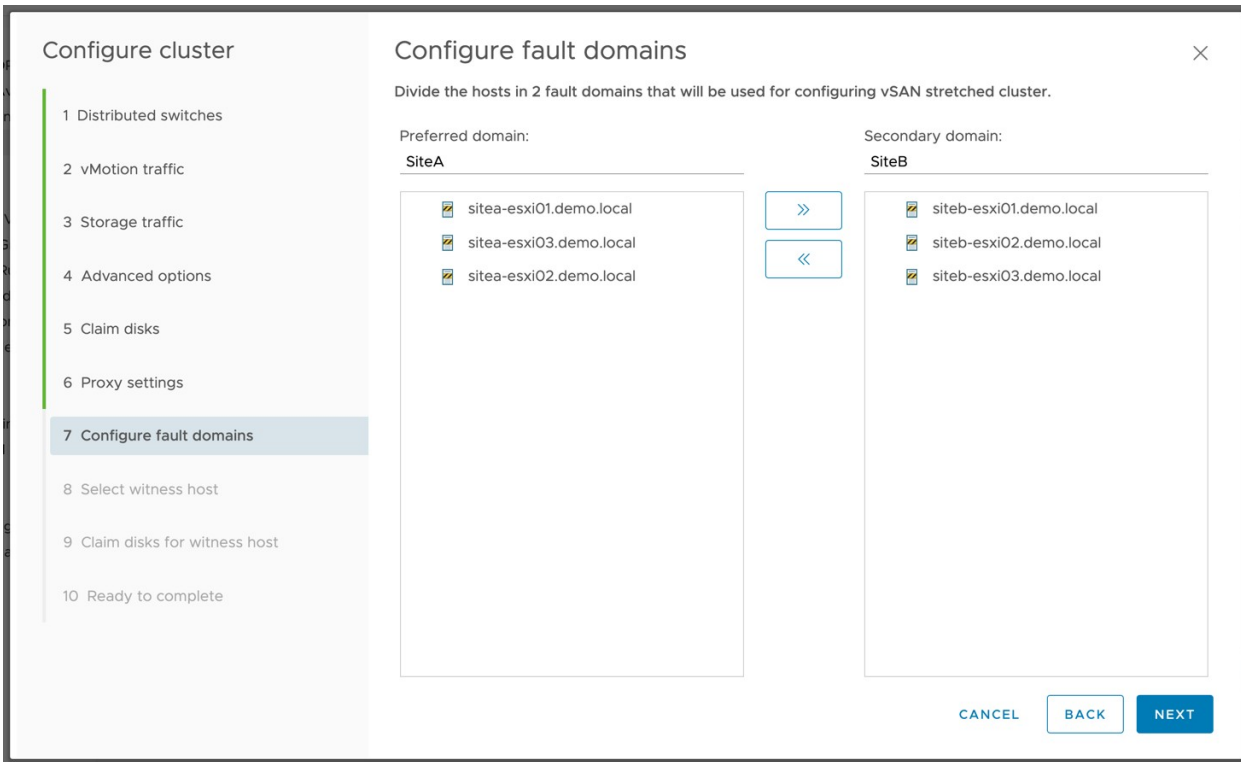
User name:

Password:

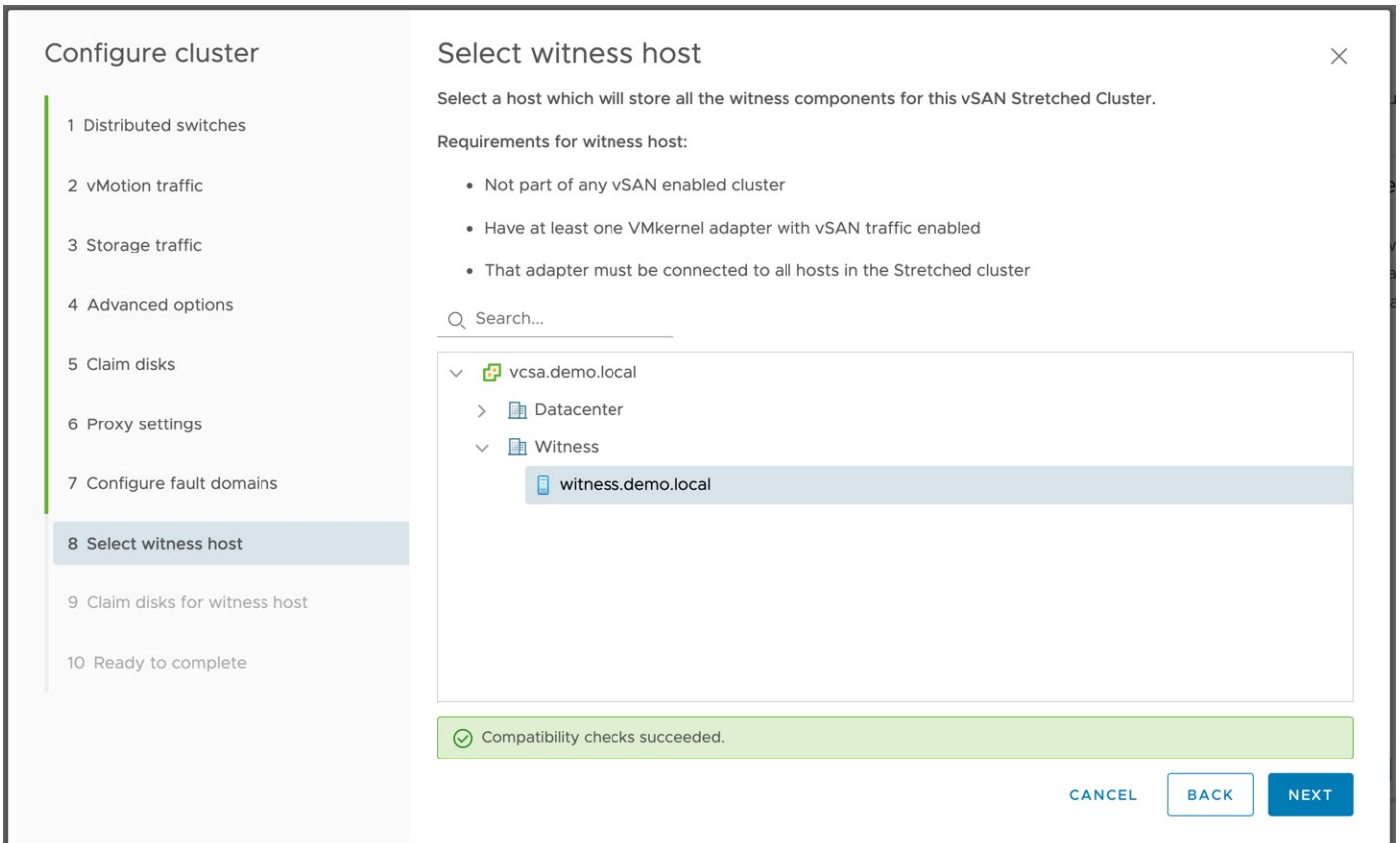
CANCEL
BACK
NEXT

Fault Domains are configured to determine which hosts will be in which site. The left site is the initial Preferred site, but this may

be swapped to the alternate site at a later time if desired.



A vSAN Witness Host is chosen to prevent split-brain scenarios when each data site is isolated from the other.



Just like a regular vSAN data node, the vSAN Witness Host must have a disk group created.

If using the vSAN Witness Appliance for a vSAN Witness Host, the Cache disk will always be the 10GB Flash Disk.

Configure cluster

- 1 Distributed switches
- 2 vMotion traffic
- 3 Storage traffic
- 4 Advanced options
- 5 Claim disks
- 6 Proxy settings
- 7 Configure fault domains
- 8 Select witness host
- 9 Claim disks for witness host**
- 10 Ready to complete

Claim disks for witness host

Select disks on the witness host to be used for storing witness components.

First, select a single disk to serve as cache tier.

| | Name | Drive Type | Capacity | Transport Type | Adapter |
|----------------------------------|-------------------------------|------------|-----------|----------------|---------|
| <input checked="" type="radio"/> | Local VMware Disk (mpx.vmh... | Flash | 10.00 GB | | |
| <input type="radio"/> | Local VMware Disk (mpx.vmh... | Flash | 350.00 GB | | |

Then, select one or more disks to serve as capacity tier.

Capacity type: Flash

| | Name | Drive Type | Capacity | Transport Type | Adapter |
|-------------------------------------|-------------------------------|------------|-----------|----------------|---------|
| <input checked="" type="checkbox"/> | Local VMware Disk (mpx.vmh... | Flash | 350.00 GB | | |

1
1 item

CANCEL
BACK
NEXT

After the vSAN Witness Host has been selected, the Wizard will complete.

Configure cluster

- 1 Distributed switches
- 2 vMotion traffic
- 3 Storage traffic
- 4 Advanced options
- 5 Claim disks
- 6 Proxy settings
- 7 Configure fault domains
- 8 Select witness host
- 9 Claim disks for witness host
- 10 Ready to complete**

Ready to complete

Physical networks
The cluster uses only one physical network

vMotion traffic
Configured static IPs for all 6 hosts in IPv4 on VLAN 30

Storage traffic
Configured static IPs for all 6 hosts in IPv4 on VLAN 40

Advanced options
The cluster is configured with the following options

- ✔ Lockdown mode is disabled on all hosts
- ✔ All hosts use NTP server: ntp.demo.local
- ✔ Enhanced vMotion Compatibility is disabled
- ✔ Host update preference: Include upgrades to new ESXi versions

vSAN Datastore
The cluster has a vSAN datastore configured out of the local disks on each of the 6 host(s)

| | |
|---------------|-----------------------|
| Claim disks | All flash disk groups |
| Cache size | 300.00 GB |
| Capacity size | 2.34 TB |

CANCEL
BACK
FINISH

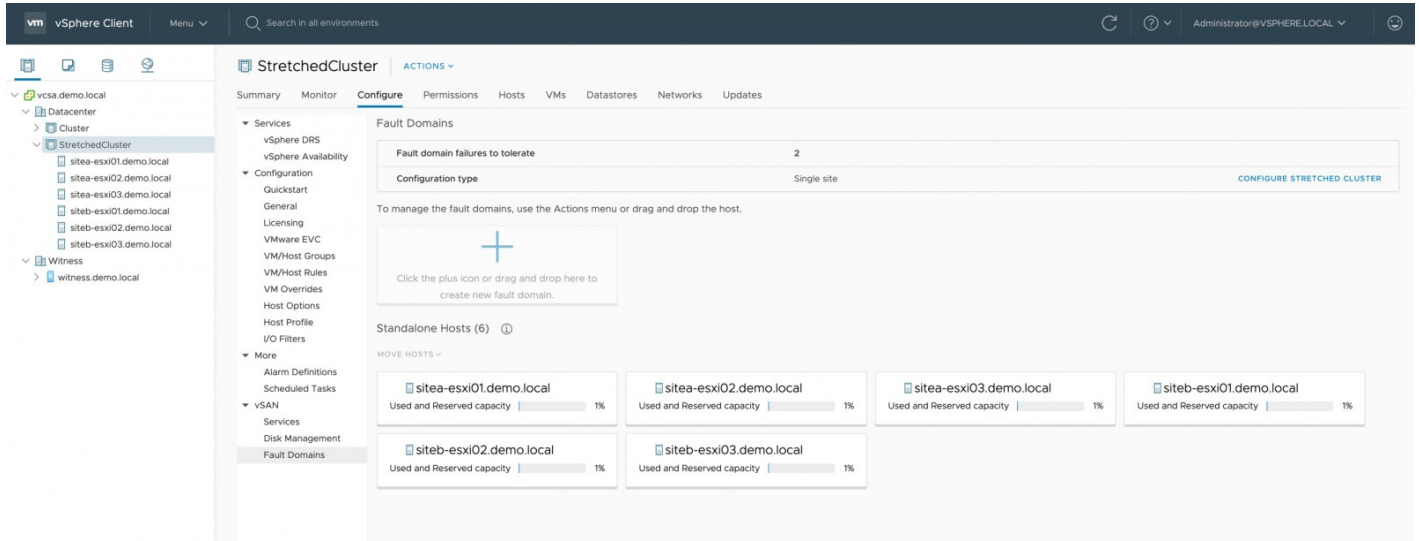
Converting a Cluster to a Stretched Cluster

The following steps should be followed to convert an existing vSAN cluster to a stretched cluster. This example is a 3+3+1 deployment, meaning three ESXi hosts at the Site A, three ESXi hosts at the Site B, and 1 vSAN Witness Host.

Consider that all hosts are properly configured and vSAN is already running.

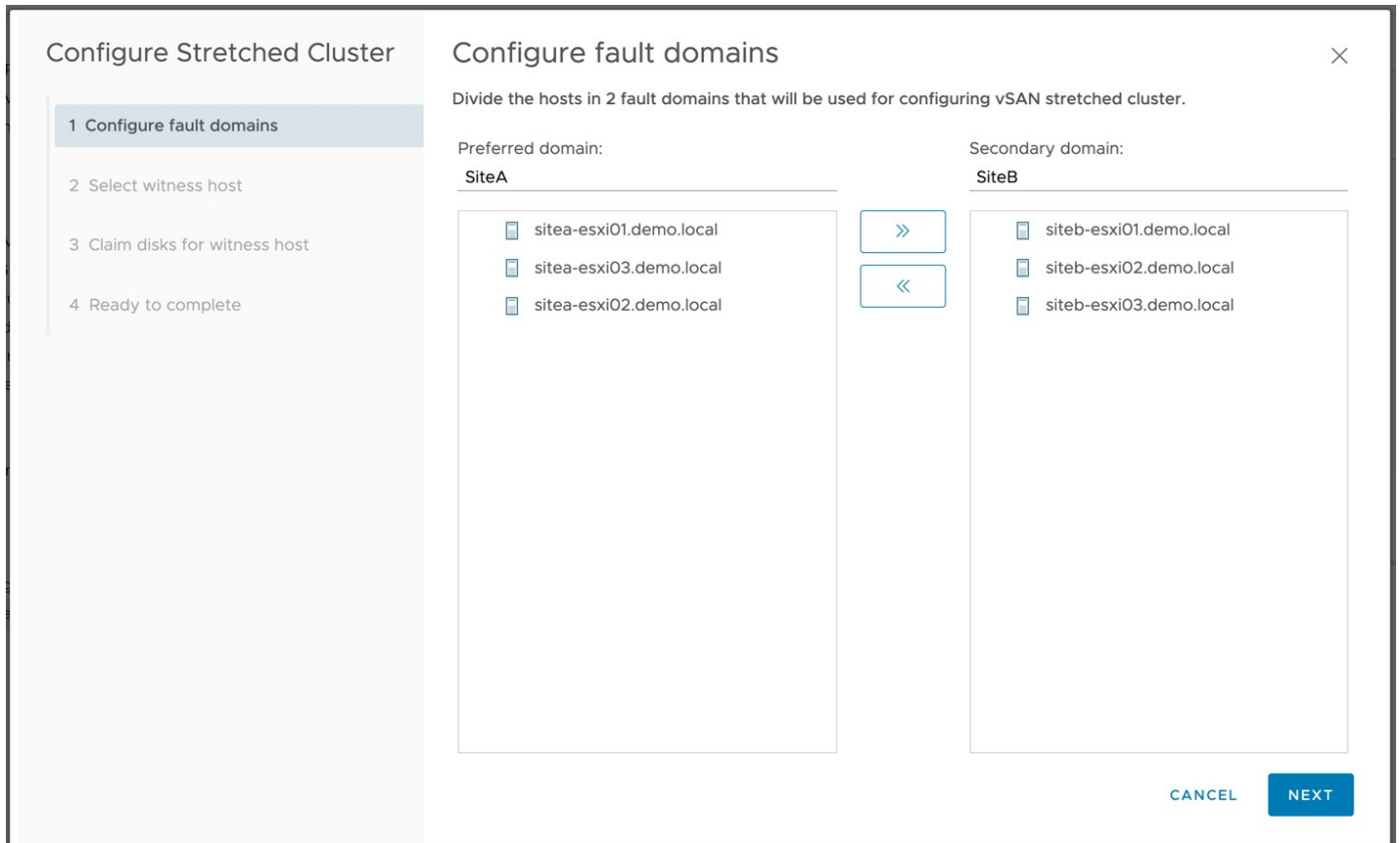
Fault Domains

Configuring the stretched cluster setting is handled through the Fault Domains menu item. Select Configure > Fault Domains



Configure the Fault Domains.

The leftmost Fault Domain is initially the Preferred Fault Domain. The alternate site may be designated as the Preferred later if desired.



Witness Host

A vSAN Witness Host is chosen to prevent split-brain scenarios when each data site is isolated.

Configure Stretched Cluster
Select witness host ✕

- 1 Configure fault domains
- 2 Select witness host
- 3 Claim disks for witness host
- 4 Ready to complete

Select a host which will store all the witness components for this vSAN Stretched Cluster.

Requirements for witness host:

- Not part of any vSAN enabled cluster
- Have at least one VMkernel adapter with vSAN traffic enabled
- That adapter must be connected to all hosts in the Stretched cluster

Search...

v

Datacenter

Witness

witness.demo.local

✔ Compatibility checks succeeded.

CANCEL
BACK
NEXT

Claim disks for witness host

Just like a regular vSAN data node, the vSAN Witness Host must have a disk group created.

If using the vSAN Witness Appliance for a vSAN Witness Host, the Cache disk will always be the 10GB Flash Disk.

Configure Stretched Cluster
Claim disks for witness host ✕

- 1 Configure fault domains
- 2 Select witness host
- 3 Claim disks for witness host
- 4 Ready to complete

Select disks on the witness host to be used for storing witness components.

First, select a single disk to serve as cache tier.

| | Name | Drive Type | Capacity | Transport Type | Adapter |
|----------------------------------|-------------------------------|------------|-----------|----------------|---------|
| <input checked="" type="radio"/> | Local VMware Disk (mpx.vmh... | Flash | 10.00 GB | | |
| <input type="radio"/> | Local VMware Disk (mpx.vmh... | Flash | 350.00 GB | | |

2 items

Then, select one or more disks to serve as capacity tier.

Capacity type: Flash

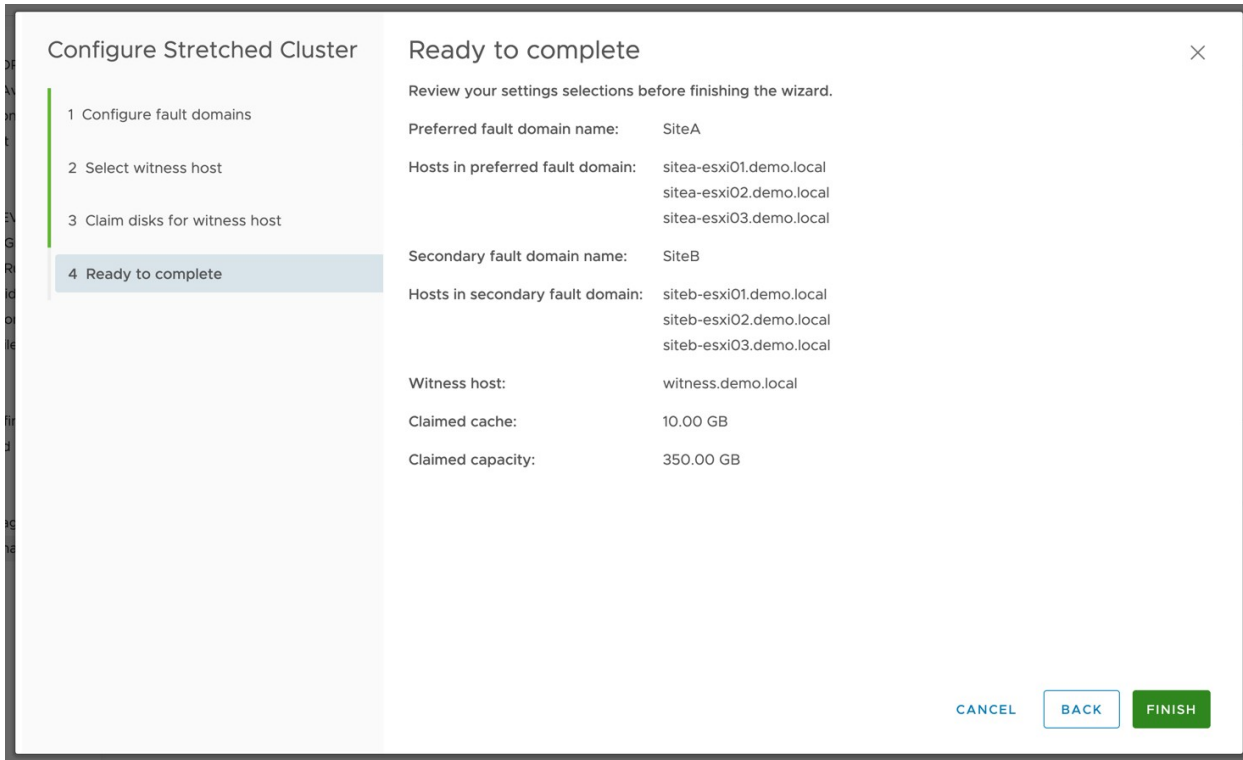
| | Name | Drive Type | Capacity | Transport Type | Adapter |
|-------------------------------------|-------------------------------|------------|-----------|----------------|---------|
| <input checked="" type="checkbox"/> | Local VMware Disk (mpx.vmh... | Flash | 350.00 GB | | |

1 item

CANCEL
BACK
NEXT

Finish

Review the vSAN Stretched Cluster configuration for accuracy and click Finish.



Set vSphere HA Advanced Settings

Select Configure > vSphere HA on the vSAN Stretched Cluster to configure the advanced settings for vSphere HA.

Failures and responses

This setting determines what happens to the virtual machines on an isolated host, i.e. a host that can no longer communicate to other nodes in the cluster, nor is able to reach the isolation response IP address.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses | Admission Control | Heartbeat Datastores | Advanced Options

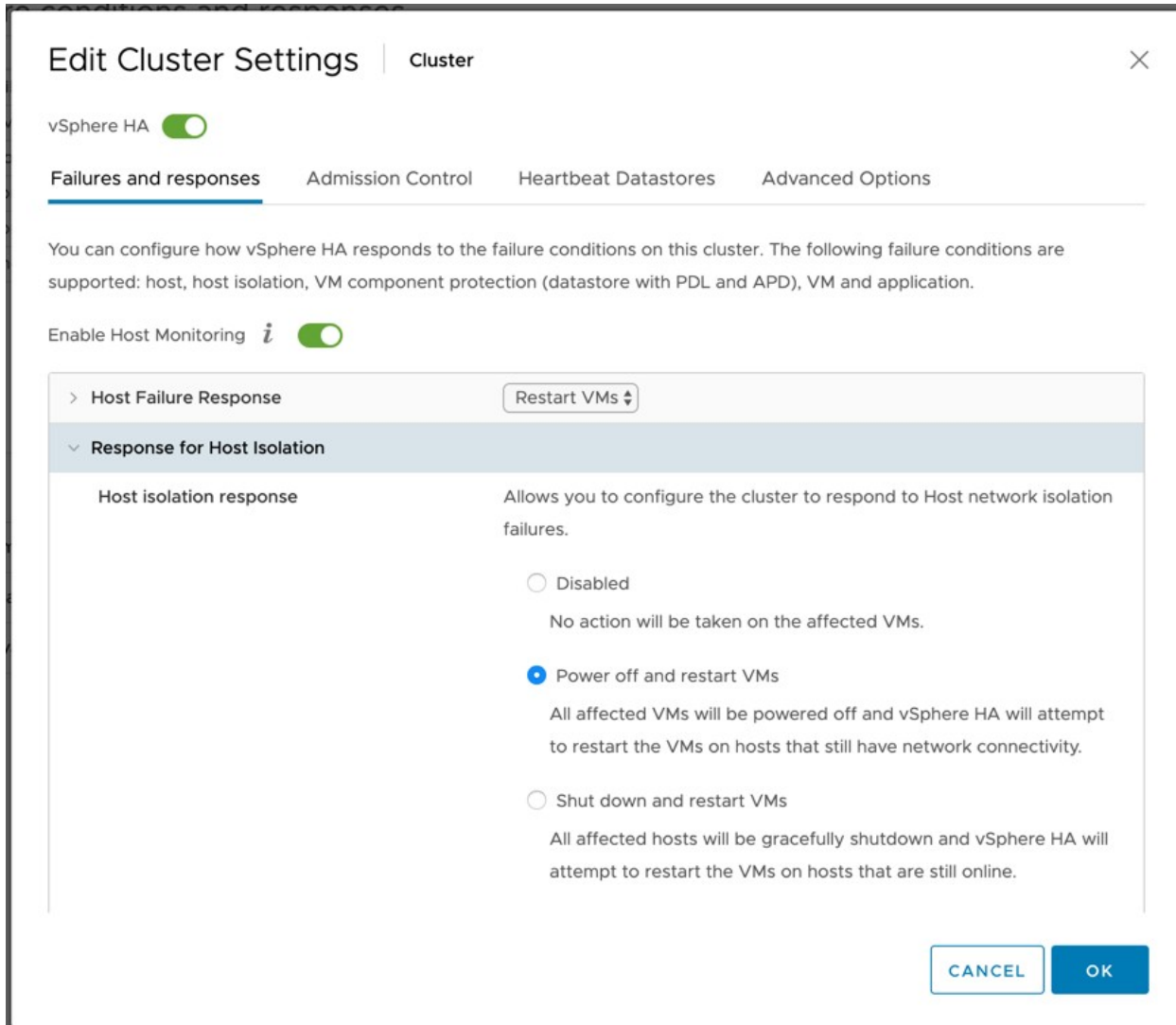
You can configure how vSphere HA responds to the failure conditions on this cluster. The following failure conditions are supported: host, host isolation, VM component protection (datastore with PDL and APD), VM and application.

Enable Host Monitoring *i*

| | |
|--------------------------------------|------------------------------------|
| > Host Failure Response | Restart VMs ▾ |
| > Response for Host Isolation | Power off and restart VMs ▾ |
| > Datastore with PDL | Disabled ▾ |
| > Datastore with APD | Disabled ▾ |
| > VM Monitoring | Disabled ▾ |

CANCEL OK

VMware recommends that the Response for Host Isolation is to Power off and restart VMs. This is because a clean shutdown will not be possible on an isolated host as the access to the vSAN Datastore and the ability to write to disk is lost.



Admission Control

Admission control ensures HA has sufficient resources to restart virtual machines after a failure. As a full 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 complete site failure. Since ESXi hosts will be equally divided across both sites in a vSAN Stretched Cluster, and to ensure that vSphere HA can restart all workloads, VMware recommends configuring the admission control policy to 50 percent for both memory and CPU.

VMware recommends using the percentage-based policy as it offers the most flexibility and reduces operational overhead. For more details about admission control policies and the associated algorithms, we would like to refer to the vSphere 6.5 Availability Guide.

The following screenshot shows a vSphere HA cluster configured with admission control enabled using the percentage-based admission control policy set to 50%.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses | **Admission Control** | Heartbeat Datastores | Advanced Options

Admission control is a policy used by vSphere HA to ensure failover capacity within a cluster. Raising the number of potential host failures will increase the availability constraints and capacity reserved.

Host failures cluster tolerates _____
Maximum is one less than number of hosts in cluster.

Define host failover capacity by: Cluster resource Percentage

Override calculated failover capacity.

Reserved failover CPU capacity: 50 % CPU

Reserved failover Memory capacity: 50 % Memory

Performance degradation VMS tolerate: 50 %

Percentage of performance degradation the VMs in the cluster are allowed to tolerate during a failure. 0% - Raises a warning if there is insufficient failover capacity to guarantee the same performance after VMs restart. 100% - Warning is disabled.

CANCEL OK

It should be noted that vSAN is not admission-control aware. There is no way to inform vSAN to set aside additional storage resources to accommodate fully compliant virtual machines running on a single site. This is an additional operational step for administrators if they wish to achieve such a configuration in the event of a failure.

Heartbeat Datastores

vSphere HA provides an additional heartbeating mechanism for determining the state of hosts in the cluster. This is in addition to network heartbeating, and is called datastore heartbeating. In many vSAN environment no additional datastores, outside of vSAN, are available, and as such in general VMware recommends disabling Heartbeat Datastores as the vSAN Datastore cannot be used for heartbeating. However, if additional datastores are available, then using heartbeat datastores is fully supported.

What do Heartbeat Datastores do, and when does it come in to play? The heartbeat datastore is used by a host which is isolated to inform the rest of the cluster what its state is and what the state of the VMs is. When a host is isolated, and the isolation response is configured to "power off" or "shutdown", then the heartbeat datastore will be used to inform the rest of the cluster when VMs are powered off (or shutdown) as a result of the isolation. This allows the vSphere HA primary node to immediately restart the impacted VMs.

To disable datastore heartbeating, under vSphere HA settings, open the Datastore for Heartbeating section. Select the option "Use datastore from only the specified list", and ensure that there are no datastore selected in the list, if any exist. Datastore heartbeats are now disabled on the cluster. Note that this may give rise to a notification in the summary tab of the host, stating that the number of vSphere HA heartbeat datastore for this host is 0, which is less than required:2. This message may be removed by following KB Article 2004739 which details how to add the advanced setting `das.ignoreInsufficientHbDatastore = true`.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses Admission Control **Heartbeat Datastores** Advanced Options

vSphere HA uses datastores to monitor hosts and virtual machines when the HA network has failed. vCenter Server selects 2 datastores for each host using the policy and datastore preferences specified below.

Heartbeat datastore selection policy:

- Automatically select datastores accessible from the hosts
- Use datastores only from the specified list
- Use datastores from the specified list and complement automatically if needed

Available heartbeat datastores

| Name | Datastore Cluster | Hosts Mounting Datastore ↓ |
|------|-------------------|----------------------------|
| | | |

CANCEL OK

Advanced Options

In a vSAN Stretched Cluster, one of the isolation addresses should reside in the site 1 data center, and the other should reside in the site 2 data center. This would enable vSphere HA to validate host isolation even in the case of a partitioned scenario (network failure between sites).

VMware recommends enabling host isolation response and specifying isolation response addresses on the vSAN network rather than the management network.

The vSphere HA advanced setting `das.usedefaultisolationaddress` should be set to false.

VMware recommends specifying two additional isolation response addresses, and each of these addresses should be site-specific. In other words, select an isolation response IP address from the Preferred Site and another isolation response IP address from the Non-Preferred Site.

The vSphere HA advanced setting used for setting the first isolation response IP address is `das.isolationaddress0` and it should be set to an IP address on the vSAN network which resides on the one site.

The vSphere HA advanced setting used for adding a second isolation response IP address is `das.isolationaddress1` and this should be an IP address on the vSAN network that resides on the alternate site.

Edit Cluster Settings | Cluster

vSphere HA

Failures and responses | Admission Control | Heartbeat Datastores | **Advanced Options**

You can set advanced options that affect the behavior of your vSphere HA cluster.

+ Add ✕ Delete

| Option | Value |
|--------------------------------|----------------|
| das.isolationaddress0 | 192.168.152.30 |
| das.isolationaddress1 | 192.168.153.30 |
| das.usedefaultisolationaddress | false |

3 items

CANCEL OK

For further details on how to configure this setting, information can be found in KB Article 1002117.

Configure Stretched Cluster Site Affinity Configuring Site Affinity.

This is achieved with VM Groups, Host Groups, and VM/Host Rules.

With these groups and rules, an administrator can specify which set of hosts (i.e., which site) a virtual machine is deployed to.

The first step is to create two host groups; the first host group will contain the ESXi hosts from the one site, and the second host group will contain the ESXi hosts in the alternate site.

In this setup example, a 3+3+1 environment is being deployed, so there are two hosts in each host group. Select the cluster object from the vSphere Inventory, select Manage, then Settings. This is where the VM/Host Groups are created.

Navigate to Cluster > Configure > VM/Host Groups. Select the option to add a group. Name the group, and ensure the group type is "Host Group" instead of "VM Group." Next, click Add to select the hosts that should be in the host group. Select the hosts from site A.

The screenshot shows the vSphere Web Client interface for a StretchedCluster. The 'Configure' tab is active, and the 'VM/Host Groups' section is selected. A dialog box titled 'Create VM/Host Group' is open, showing the following details:

- Name:** HostsSiteA
- Type:** Host Group
- Members:**
 - sitea-esxi01.demo.local
 - sitea-esxi02.demo.local
 - sitea-esxi03.demo.local

The dialog also includes '+ Add...' and '- Remove' buttons for managing the members list, and 'CANCEL' and 'OK' buttons at the bottom.

Once the hosts have been added to the Host Group, click OK. Review the settings of the host group, and click OK to create it:

This step will need to be repeated for the alternate site. Create a host group for the alternate site and add the ESXi hosts from the alternate site to the host group.

The screenshot shows the vSphere configuration interface for a StretchedCluster. The left sidebar contains a navigation menu with categories like Services, Configuration, More, and vSAN. The 'VM/Host Groups' option is selected under Configuration. The main panel displays the 'VM/Host Groups' configuration page, which includes a table of existing host groups and a section for adding members to a specific group.

VM/Host Groups

+ Add... - Delete

| Name | Type |
|------------|------------|
| HostsSiteA | Host Group |
| HostsSiteB | Host Group |

+ Add... - Remove

HostsSiteA Group Members

| |
|-------------------------|
| sitea-esxi03.demo.local |
| sitea-esxi01.demo.local |
| sitea-esxi02.demo.local |

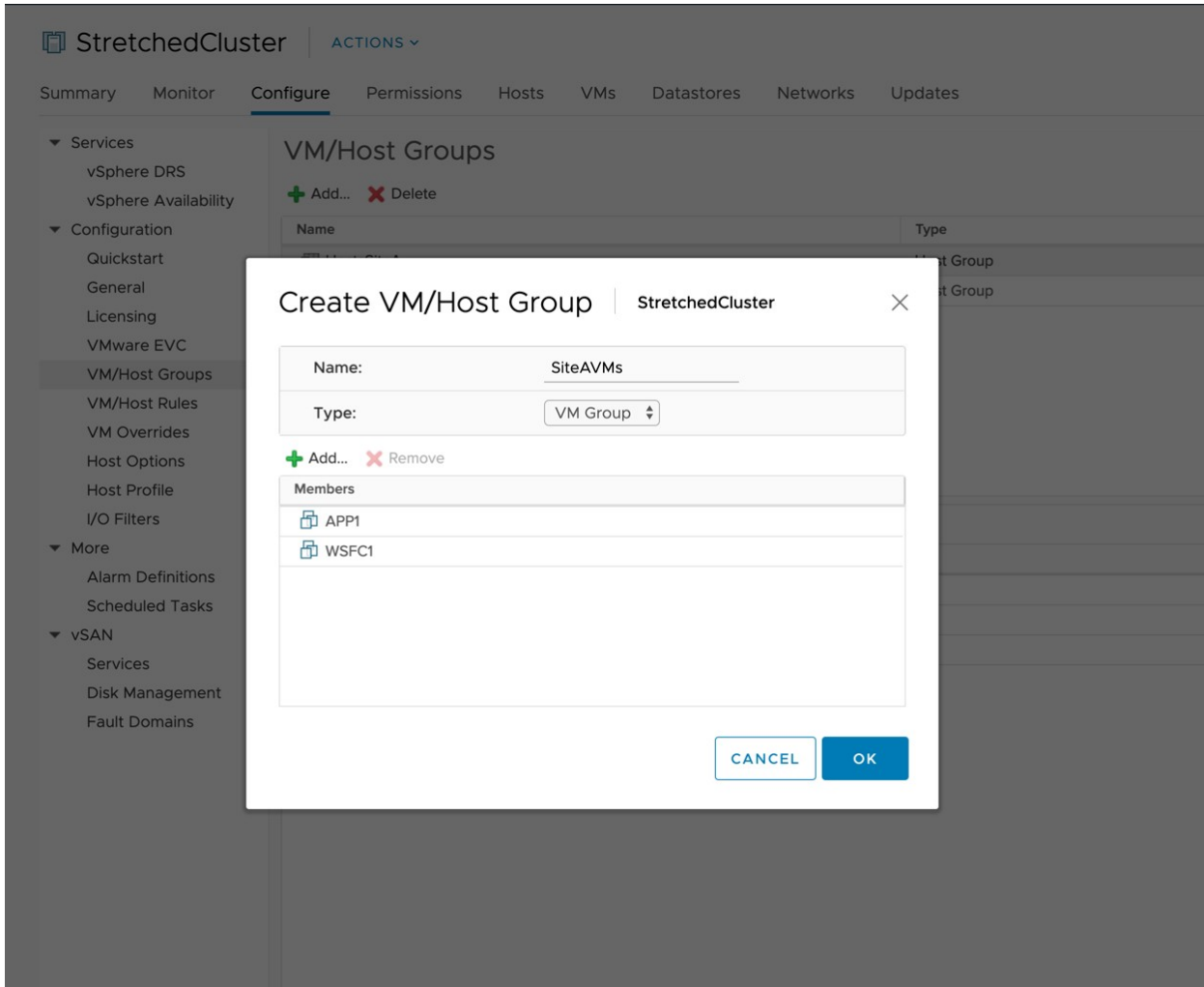
When host groups for both data sites have been created, the next step is to create VM groups. However, virtual machines should be created on the cluster before you can do this.

Create VM Groups

Once the host groups are created, the initial set of virtual machines should now be created.

Wait to power on the virtual machines. Once the virtual machines are in the inventory, you can proceed with creating the VM Groups.

Create the VM Groups for the Preferred site, Site A in this case. Select the virtual machines that should run in Site A.



Create a second VM Group for the virtual machines that should reside on the alternate site.

The screenshot shows the vSphere configuration interface for a Stretched Cluster. The main navigation bar includes 'Summary', 'Monitor', 'Configure', 'Permissions', 'Hosts', 'VMs', 'Datastores', 'Networks', and 'Updates'. The 'Configure' tab is active, and the left-hand navigation pane shows the 'VM/Host Groups' section selected under 'Configuration'. The main content area is titled 'VM/Host Groups' and contains a table with the following data:

| Name | Type |
|------------|------------|
| HostsSiteA | Host Group |
| HostsSiteB | Host Group |
| SiteAVMs | VM Group |
| SiteBVMs | VM Group |

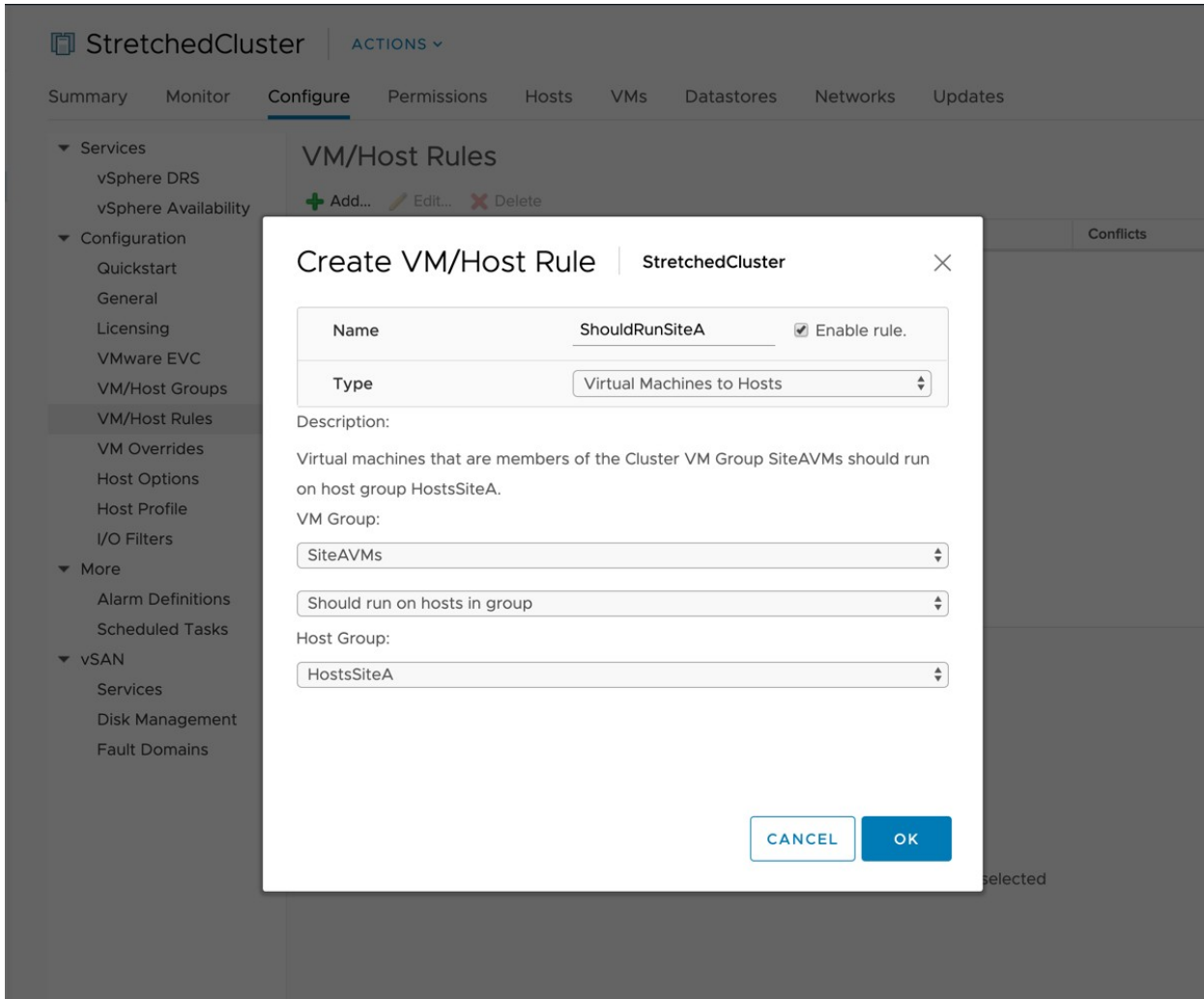
Below the table, there is a section for 'SiteBVMs Group Members' with the following entries:

| Name |
|-------|
| WSFC2 |
| APP2 |

Create VM/Host Rules

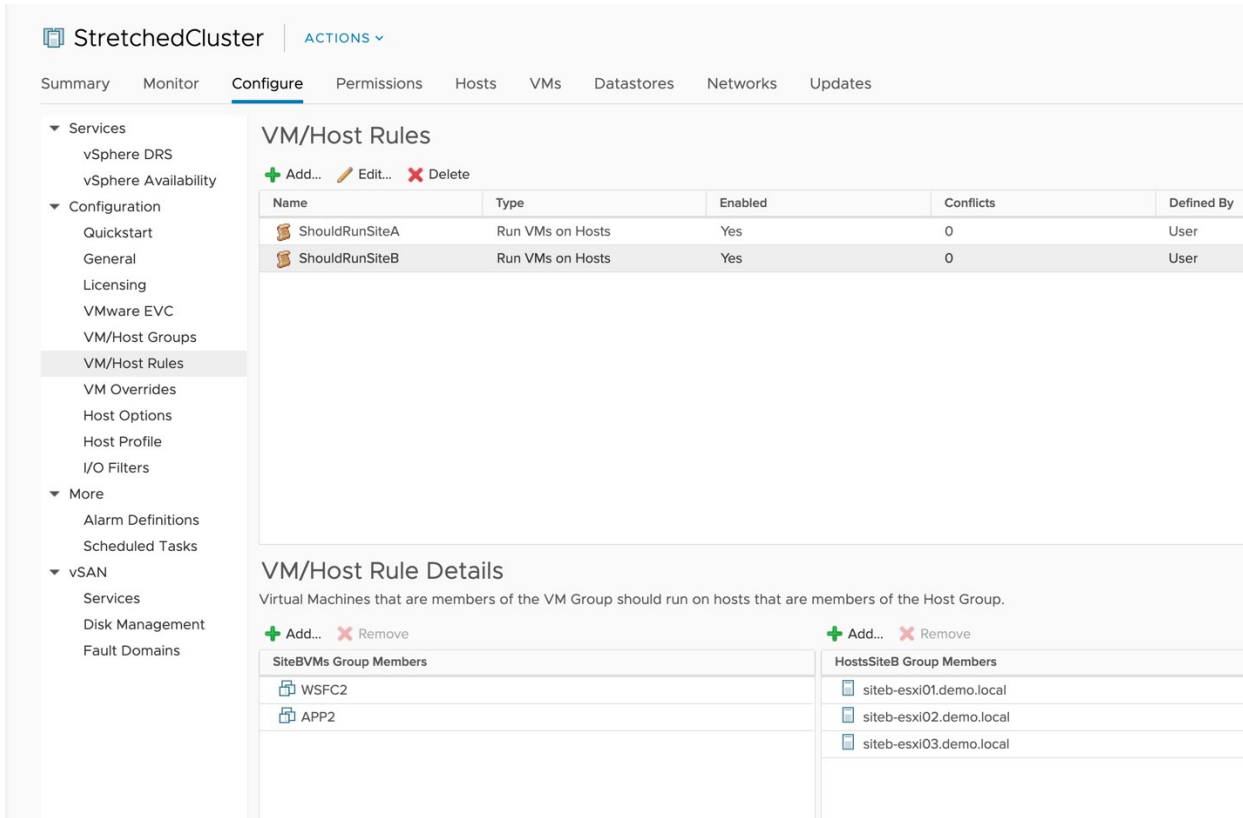
Now that the host groups and VM groups are created, it is time to associate them with host groups and ensure that particular VMs run on a particular site. Navigate to the VM/Host rules to associate a VM group with a host group.

In the example shown below, The VMs in the Site A VMs VM group with the host group called SiteAHosts, which will run the virtual machines in that group on the hosts in Site A.



This is a “should” rule. We use a “should” rule to allow vSphere HA to start the virtual machines on the other side of the stretched cluster in case of a site failure.

A VM/Host “should” rule must also be created for Site B.



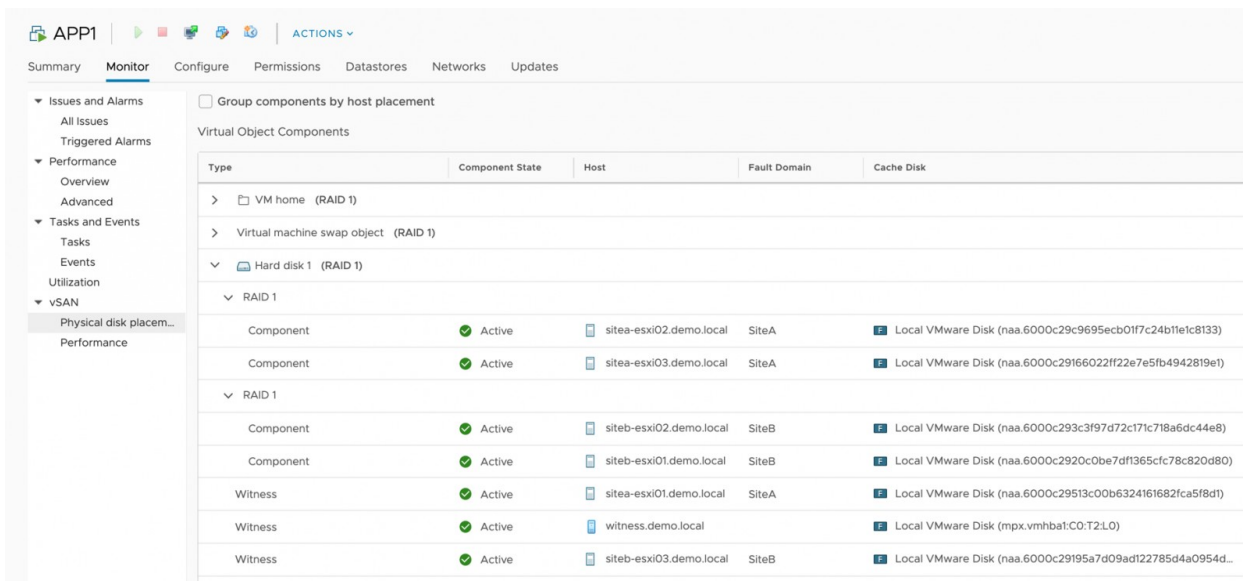
Note: DRS will be required to enforce the VM/Host Rules. Without DRS enabled, the soft “should” rules have no effect on placement behavior in the cluster.

That completes the setup of the vSAN Stretched Cluster.

The final steps are to power up the virtual machines created earlier and examine the component layout.

Verifying vSAN Stretched Cluster Component Layouts

In the example below, sitea-esxi01, sitea-esxi02, & sitea-esxi03 reside in Site A, and siteb-esx01, siteb-esxi02, and siteb-esx03 reside in Site B. Witness.demo.local is the witness. The layout shows that the VM has been deployed correctly.



The illustration shows:

- One mirrored copy of the data and a witness component reside on storage in Site A
- A second mirrored copy of the data and a witness component reside on storage in Site B

- A Witness component resides on the vSAN Witness Host

Warning: Disabling and re-enabling vSAN in a stretched cluster environment has the following behaviors:

- The witness configuration is not persisted. When recreating a vSAN Stretched Cluster, the vSAN Witness Host must be re-configured.
- If using the same vSAN Witness as before, the vSAN Witness Host's disk group will need to be deleted before it may be added back.

This can be done by:

- Using the `esxcli vsan storage remove` command From an ESXi host console session (local or SSH) Using the vSphere CLI
- Executing a similar command remotely using PowerCLI.
- Removing the partition from the vSAN Witness Host cache disk using the vSphere UI. Fault Domains are persisted, but vSAN does not know which FD is to be designated as Preferred.

Fault Domains are persisted, but vSAN does not know which FD is to be designated as Preferred.

Upgrading an older vSAN Stretched Cluster

Upgrading a vSAN Stretched Cluster is very easy. It is essential to follow a sequence of steps to ensure the upgrade goes smoothly.

Upgrade vCenter Server

As with any vSphere upgrades, it is typically recommended to upgrade the vCenter Server first.

Upgrade Hosts in Each Site

Upgrading hosts at each site is the next task to be completed. There are a few considerations to remember when performing these steps.

As with any upgrade, hosts must be put in maintenance mode, remediated, upgraded, and rebooted. It is important to consider the amount of available capacity at each site. In sites that have sufficient available capacity, it would be desirable to choose the “full data migration” vSAN data migration method. This method is preferred when site locality is important for read operations. When the “ensure accessibility” method is selected, read operations will traverse the inter-site link. Some applications may be more sensitive to the additional time required to read data from the alternate site.

With vSphere DRS in place, as hosts are put in maintenance mode, it is important to ensure the previously described VM/host groups and VM/host rules are in place. These rules will ensure that virtual machines are moved to another host in the same site. If DRS is set to “fully automated” virtual machines will vMotion to other hosts automatically, while “partially automated” or “manual” will require the virtualization admin to vMotion the virtual machines to other hosts manually.

It is recommended to sequentially upgrade hosts at each site first, followed by sequentially upgrading the hosts at the alternate site. This method will introduce the least amount of additional storage traffic. While it is feasible to upgrade multiple hosts simultaneously across sites when there is additional capacity, as data is migrated within each site, this method will require additional resources.

Upgrade the Witness Appliance

After both data sites have been upgraded, the vSAN Witness Host will also need to be upgraded. The vSAN Witness Host can be upgraded using vSphere Update Manager or any other approved method to update ESXi.

The vSAN Health Check may show an On-Disk Format inconsistency, but it no longer reports that the vSAN Witness Host is different version. Different releases of vSAN often have updated On-Disk Format versions.

The screenshot shows the vSAN Health page in the vSphere Client. The left-hand navigation pane is expanded to 'vSAN' > 'Health'. The main content area shows the 'Cluster' health status, which is overall green but has a yellow warning icon. One of the health checks, 'Disk format version', is highlighted with a red rectangular box. This check shows a yellow warning icon and a red border, indicating an issue. Other health checks like 'ESXi vSAN Health service installation', 'vSAN Health Service up-to-date', and 'vSAN extended configuration in sync' are all green.

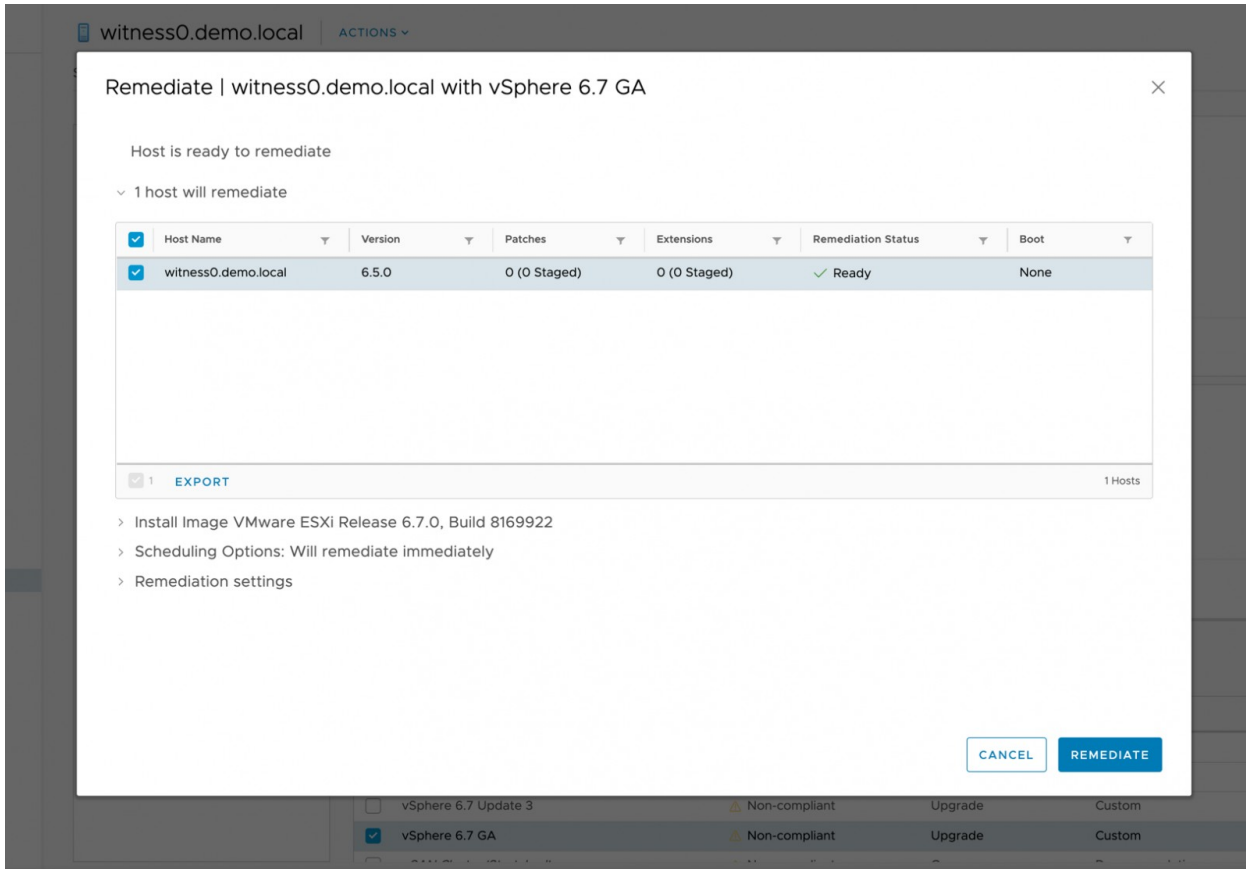
Expanding this Disk format version error does show that the vSAN Witness Host needs to be upgraded.

This screenshot shows the 'Disk format version' error expanded. A modal window titled 'Detailed vSAN disks format status' is open, displaying a table of disk information. The table has four columns: 'vSAN host', 'Disks with older format', 'Check Result', and 'Recommendation'. The 'witness0.demo.local' host is highlighted with a red box, showing a yellow warning icon in the 'Check Result' column and the recommendation 'Perform software upgrade first before re...'. Other hosts like 'siteb-02.demo.local', 'sitea-02.demo.local', 'sitea-01.demo.local', and 'siteb-01.demo.local' also show yellow warning icons and recommendations for on-disk format upgrades.

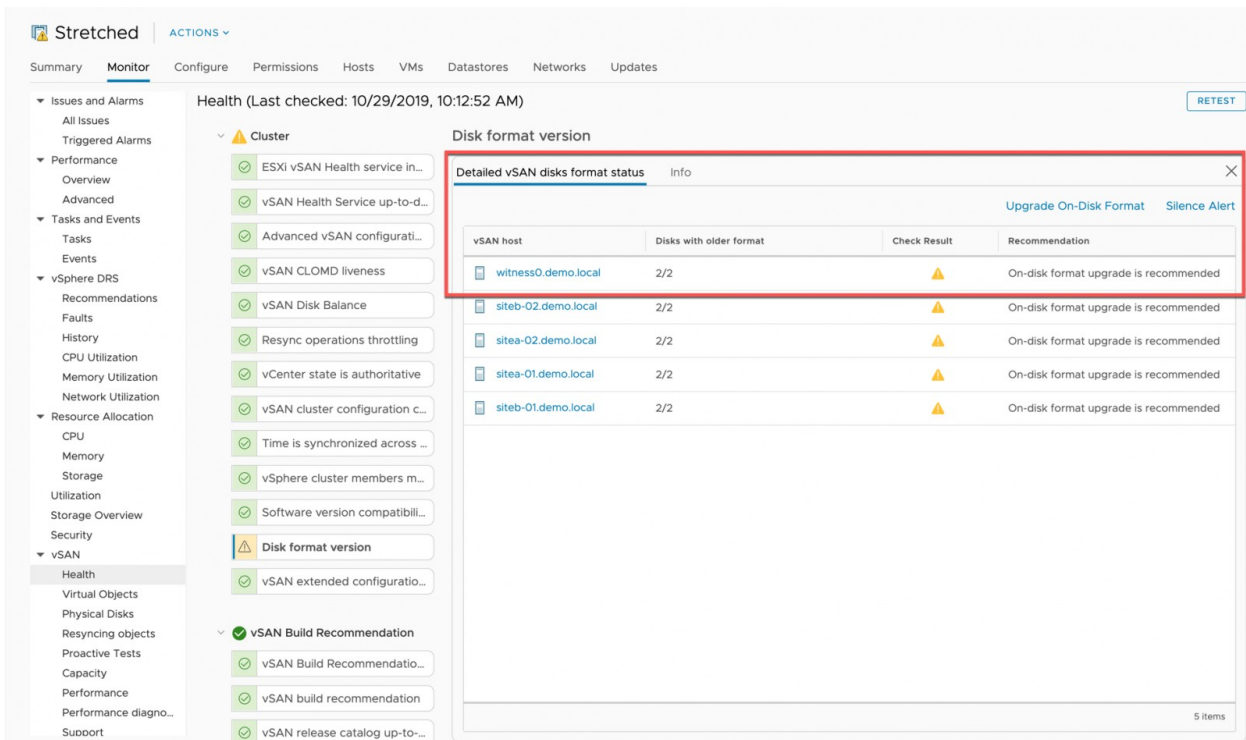
| vSAN host | Disks with older format | Check Result | Recommendation |
|---------------------|-------------------------|--------------|---|
| witness0.demo.local | 2/2 | ⚠ | Perform software upgrade first before re... |
| siteb-02.demo.local | 2/2 | ⚠ | On-disk format upgrade is recommende... |
| sitea-02.demo.local | 2/2 | ⚠ | On-disk format upgrade is recommende... |
| sitea-01.demo.local | 2/2 | ⚠ | On-disk format upgrade is recommende... |
| siteb-01.demo.local | 2/2 | ⚠ | On-disk format upgrade is recommende... |

When using a vSAN Witness Appliance as the vSAN Witness Host, it is essential to remember to use a VMware-provided vSphere ISO to upgrade and not an OEM-provided vSphere ISO. OEM-provided vSphere ISOs often have additional OEM vendor software specific to their systems. This additional software is not required on the vSAN Witness Appliance.

Upgrading the vSAN Witness Host can easily be accomplished by vSphere Update Manager.



With the vSAN Witness Host upgraded to a newer release, the On-Disk Format can now be upgraded to a format consistent with the vSAN release.



Upgrading the On-Disk Format is relatively quick and does not require data to be evacuated from disk groups or a rolling reformat. The only exception to this rule is moving from an On-Disk Format version previous to version 3.

The screenshot shows the vSAN Stretched Cluster Configuration UI. The 'Configure' tab is active, and the 'CLAIM UNUSED DISKS' section is expanded. A table lists disk groups for various hosts, including witness0, sitea-01, siteb-01, and siteb-02. Each row shows the disk group name, the number of disks (2 of 2), the state (Connected or Mounted), the vSAN Health Status (Healthy), the disk type (All flash), the fault domain (Sec), the network partition group (Group 1), and the disk format version (5). At the top right of the configuration area, there are 'UPGRADE' and 'PRE-CHECK UPGRADE' buttons. A status message at the top indicates that the cluster is ready to upgrade.

The On-Disk Format upgrade can be launched from the vSAN Health Check, or it can be launched from the Disk Management menu in the vSAN Cluster Configuration UI. VMware recommends using the PRE-CHECK UPGRADE option before initiating an upgrade.

The final step in the upgrade process will be to upgrade the on-disk format.

The screenshot shows a dialog box titled 'Upgrade | Stretched'. It contains a checkbox for 'Allow Reduced Redundancy' which is currently unchecked. Below the checkbox, there is a warning icon and a paragraph of text: 'Upgrading the vSAN on-disk format is a long running operation. Once you upgrade the on-disk format, you cannot rollback software on the hosts or add certain older hosts to the cluster, as explained in this KB . This operation upgrades one disk group at a time.' At the bottom right of the dialog, there are two buttons: 'CANCEL' and 'UPGRADE'.

Once the on-disk format is complete, the cluster has been upgraded.

Management and Maintenance

The following section of the guide covers considerations related to the management and maintenance of a vSAN Stretched Cluster configuration.

Maintenance Mode Consideration

vSAN Stretched Cluster configurations have two scenarios: maintenance mode on a Site Host and maintenance mode on the vSAN Witness Host.

Maintenance Mode on a Site Host

Maintenance mode in vSAN Stretched Clusters is site-specific. All maintenance modes (Ensure Accessibility, Full data migration, and No data migration) are supported. However, to do a Full Data Migration, there should be enough resources in the same site to facilitate the rebuilding of components on the remaining nodes on that site.

Maintenance Mode on the vSAN Witness Host

Maintenance mode on the vSAN Witness Host should be infrequent as it will not run any virtual machines. When maintenance mode is performed on the vSAN Witness Host, the witness components cannot be moved to either site. When the vSAN Witness Host is put in maintenance mode, it behaves as the No data migration option would on-site hosts. It is recommended to check that all virtual machines are in compliance and that there is no ongoing failure before doing maintenance on the vSAN Witness Host.

While the vSAN Witness Host is in maintenance mode, the cluster cannot survive a site failure. Maintenance of the vSAN Witness Host should be kept to the

Monitoring

Native VMware Aria Operations dashboards built into vCenter Server can display intelligence for vSAN stretched clusters. The VMware Aria Operations dashboard view knows the host's vSAN stretched cluster configuration and separates the hosts into two sites. A deeper level of insights and analytics has been introduced with the VMware Aria Operations dashboard for a stretched cluster.

For more details, please read the VMware Aria Operations blog post [here](#).

Updates using vLCM

vSphere Lifecycle Manager (vLCM) is a solution for unified software and firmware lifecycle management. vLCM is enhanced with firmware support for Lenovo ReadyNodes, awareness of vSAN stretched cluster and fault domain configurations, additional hardware compatibility pre-checks, and increased scalability for concurrent cluster operations. In vSAN 7 Update 3, vLCM supports topologies that use dedicated witness host appliances. Both 2-node and stretched cluster environments can be managed and updated by vLCM, guaranteeing a consistent, desired state of all hosts participating in a cluster using these topologies. It also performs updates in a recommended order for easy cluster upgrades.

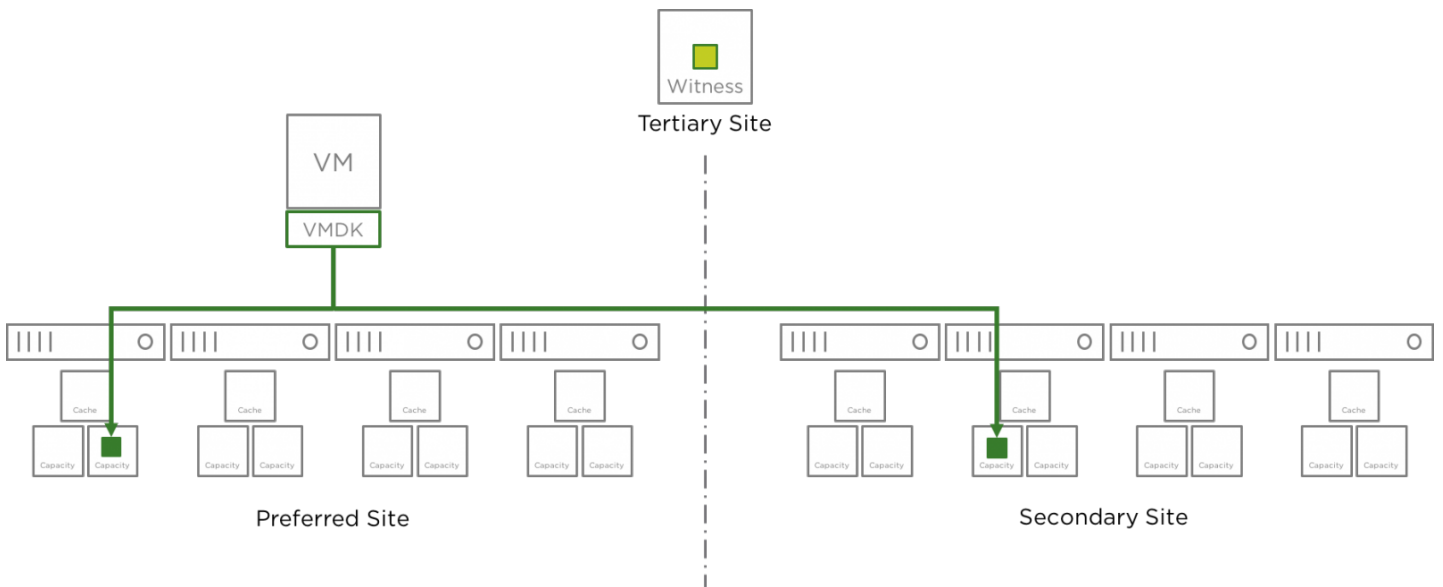
Failure Scenarios

In this section, we will discuss the behavior of the vSAN Stretched Cluster when various failures occur.

Failure Scenarios and Component Placement

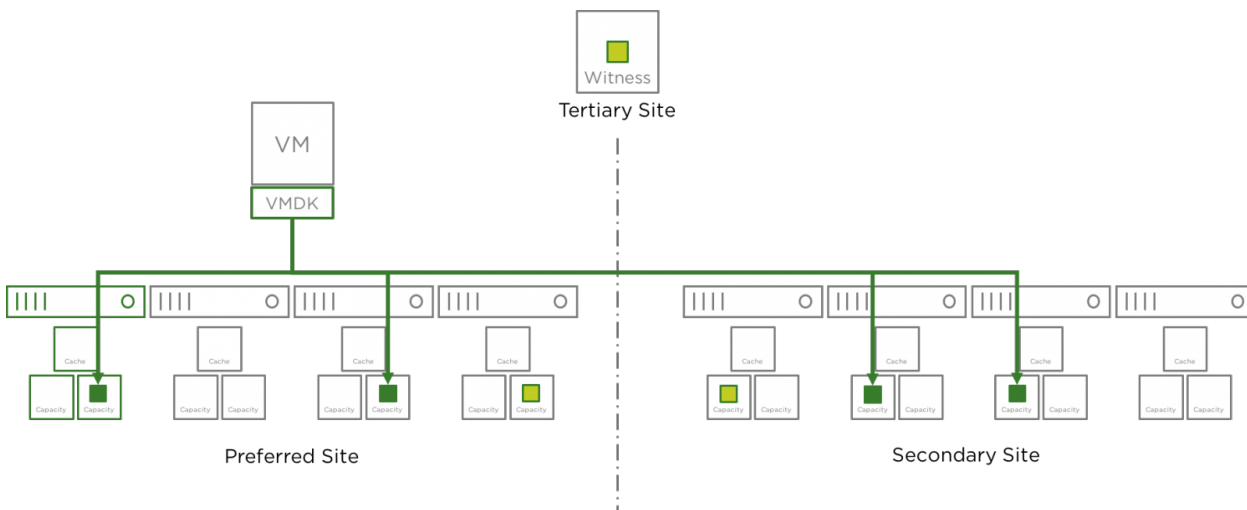
Understanding component placement is paramount in understanding failure scenarios. The illustration shows the placement of a vSAN Object's components in a Stretched Cluster Scenario.

The virtual machine's virtual disk (vmdk) has one component placed in the Preferred Site, one in the Secondary Site, and a Witness component in the Tertiary Site that houses the vSAN Witness Host.



The illustration shows a storage policy that will protect across sites but not within a site.

Below is a cluster with data protection (mirroring) across sites and local data protection (mirroring in this case) within each data site.



vSAN Stretched Clusters can support up to a single complete site failure and a policy-based maximum of host failures within a site.

If a site has more failures than the local protection policy will allow, then the site is considered failed. It is important to remember that the vSAN Witness Host, residing in the Tertiary site, is only a single host. Because of this, a failure of the vSAN Witness Host is also considered a site failure.

The scenarios in this section will cover different failure behaviors.

Individual Host Failure or Network Isolation

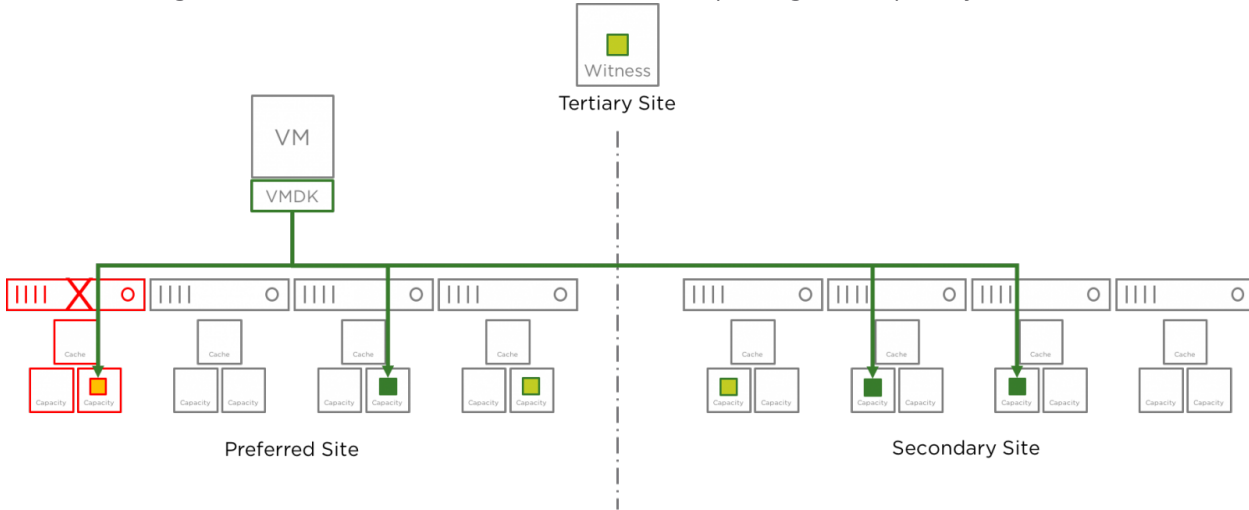
What happens when a host fails or is network isolated?

vSAN will mark the component absent. The VM will continue to run or it will be rebooted by vSphere HA if the VM was running on the host that went offline.

If the policy includes Local Protection, reads will be serviced by the remaining components within the same site.

- The component will be rebuilt within the same site after 60 minutes if an additional host is available and the failed host does not return online.
- If no additional hosts are available within the site, the component will only be rebuilt if the failed/isolated host returns online.
- When multiple hosts containing the components fail or are isolated, reads will be serviced across the inter-site link.

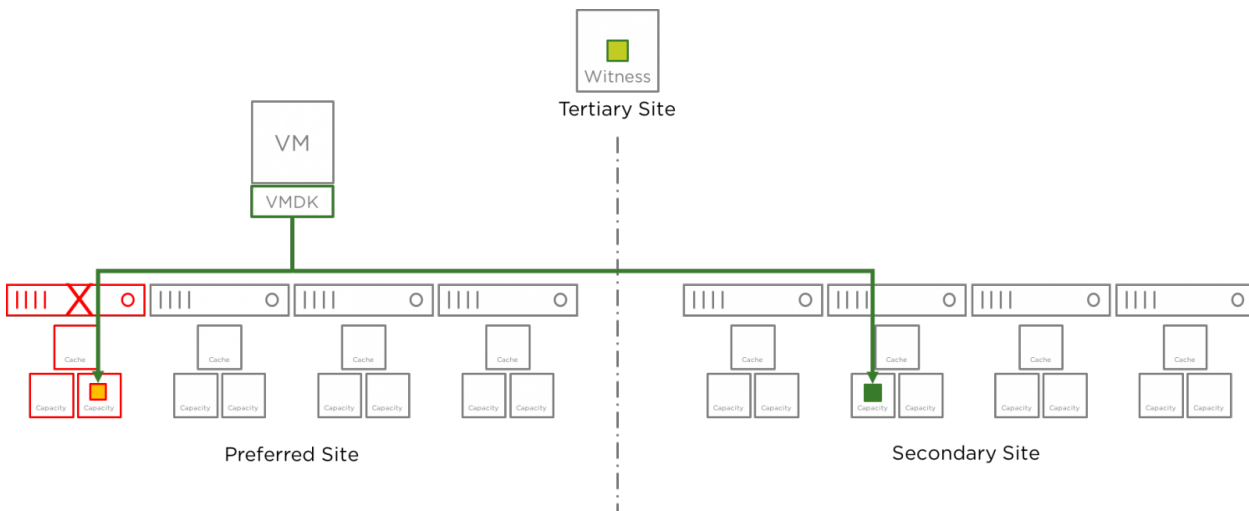
This can be a significant amount of traffic on the inter-site link depending on the quantity of data on the failed host.



If the policy does not include Local Protection, reads will be serviced across the inter-site link.

- This can be a significant amount of traffic on the inter-site link depending on the quantity of data on the failed host.
- The component will be rebuilt within the same site as the failed/isolated hosts after 60 minutes if there is an alternate

If there are no additional hosts available, or hosts are at capacity, the component will only be rebuilt if the failed/isolated host comes back online.



Individual Drive Failure

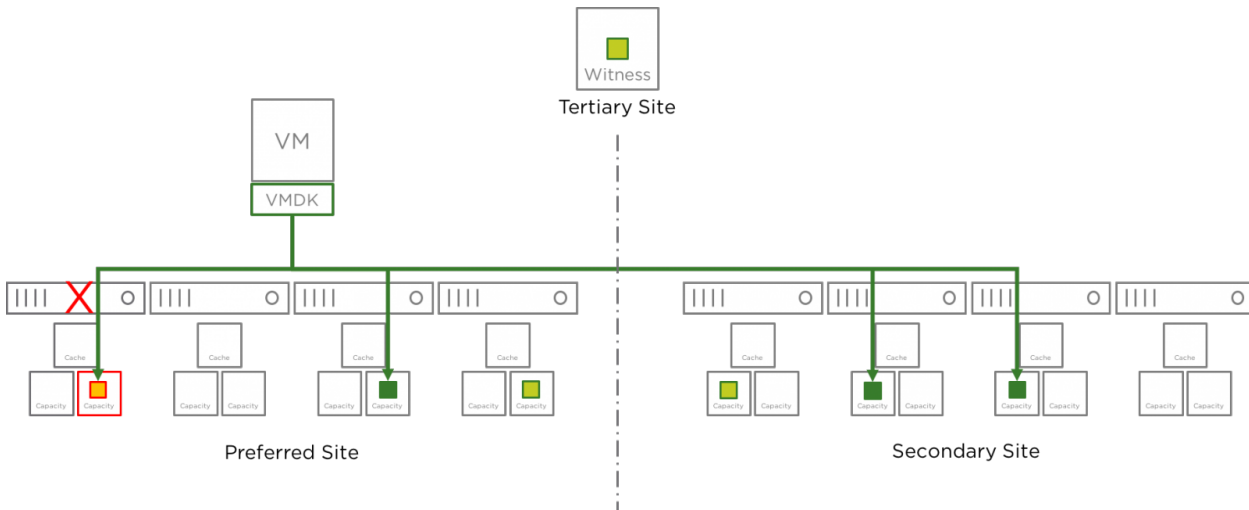
What happens when a drive fails?

vSAN will mark the component absent. The VM will continue to run.

If the policy includes Local Protection, reads will be serviced by the remaining components within the same site.

- The component will be rebuilt within the same site after 60 minutes if an additional host is available and the failed host does not come back online.
- If no additional hosts are available within the site, the component will only be rebuilt if the failed/isolated host comes back online.
- When multiple hosts containing the components fail or are isolated, reads will be serviced across the inter-site link.

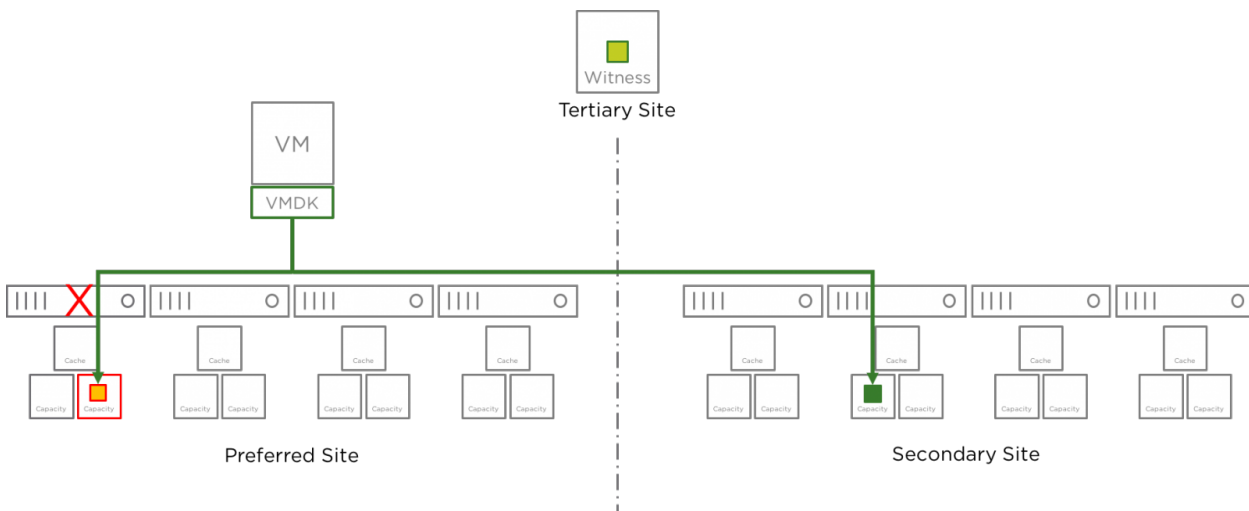
This can be a significant amount of traffic on the inter-site link depending on the data on the failed host.



If the policy does not include Local Protection, reads will be serviced across the inter-site link.

- This can be a significant amount of traffic on the inter-site link depending on the amount of data on the failed host.
- The component will be rebuilt within the same site as the failed/isolated hosts after 60 minutes if there is an alternate

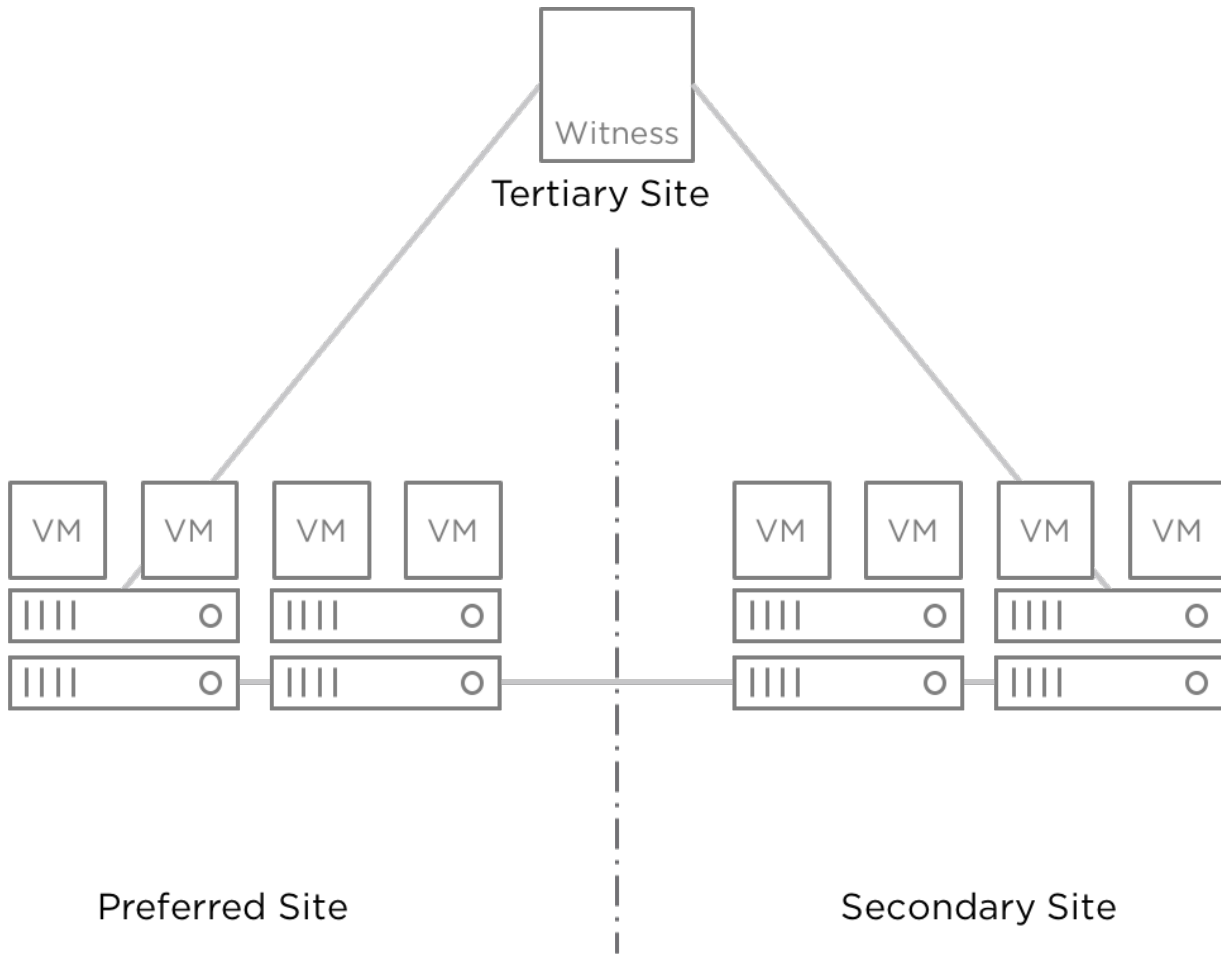
If there are no additional hosts available, or hosts are at capacity, the component will only be rebuilt if the failed/isolated host comes back online.



Site Failure or Network Partitions

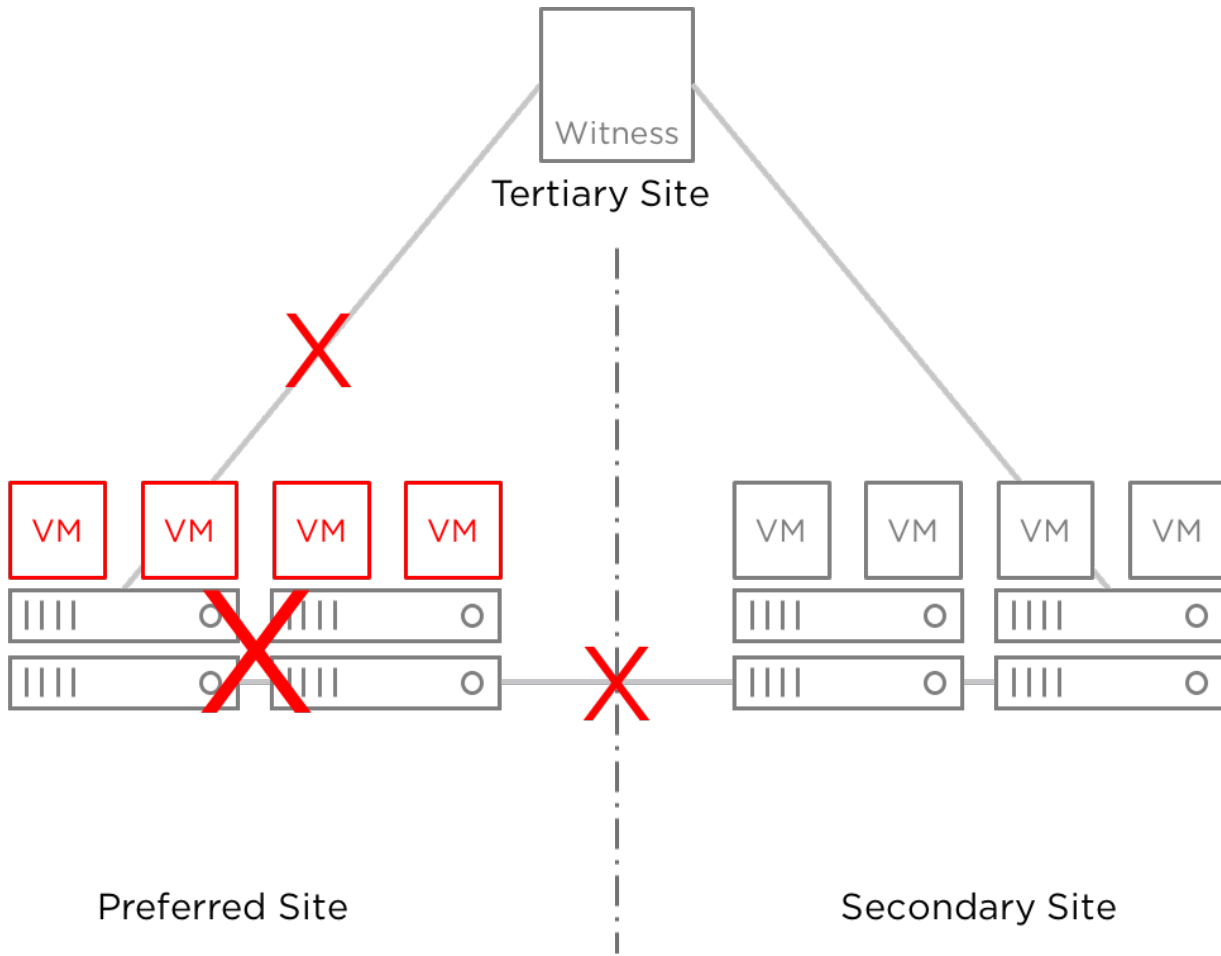
What happens when sites go offline or lose connectivity?

A typical vSAN Stretched Cluster configuration can be seen here:

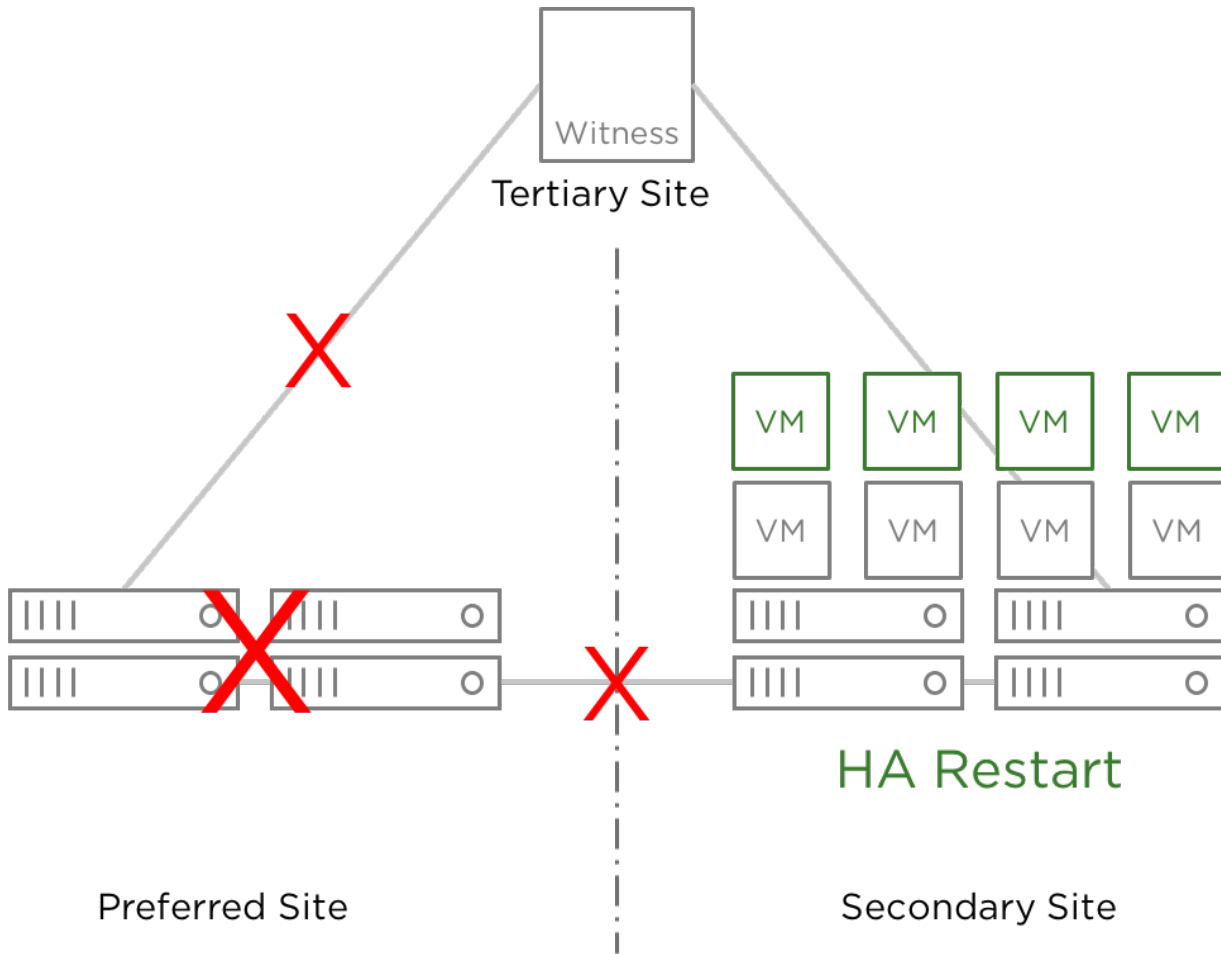


Preferred Site Failure or Completely Partitioned

In the event the Preferred Site fails or is partitioned, vSAN powers the virtual machines running in that site off. The reason for this, is because the virtual machine's components are not accessible due to the loss of quorum. The vSAN Stretched Cluster has now experienced a single site failure. The loss of either site in addition to the witness is two failures, will take the entire cluster offline.

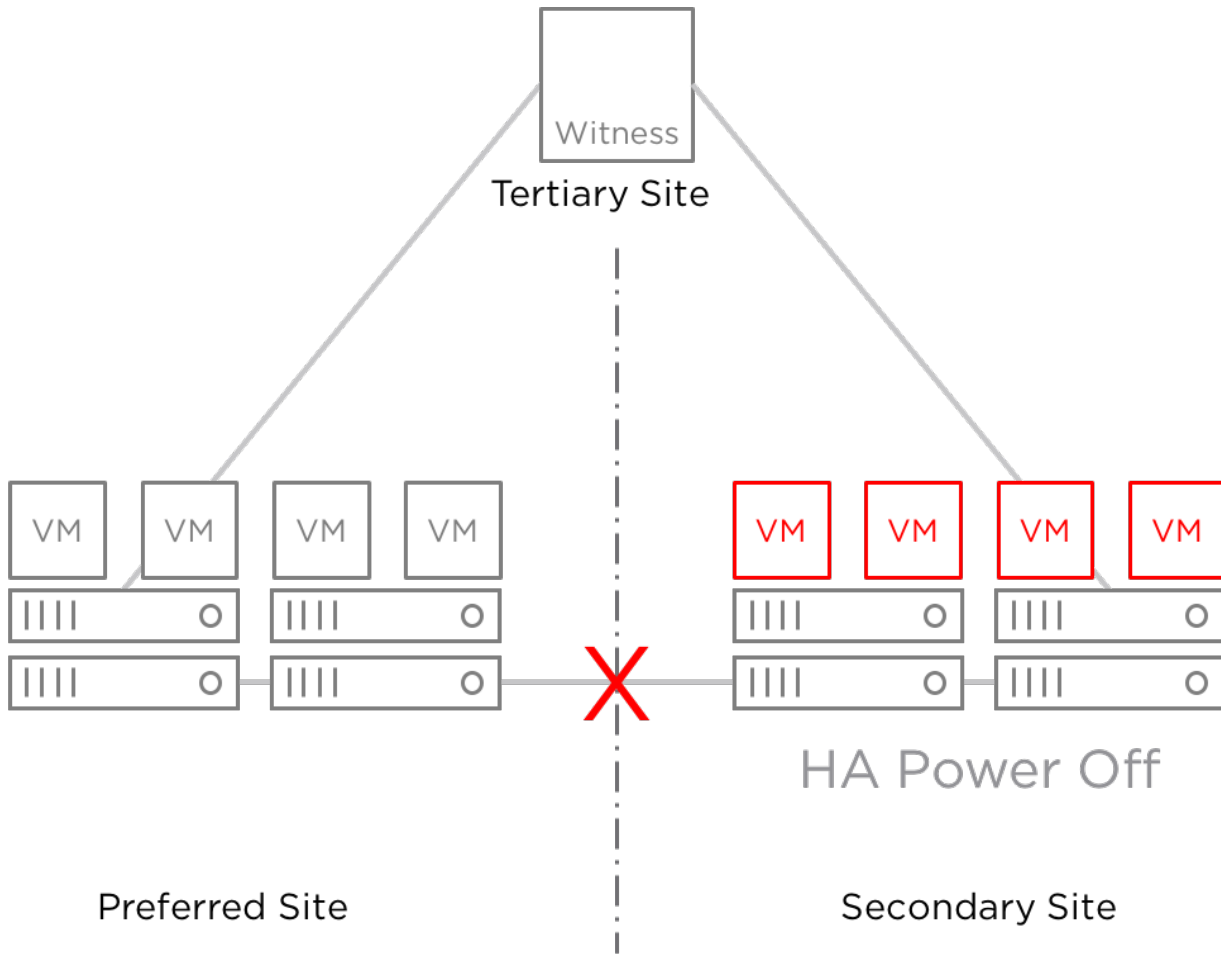


An HA primary node will be elected in the Secondary Site, which will validate which virtual machines are to be powered on. Because quorum has been formed between the vSAN Witness Host and the Secondary Site, virtual machines in the Secondary Site will have access to their data, and can be powered on.

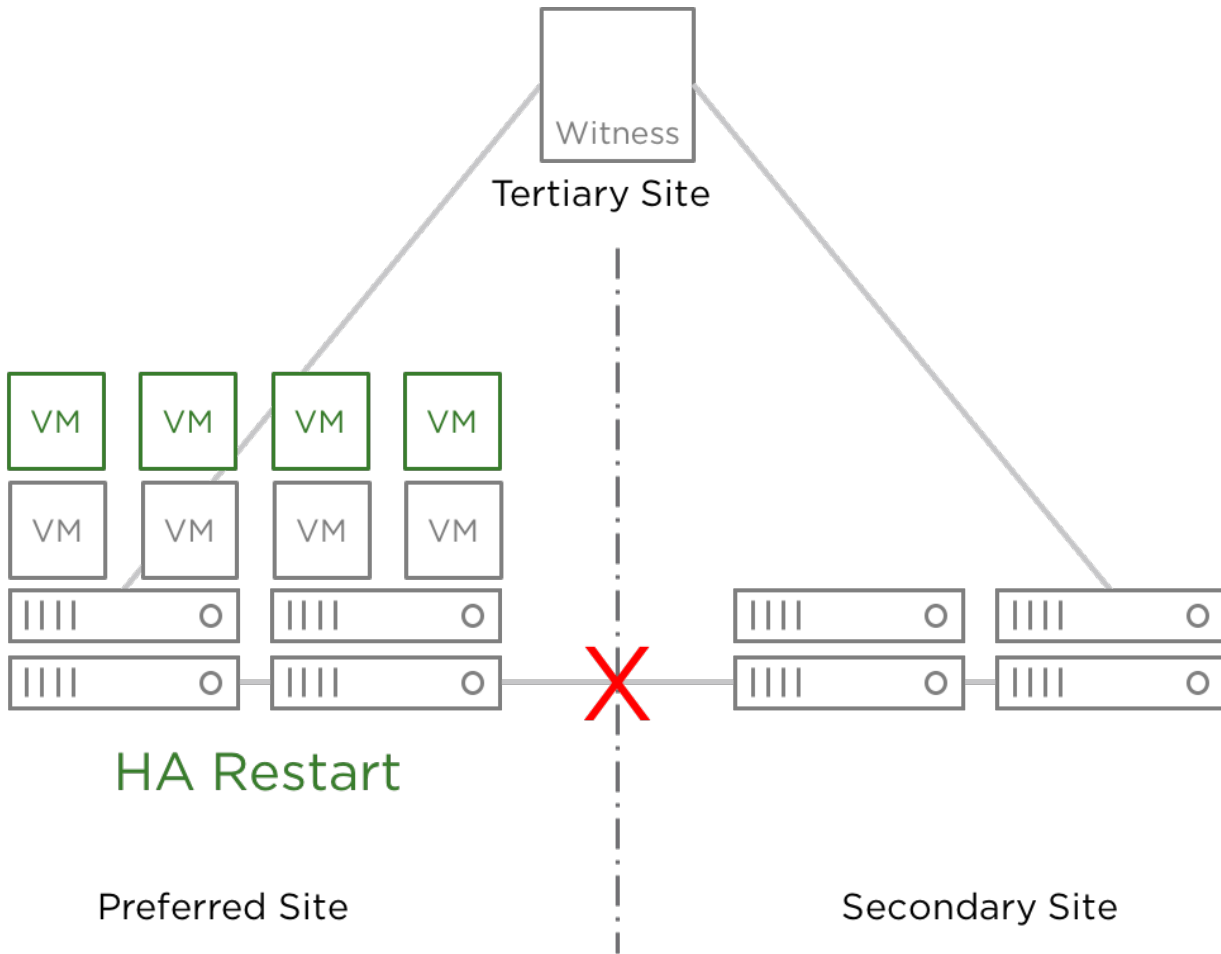


Secondary Site Failure or Partitioned

In the event the Secondary Site fails or is partitioned, vSAN powers the virtual machines running in that site off. The reason for this, is because the virtual machine's components are not accessible due to the loss of quorum. The vSAN Stretched Cluster has now experienced a single site failure. The loss of either site in addition to the witness is two failures, will take the entire cluster offline.

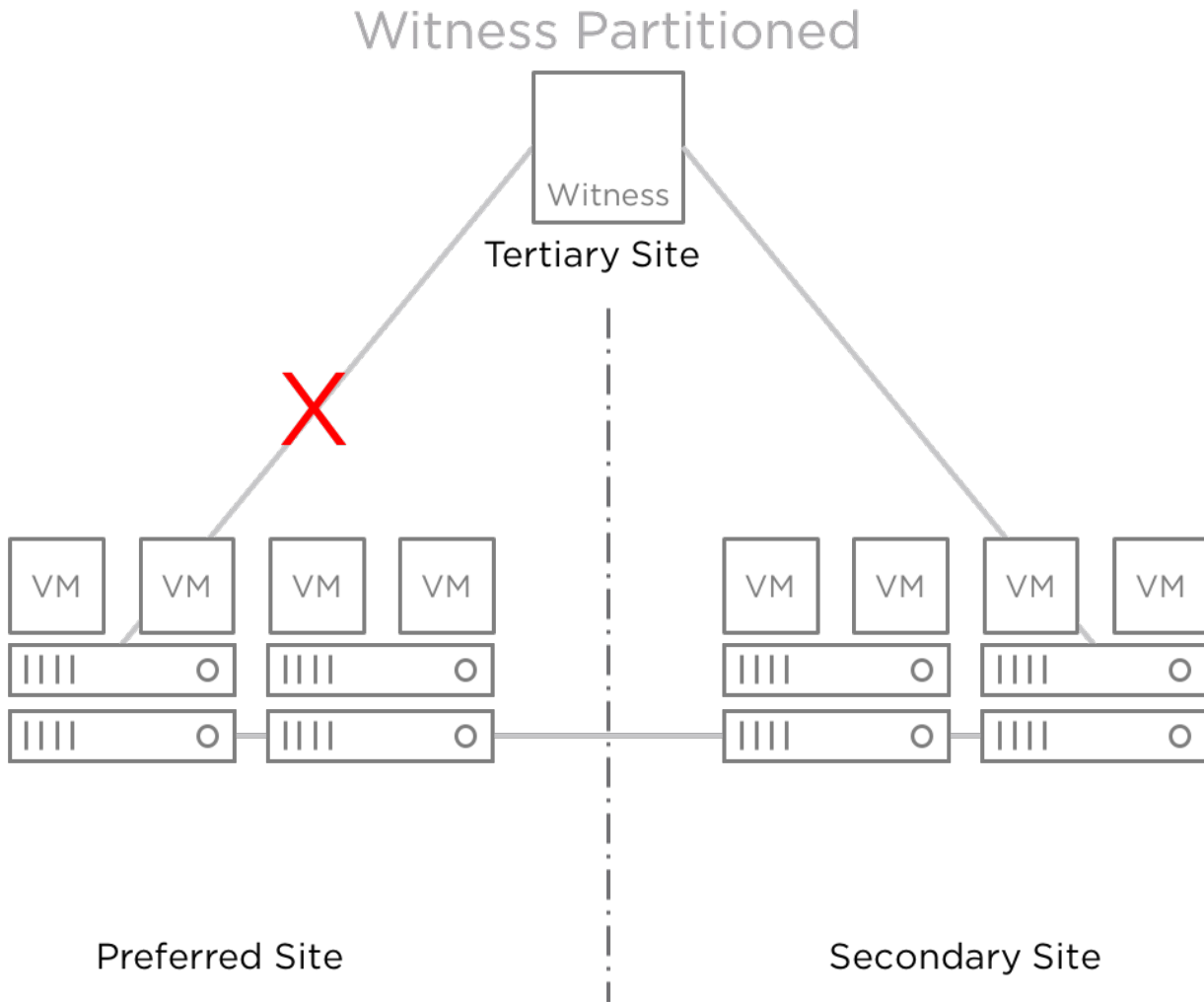


The HA primary node on the Preferred Site, will validate which virtual machines are to be powered on. Virtual machines which have been moved to the Preferred Site will now have access to their data, and can be powered on.

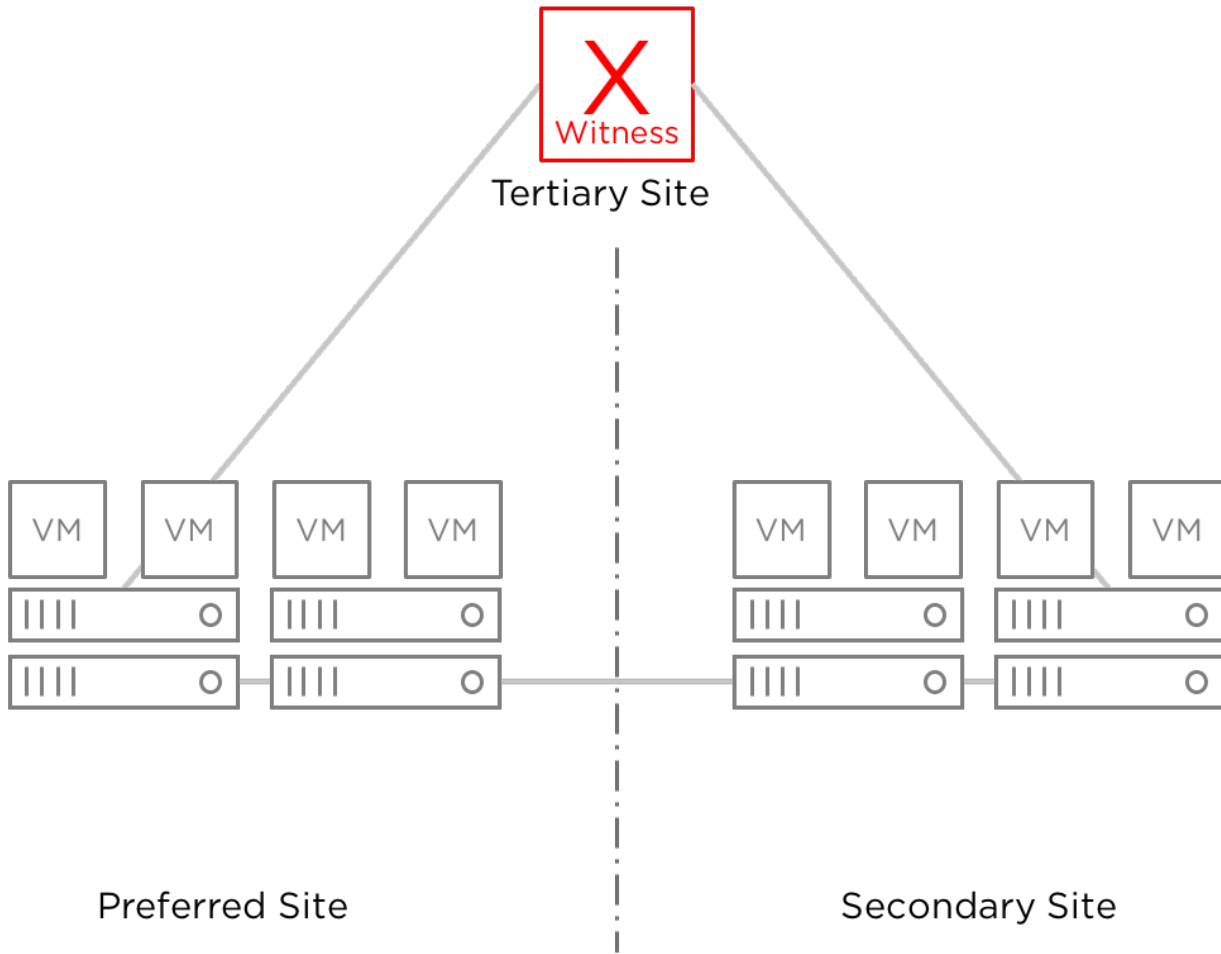


vSAN Witness Host Failure or Partitioned

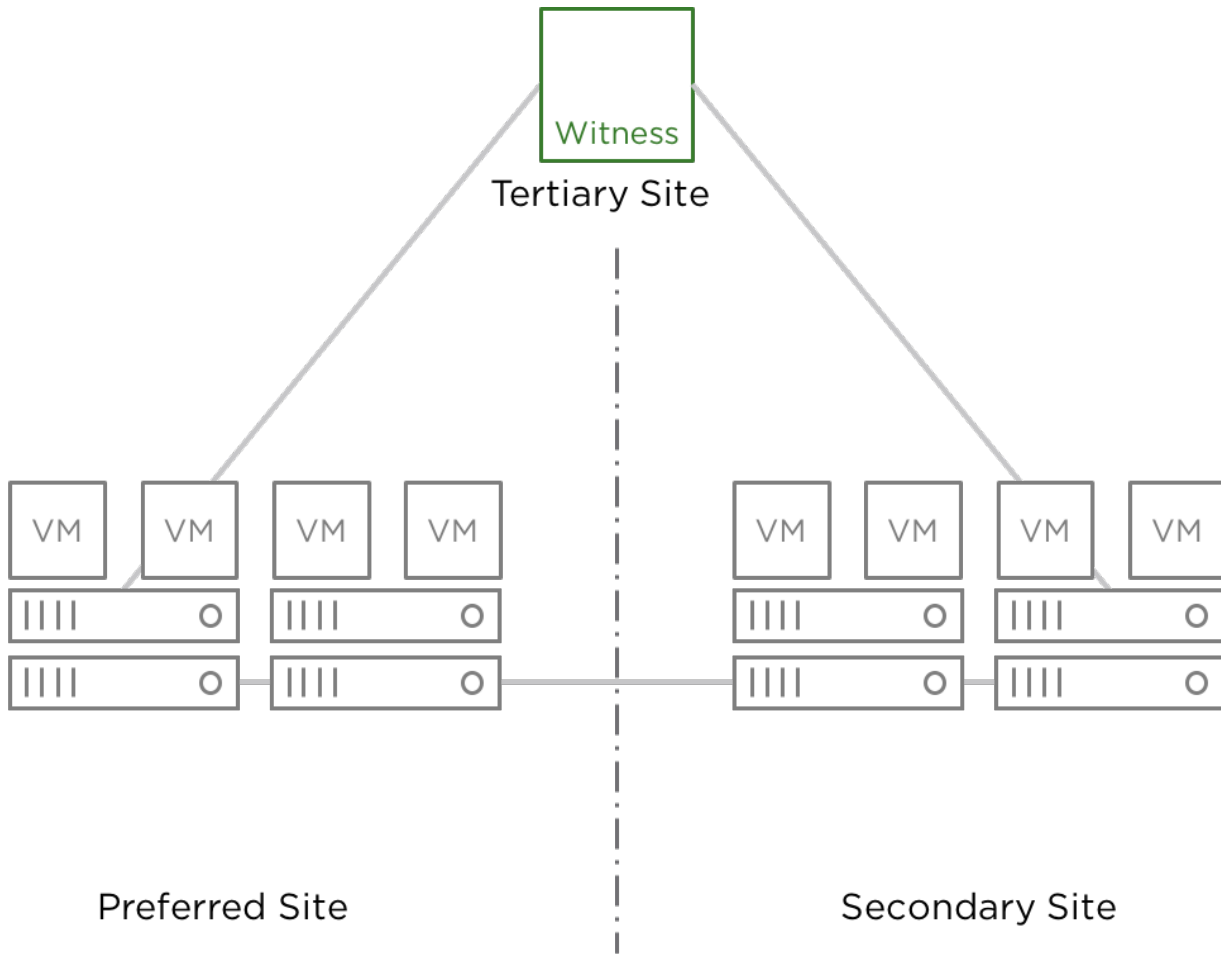
Virtual machines running in both of the main sites of a Stretched Cluster are not impacted by the vSAN Witness Host being partitioned. Virtual machines continue to run at both locations. The vSAN Stretched Cluster has now experienced a single site failure. The loss of the secondary site, in addition to the witness, is two failures that will take the entire cluster offline.



In the event the vSAN Witness Host has failed, the behavior is the same as if the vSAN Witness Hosts were partitioned from the cluster. Virtual machines continue to run at both locations. Because the vSAN Stretched Cluster has now experienced a single site failure, it is important to either get the vSAN Witness Host back online, or deploy a new one for the cluster.

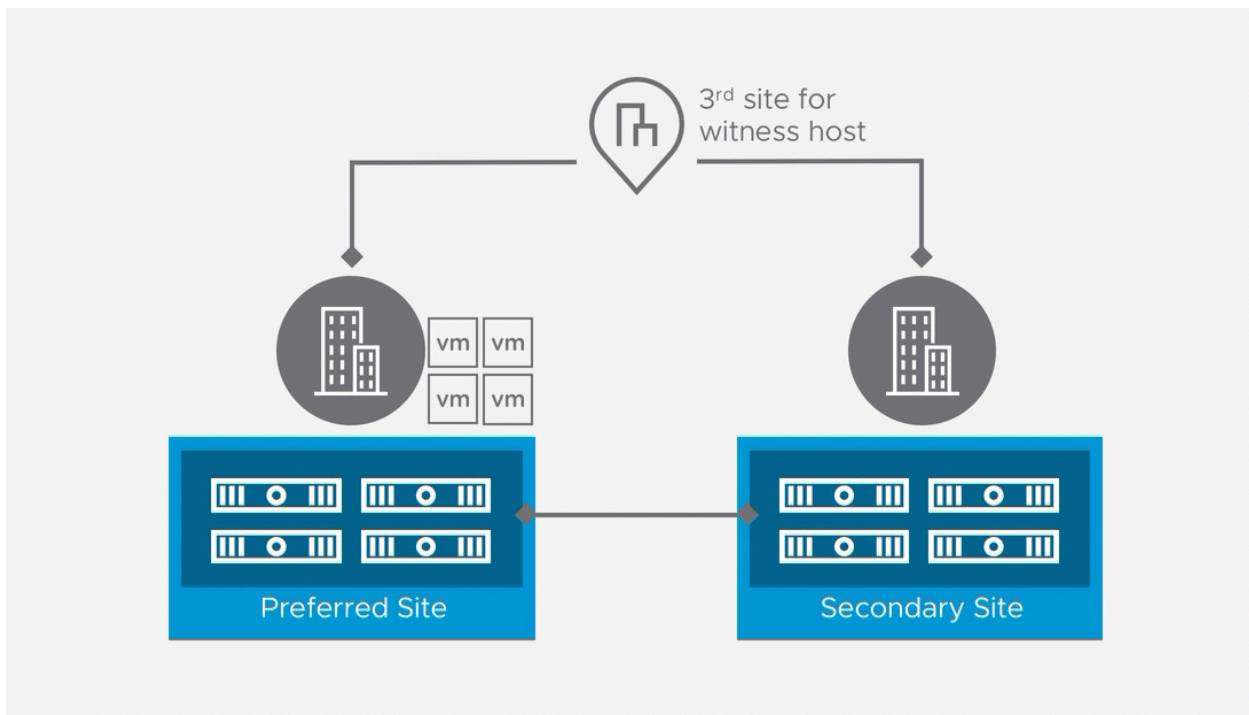


When the existing vSAN Witness Host comes back online or a new vSAN Witness Host is deployed, metadata changes are resynchronized between the main Stretched Cluster sites and the vSAN Witness Host. The amount of data that needs to be transmitted depends on a few items such as the number of objects and the number of changes that occurred while the vSAN Witness Host was offline. However, this amount of data is relatively small considering it is metadata, not large objects such as virtual disks.



Intelligent site continuity for stretched clusters

A number of improvements have been made to the failure response logic when tracking the "fitness" of a site for failover/failback in the event that links to the witness and adjacent data site become available at different times, illustrated below.



This enhancement addresses a scenario in which the preferred site is completely isolated (a break in the ISL link, and to the witness site), and vSAN properly fails over to the secondary site. In the event that the witness site regains connectivity to the preferred site, vSAN 6.7 will properly track the “fitness” of the site and maintain the Secondary Site as the active site until the ISL is recovered. This helps prevent a false positive of the primary site being back up, and failing back over to it, and attempting to use stale components

Multiple Simultaneous Failures

What happens if there are failures at multiple levels?

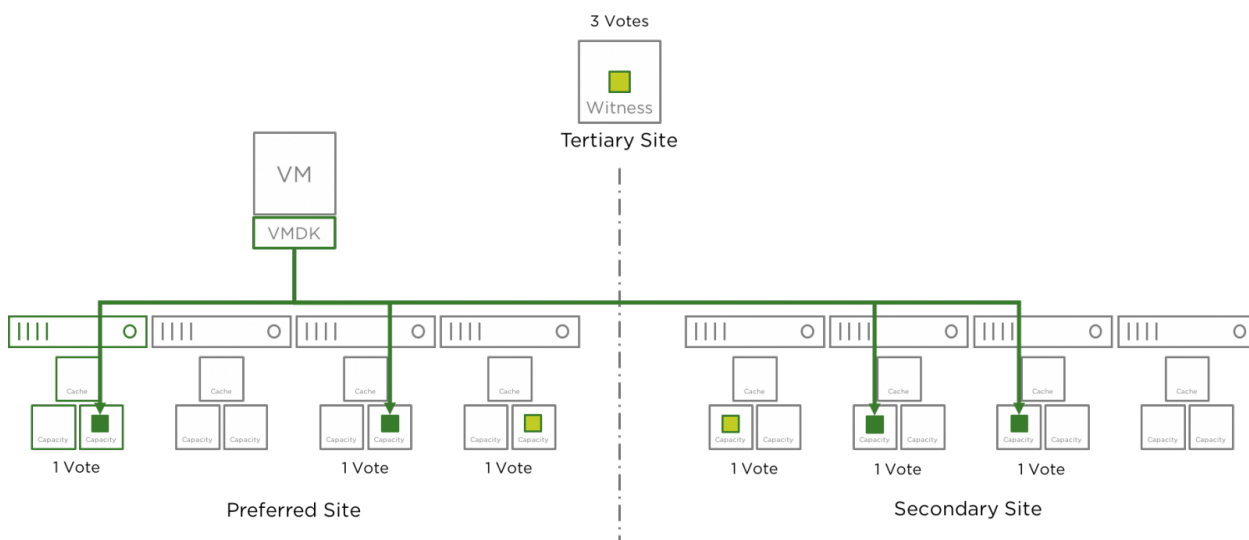
The scenarios thus far have only covered situations where a single failure has occurred.

Reviewing the vSAN Storage Policy Rule changes, Number of Failures to Tolerate became Primary Number of Failures to Tolerate , and is directly associated with site availability. Secondary Number of Failures to Tolerate was introduced with Local Protection, which works along with Failure Tolerance Method to determine data layout/placement within a Stretched Cluster site.

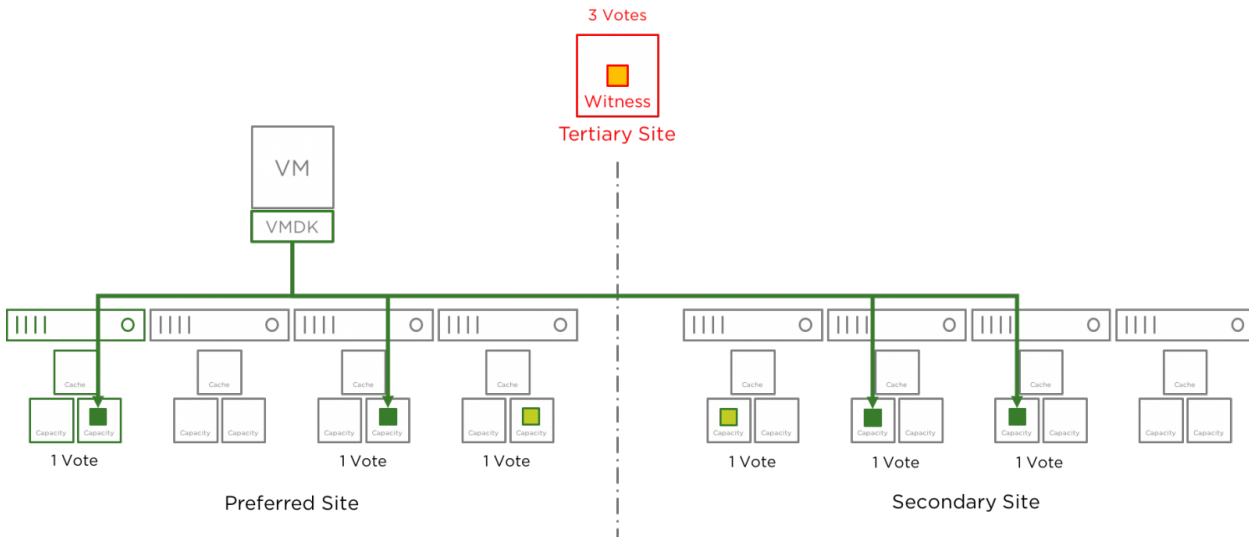
Votes and their contribution to object accessibility

The vSAN Design Guide goes into further detail about how component availability determines access to objects. In short, each component has a vote, and a quorum of votes must be present for an object to be accessible. Each site will have an equal number of votes and there will be an even distribution of votes within a site. If the total number of votes is an even number, a random vote will be added.

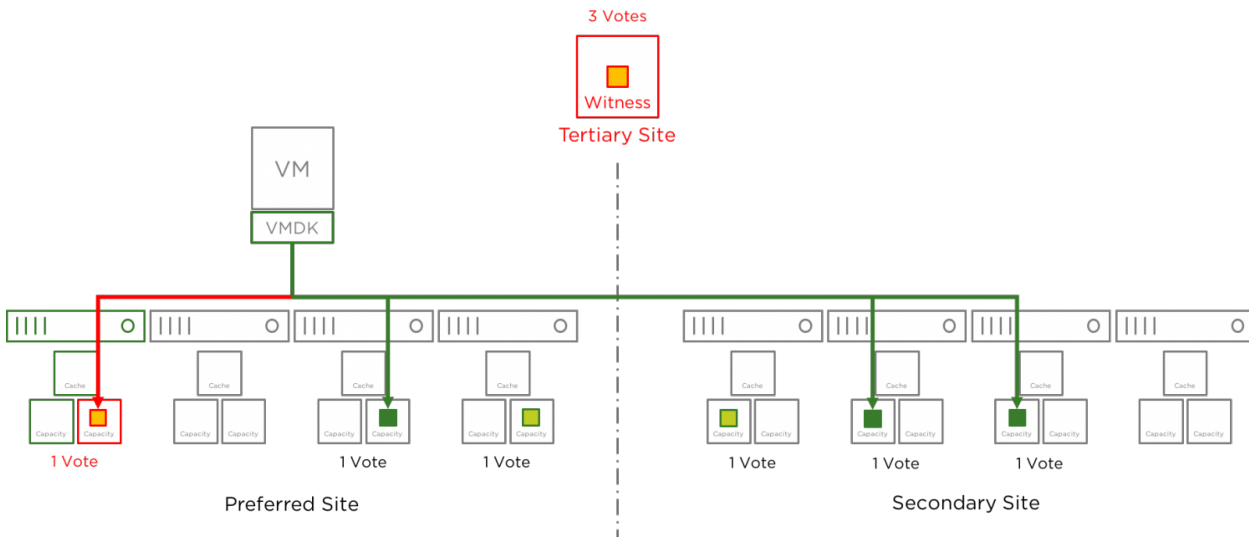
In the illustration below, an 8 Node vSAN Stretched Cluster (4+4+1) has an object with PFTT=1 (Mirrored across sites) and SFTT=1/FTM Mirroring (Local Protection). Each site has 3 votes each in this case, with a total of 9 votes.



If the vSAN Witness Host fails, 3 votes are no longer present, leaving only 6 accessible. Only 66% of the components are available, so the vmdk is still accessible.

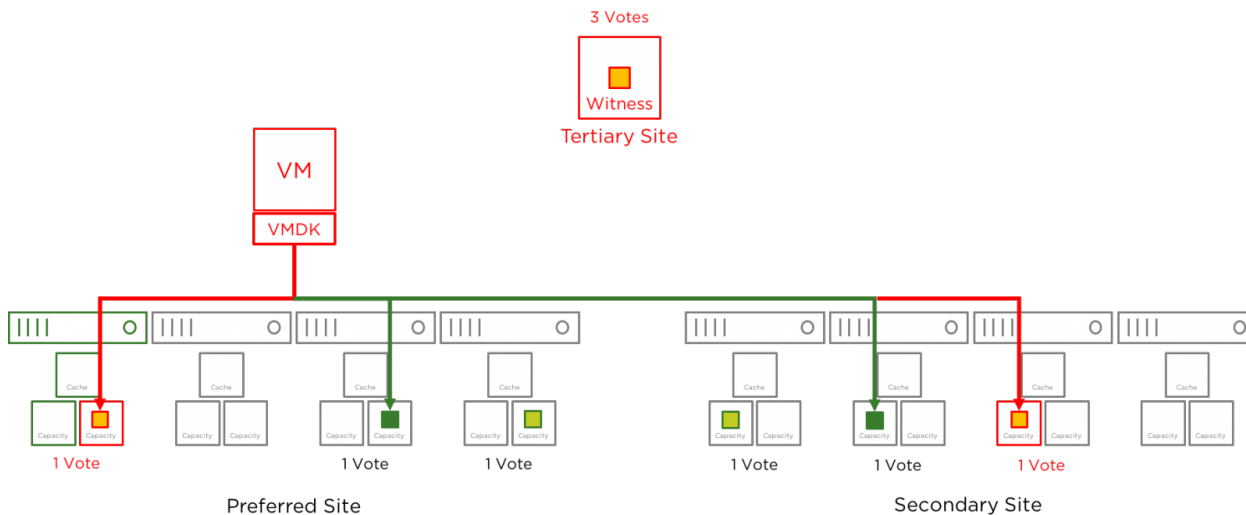


With the loss of a vSAN Disk, 4 votes are no longer present, leaving only 5 accessible. Because 55.5% of the components are available, the vmdk is still accessible.

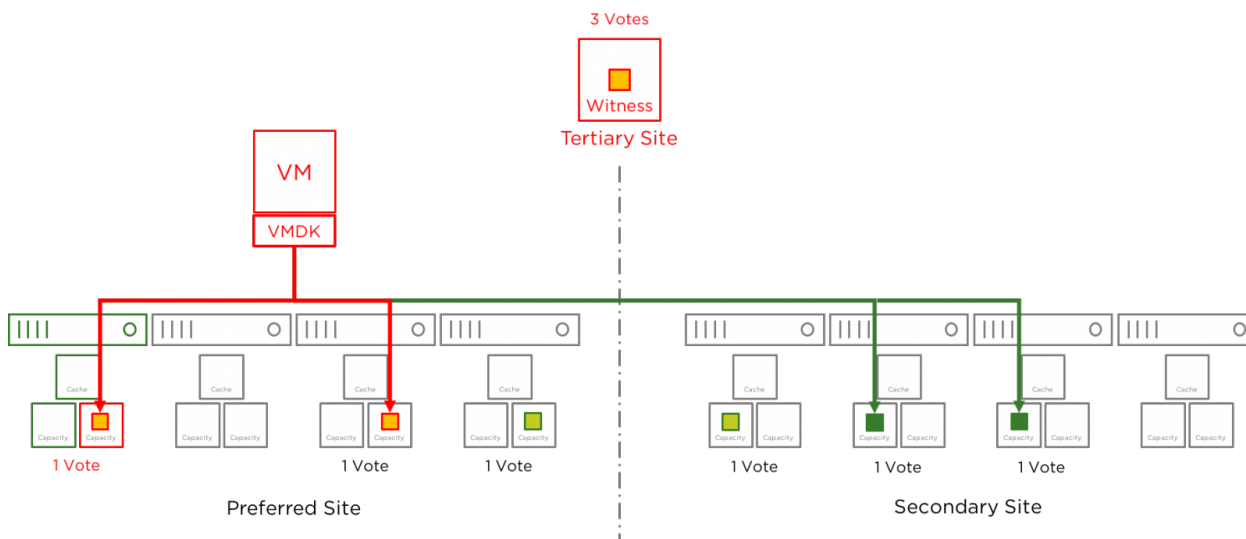


If a component in another location, on either site, regardless of having Local Protection enabled, were to fail, the vmdk would be inaccessible.

Below, the illustration shows the vSAN Witness Host offline, 1 failure in the Preferred Site, and 1 failure in the Secondary Site.



In this illustration, the vSAN Witness Host offline and 2 failures in the Preferred Site.



In both cases, only 44.4% of the votes are present (4 of 9), resulting in the virtual machine's vmdk being inaccessible.

It is important to understand that components, the weight of their votes, and their presence/absence determine the accessibility of the vSAN object.

In each failure case above, restoring access to the existing vSAN Witness would make the object accessible. Deploying a new vSAN Witness would not because the components would not be present.

Improved resilience for simultaneous site failures

Often, administrators need to place an entire stretched cluster site in maintenance mode for upgrades or planned tests. An entire site might also go down for unexpected reasons like power outages. When one of the sites is offline, the VMs will be running on the surviving site of the stretched cluster or the 2 Node cluster. During this time, the cluster becomes more vulnerable to additional site failures. In vSAN 7 Update 3, if the cluster experiences a witness site fault after one of the sites has already been deemed unavailable, vSAN will continue maintaining data availability for the workloads running inside the stretched cluster or the 2 Node cluster.

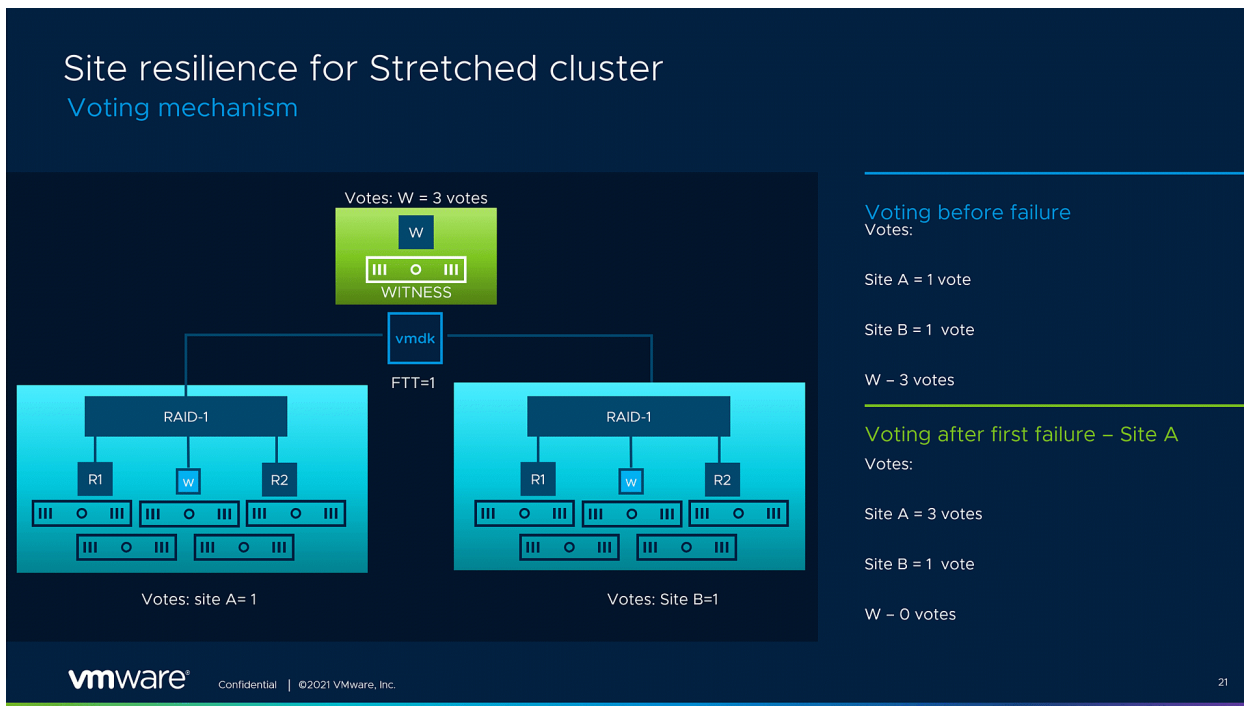
What are the changes in the voting mechanism?

In versions previous to vSAN 7 Update 3, if one of the sites becomes offline or inactive, and then the witness site goes down due to a planned or unplanned event, this results in the unavailability of the data residing on the remaining site due to lack of quorum.

This resilience enhancement has been achieved by modifying the voting mechanism in the cluster. These changes enable vSAN to assign most votes to the VM object replicas on the surviving site. These readjustments remove the dependency on the witness site. Thus, the remaining site objects can form a quorum immediately after the first site has been deemed offline, including being

offline in maintenance mode, and maintain resilience for the VMs running inside the stretched cluster even if the witness site goes down unexpectedly. The original voting mechanism is restored once all the hosts are back to operational.

The graphic below showcases a possible voting mechanism for a single VMDK object. For simplicity, we're going to take an example with an object that has a Site disaster tolerance =1 with RAID 1 applied. Note that these vote numbers are only a representation of the voting ratio. As a starting point, both sites' objects, on Site A and Site B, possess an equal number of votes, 1 vote each. The witness site object originally gets the most votes - 3 votes. As described, the adjustments in the voting mechanism will be triggered by a planned or unplanned site failure. After the fault event, most of the votes will be assigned to the remaining site, in this example, 3 votes will be assigned to Site A. Now, this site can form a quorum and maintain site resiliency for the VMs running inside the stretched cluster, even in case of an additional witness site failure.



Avoiding split-brain scenario

vSAN resolves the risk of a potential split-brain scenario by allowing the witness site to record the information about the new vote's layout. This awareness allows the witness site to reject any attempts for a sub-cluster creation with the originally failing Site B, because it knows Site B does not have enough quorum to form a cluster with the witness site.

New object status has been introduced in the UI to indicate the changed condition of the VM objects after the fault event:

- Without Improved Site resiliency - the object status is "Inaccessible."
- With Improved Site resiliency - the object status becomes "Reduced availability with no rebuild -daily timer."

A detailed demo of the feature can be found [here](#).

Enhancement specifics requirement and limitations:

- Remember that this feature is not applicable for VM objects whose policy consists of no replication level, meaning having a policy with PFTT = 0.
- The enhancement applies only to stretched clusters and 2 Node clusters topologies.
- vSAN will only adjust the votes if the VM object has lost all its replicas and no durability components are available for this object replica.
- This adjustment can only be performed when the witness failure occurs after the first site is pronounced offline.
- This feature improves availability for non-simultaneous planned/unplanned failures.
- No on-disk format change is required, but all hosts in the cluster must be updated with an object format version update.
- The enhancement behavior will remain the same whether it was a planned or unplanned event that provoked the unavailability of the sites.

Recovering from a Complete Site Failure

The descriptions of the host failures previously, although related to a single host failure, are also complete site failures. VMware has modified some of the vSAN behavior when a site failure occurs and subsequently recovers. In the event of a site failure, vSAN will now wait for additional time for “all” hosts to become ready on the failed site before it syncs components. The main reason is that if only some subset of the hosts come up on the recovering site, then vSAN will start the rebuild process. This may result in the transfer of data that already exists on the nodes that might become available at some point in time later on.

VMware recommends that when recovering from a failure, especially a site failure, all nodes in the site should be brought back online together to avoid costly resync and reconfiguration overheads. The reason behind this is that if vSAN bring nodes back up at approximately the same time, then it will only need to synchronize the data that was written between the time when the failure occurred and the when the site came back. If instead nodes are brought back up in a staggered fashion, objects might to be reconfigured and thus a significant higher amount of data will need to be transferred between sites.

How Read Locality is Established After Failover

A common question is how read locality is maintained when there is a failover. This guide has already described read locality and how in a typical vSAN deployment, a virtual machine reads equally from all of its replicas in a round-robin format. In other words, if a virtual machine has two replicas due to being configured to tolerate one failure, 50% of the reads come from each replica.

This algorithm has been enhanced for Stretched Clusters so that 100% of the reads comes from vSAN hosts on the local site, and the virtual machine does not read from the replica on the remote site. This avoids any latency that might be incurred by reading over the link to the remote site. The result of this behavior is that the data blocks for the virtual machine are also cached on the local site.

The virtual machine is restarted on the remote site during a failure or maintenance event. The 100% rule continues in the event of a failure. This means that the virtual machine will now read from the replica on the site to which it has failed over. One consideration is that there is no cached data on this site, so cache will need to warm for the virtual machine to achieve its previous performance levels.

When the virtual machine starts on the other site, either as part of a vMotion operation or a power on from vSphere HA restarting it, vSAN instantiates the in-memory state for all the objects of said virtual machine on the host where it moved. That includes the “owner” (coordinator) logic for each object. The owner checks if the cluster is set up in a “stretch cluster” mode and, if so, which fault domain it runs in. It then uses a different read protocol. Instead of the default round-robin protocol across replicas (at the granularity of 1MB), it sends 100% of the reads to the replica on the same site (but not necessarily the same host) as the virtual machine.

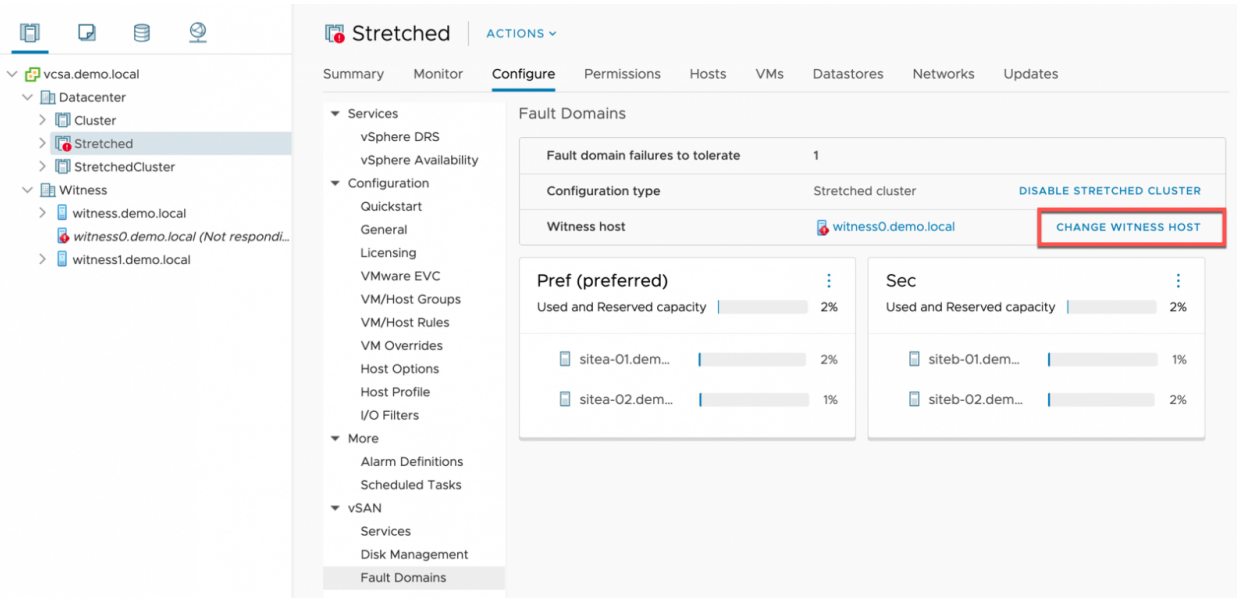
Replacing a Failed Witness Host

If a vSAN Witness Host fails in the vSAN Stretched Cluster, a new vSAN Witness Host can easily be replaced from the UI.

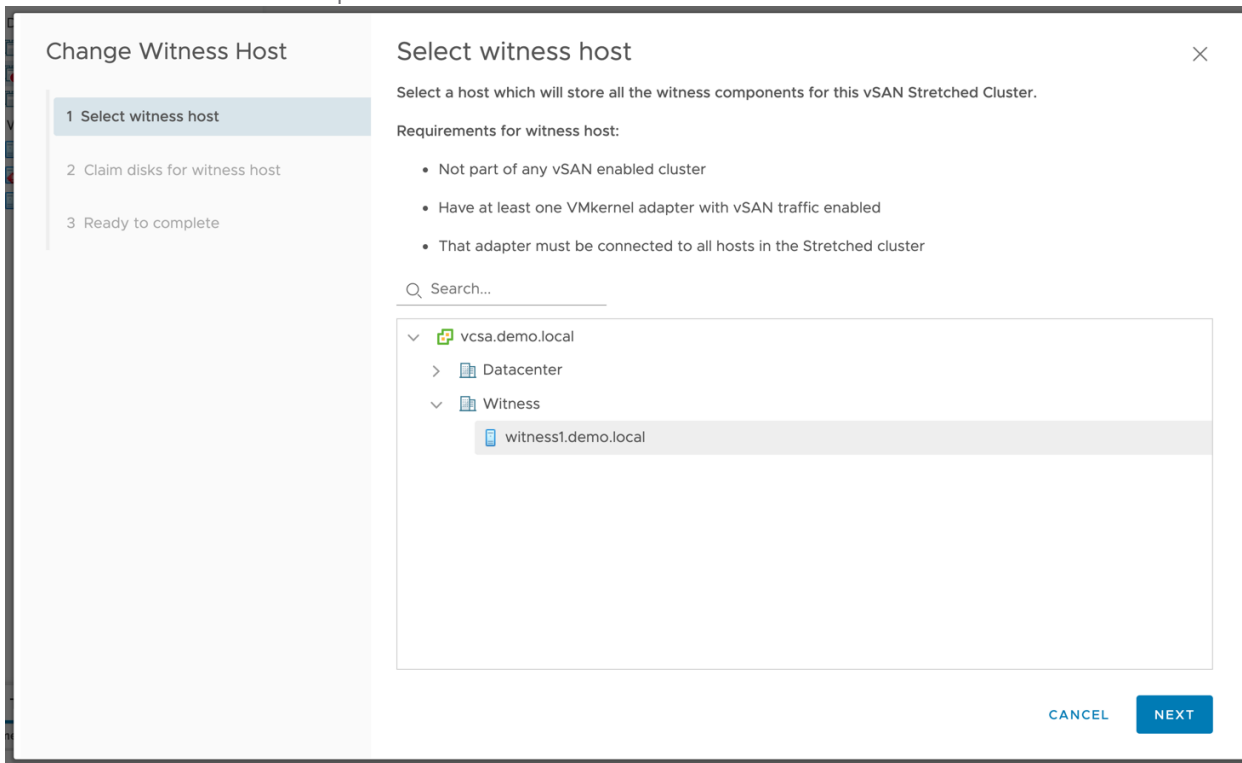
Changing the vSAN Witness Host

Navigate to Cluster > Configure > Fault Domains.

The illustration shows that witness0.demo.local is not responding.

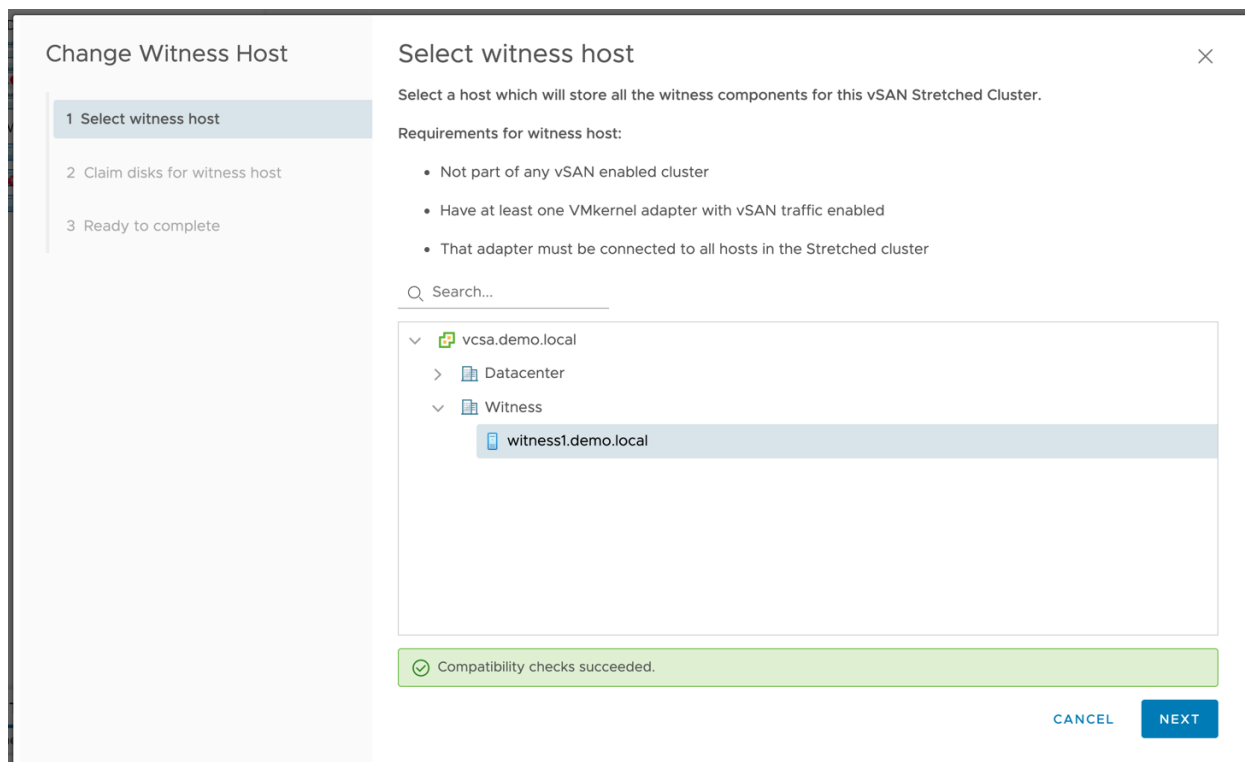


The CHANGE WITNESS HOST option will launch a wizard to select a new vSAN Witness Host.



The first illustration shows that witness1.demo.local has been deployed. The Change Witness Host wizard confirms that another cluster is not using it.

This wizard will not deploy a vSAN Witness Appliance, so a vSAN Witness Appliance will need to be deployed before invoking this wizard.



After selecting an available vSAN Witness Host, the disk group will need to be created.

When claiming disks for the vSAN Witness Host Disk Group, be certain that enough capacity is available for the number of components in the vSAN Stretched Cluster.

Profile Capacities:

- Tiny - 750 Components - 15GB Capacity Device
- Normal - 22,000 Components - 350GB Capacity Device
- Large - 45,000 Components - 3x350GB Capacity Devices

Change Witness Host

- 1 Select witness host
- 2 Claim disks for witness host
- 3 Ready to complete

Claim disks for witness host

Select disks on the witness host to be used for storing witness components.

First, select a single disk to serve as cache tier.

| <input type="radio"/> | Name | Drive Type | Capacity | Transport Type | Adapter |
|----------------------------------|-------------------------------|------------|----------|----------------|---------|
| <input checked="" type="radio"/> | Local VMware Disk (mpx.vmh... | Flash | 10.00 GB | | |
| <input type="radio"/> | Local VMware Disk (mpx.vmh... | Flash | 15.00 GB | | |

2 items

Then, select one or more disks to serve as capacity tier.

Capacity type: Flash

| <input checked="" type="checkbox"/> | Name | Drive Type | Capacity | Transport Type | Adapter |
|-------------------------------------|-------------------------------|------------|----------|----------------|---------|
| <input checked="" type="checkbox"/> | Local VMware Disk (mpx.vmh... | Flash | 15.00 GB | | |

1 item

CANCEL
BACK
NEXT

Note: the 10GB Flash Device will always be selected as the Cache Device.

Change Witness Host

- 1 Select witness host
- 2 Claim disks for witness host
- 3 Ready to complete

Ready to complete

Review your settings selections before finishing the wizard.

Witness host: witness1.demo.local

Claimed cache: 10.00 GB

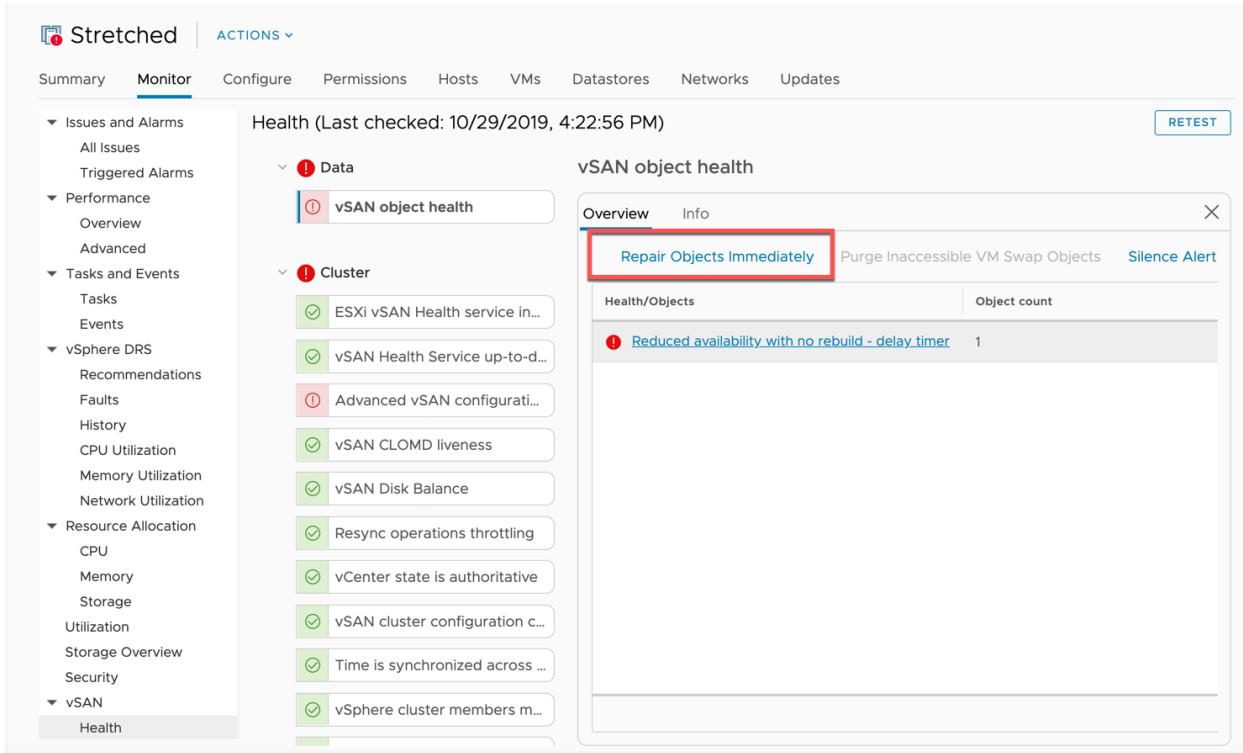
Claimed capacity: 15.00 GB

CANCEL
BACK
FINISH

On completion, verify that the vSAN Health Check failures have been resolved.

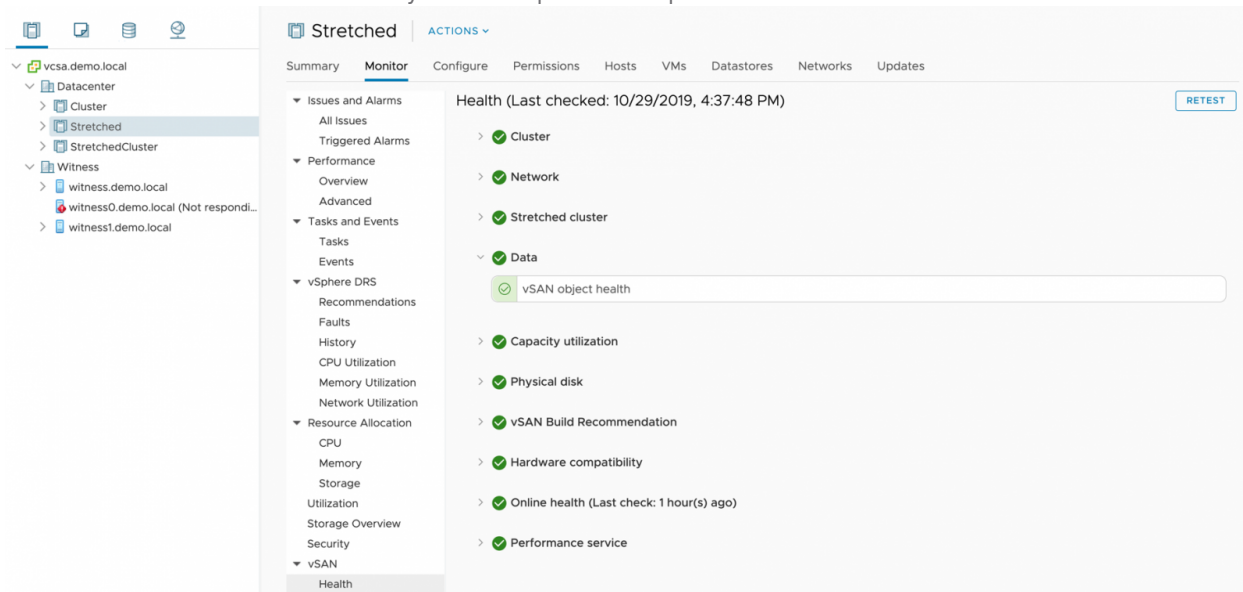
Note that the vSAN Object health test may continue to fail as the witness component of VM still remains "Absent" if the vSAN Witness Host is replaced before the CLOMD (Cluster Level Object Manager Daemon) timer expires (default of 60 minutes).

Run the vSAN Health Check to ensure there are no vSAN Object Health errors. If there are any errors, and the Repair Object Immediately option is available, perform it.



Witness components will be rebuilt on the new vSAN Witness Host.

Rerun the health check tests and they should all pass at this point.



VM Provisioning When a Site is Down

If a cluster fails, i.e., one of the sites is down, new virtual machines can still be provisioned. The provisioning wizard will warn the administrator that the virtual machine does not match its policy as follows:

New Virtual Machine

- ✓ 1 Select a creation type
- ✓ 2 Select a name and folder
- ✓ 3 Select a compute resource
- 4 Select storage**
- 5 Select compatibility
- 6 Select a guest OS
- 7 Customize hardware
- 8 Ready to complete

Select storage
Select the storage for the configuration and disk files

Encrypt this virtual machine (Requires Key Management Server)

VM Storage Policy: vSAN Default Storage Policy

| Name | Capacity | Provisioned | Free | Type |
|-------------------------------------|-----------|-------------|-----------|------|
| Storage Compatibility: Incompatible | | | | |
| NFS | 10.81 TB | 8.39 TB | 5.09 TB | |
| vsanDatastoreSC | 199.98 GB | 3.03 GB | 196.96 GB | |

Compatibility

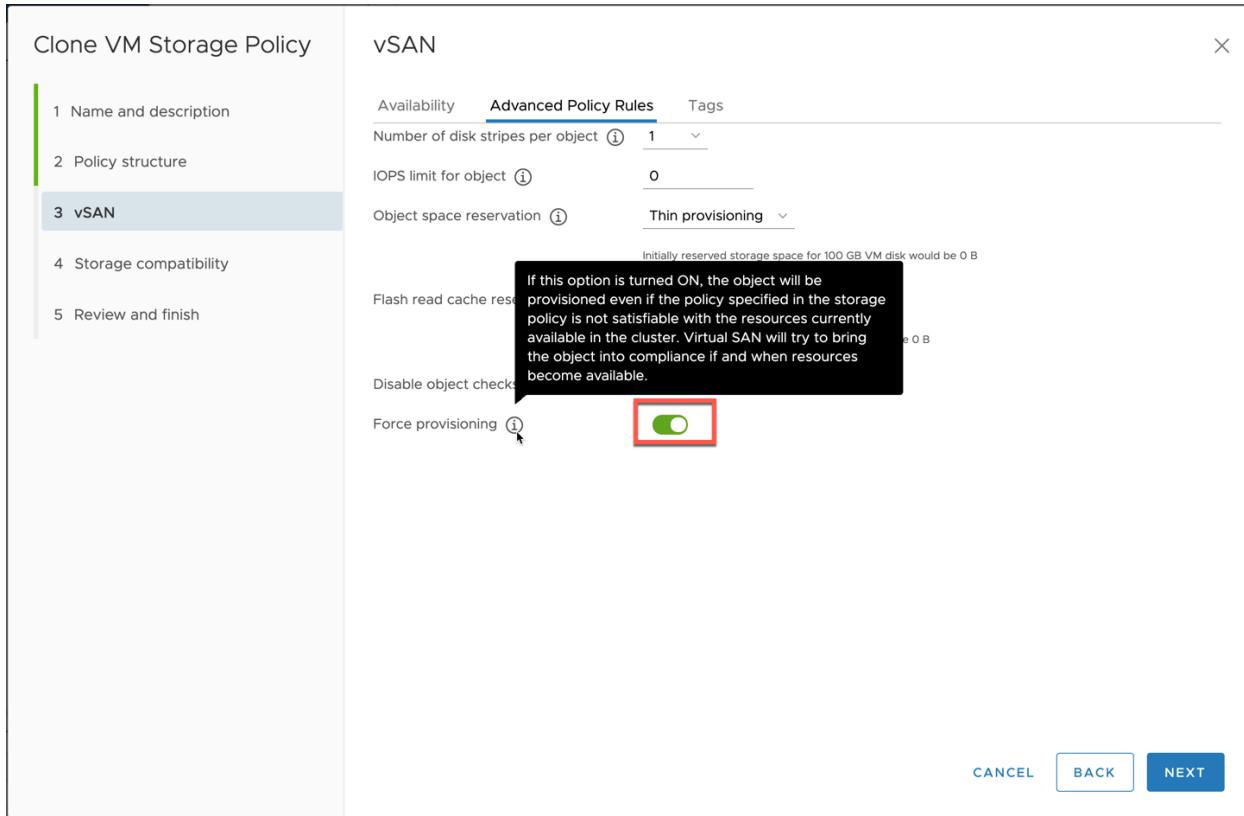
⚠ Datastore does not match current VM policy.

CANCEL BACK NEXT

In this case, when one site is down and there is a need to provision virtual machines, the ForceProvision capability is used to provision the VM.

If a policy that has the Force Provisioning Rule already exists, that policy may be used. If a policy does not exist with the Force Provisioning Rule, one will have to be created.

Cloning an existing policy, and selecting Force Provisioning in the Advanced Policy Rules will create a similar policy and will allow vSAN Objects to be provisioned when one site is offline.



Virtual Machines provisioned with a Storage Policy with the Force Provisioning Rule enabled will be deployed initially with no redundancy.

When the cluster connectivity & availability issues have been rectified, vSAN will automatically update the vSAN object configuration to meet the desired Storage Policy Rules. VMware recommends applying the originally desired Storage Policy to the vSAN Object after it meets the originally desired Storage Policy configuration.

Efficient inter-site resync for stretched clusters

Improvements were made to the resync mechanism in stretched clusters that further reduces the time from component failure to object compliance. One particular example is the use of a "proxy owner" host for objects that need to be resynced across sites following a failure.

Take the example of an object with the storage policy: PFTT=1, SFTT=1.

Let us assume that Site B (Secondary) has had a failure and that all components on that site need to be rebuilt, in previous versions of vSAN the resync would read the Primary site's data and synchronize it multiple times across the inter-site link (ISL), so if two components needed to be rebuilt on the Secondary site the data would be transmitted twice across the ISL raising the bandwidth utilization on the ISL for the duration of the resync.

Resyncs made to the remote site will now be copied once to a proxy host and then copied within that site to the other hosts from the proxy host. This lowers the traffic cost for resyncs across sites from one copy per component (could be multiple if using R5/R6) down to a single copy.

The above figure illustrates how components are rebuilt with the proxy owner improvement for inter-site resyncs. First, a copy is made to Site B, then the object is re-created within the site and finally, the data is copied from the proxy host to any other hosts within the site.

Failure Scenario Matrices

Hardware/Physical Failures

| Scenario | vSAN Behavior | Impact/Observed VMware HA Behavior |
|---|--|--|
| Cache disk failure | 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) | 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) | Disk marked as failed and all components present on it will rebuild on another disk. | VM will continue running. |
| Disk Group failure/offline | All components present on the Disk Group will rebuild on another Disk Group. | VM will continue running. |
| RAID/HBA card failure | 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 policies configured with PFTT=0 and Affinity=Secondary | No component reconfiguration across sites is required due to PFTT=0. In addition, as site affinity is set to secondary the VM will not be HA restarted on the preferred site, rather it will be allowed to run in place on the secondary site. This is because the object is in compliance with its policy and does not need site quorum to run. | VM continues running. |
| Dual site loss | Cluster offline. | VMs stop running. HA cannot restart VMs until quorum is re-established. |

Policy Implications

The below table is built based on the following example policy configuration:

PFTT = 1, SFTT = 2, FTM = R5/6

| Scenario | vSAN Behavior | Impact/Observed VMware HA Behavior |
|---|---|--|
| Single site failure (PFTT) | 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, host failure on one site (SFTT) | All components present will rebuild on their respective fault domain. | Disk and disk group 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, host failure on one site (SFTT) | 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 (PFTT) and single disk, disk group, host failure across remaining sites (SFTT) | 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 (PFTT) and dual disk, disk group, host failure across remaining sites (SFTT) | 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 (PFTT) and triple disk, disk group, host failure across remaining sites (SFTT) | 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 single PFTT failure (for example the witness) and the dual SFTT failure. The policy specifies "SFTT=2" which, during a PFTT 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 SFTT devices on remaining sites. | VMs will stop running and the cluster will be offline until a site is brought back online. Due to the PFTT and SFTT violations (single site failure and triple SFTT failure) the cluster's objects will have lost quorum and as such cannot run. |
| Dual site failure (PFTT) | 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 brought back online. |
| Single disk, disk group, host failure on one site (SFTT) and Dual disk, disk group, host failure on another site (SFTT) | 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 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, host failure on one site (SFTT) and Dual disk, disk group, host failure on another site (SFTT) | Cluster offline. A site will need to come back online to bring vSAN online. This results from the dual SFTT failure and the policy specifying "SFTT=2", after which the site is marked as failed. This can be achieved by replacing the failed SFTT devices on either site. | VMs will stop running, and the cluster will be offline until a site is back online. |
| 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. |

A particular case should be called out for policies configured with PFTT=0 and Affinity=Secondary

| Scenario | vSAN Behavior | Impact/Observed vSAN Behavior |
|--|--|-------------------------------|
| Sites are partitioned due to ISL loss. | No component reconfiguration across sites is required due to PFTT=0. In addition, as site affinity is set to secondary, the VM will not be HA restarted on the preferred site. Rather it will be allowed to run in place on the secondary site. The object complies with its policy and does not need a site quorum. | VM continues running. |

Additional Resources

Links to other vSAN resources

- [vSAN Documentation](#)
- [vSAN Resource Page](#)
- [vSAN Design Guide](#)
- [vSAN Proof Of Concept Guides](#)
- [vSAN Stretched Cluster Bandwidth Sizing Guidance](#)

Location of the vSAN Witness Appliance OVA

The vSAN Witness Appliance OVA is located on the Drivers & Tools tab of the vSAN download page. You will find a section called VMware vSAN tools, plug-ins, and appliances there. The "Stretch Cluster Witness VM OVA" is located here. The URL for vSAN 8.0 is:

https://customerconnect.vmware.com/downloads/info/slug/datacenter_cloud_infrastructure/vmware_vsan/8_0#drivers_tools

