



VMware® Virtual SAN™ 6.2 Stretched Cluster & 2 Node Guide

Jase McCarty
Storage and Availability Business Unit
VMware
v 6.2.0 / March 2016 / version 0.30

Contents

INTRODUCTION.....	5
SUPPORT STATEMENTS.....	7
VSPHERE VERSIONS.....	7
VSPHERE & VIRTUAL SAN.....	7
HYBRID AND ALL-FLASH SUPPORT.....	7
ON-DISK FORMATS.....	7
WITNESS HOST AS AN ESXI VM.....	8
FEATURES SUPPORTED ON VSAN BUT NOT VSAN STRETCHED CLUSTERS.....	8
FEATURES SUPPORTED ON VMSC BUT NOT VSAN STRETCHED CLUSTERS.....	9
NEW CONCEPTS IN VIRTUAL SAN - STRETCHED CLUSTER.....	10
VIRTUAL SAN STRETCHED CLUSTERS VERSUS FAULT DOMAINS.....	10
THE WITNESS HOST.....	10
READ LOCALITY IN VIRTUAL SAN STRETCHED CLUSTER.....	10
REQUIREMENTS.....	12
VMWARE VCENTER SERVER.....	12
A WITNESS HOST.....	12
NETWORKING AND LATENCY REQUIREMENTS.....	13
<i>Layer 2 and Layer 3 support.....</i>	<i>13</i>
<i>Supported geographical distances.....</i>	<i>13</i>
<i>Data site to data site network latency.....</i>	<i>13</i>
<i>Data site to data site bandwidth.....</i>	<i>13</i>
<i>Data Site to witness network latency.....</i>	<i>14</i>
<i>Data Site to witness network bandwidth.....</i>	<i>14</i>
<i>Inter-site MTU consistency.....</i>	<i>14</i>
CONFIGURATION MINIMUMS AND MAXIMUMS.....	15
VIRTUAL MACHINES PER HOST.....	15
HOSTS PER CLUSTER.....	15
WITNESS HOST.....	15
NUMBER OF FAILURES TO TOLERATE.....	15
FAULT DOMAINS.....	16
DESIGN CONSIDERATIONS.....	17
WITNESS HOST SIZING - COMPUTE.....	17
WITNESS HOST SIZING - MAGNETIC DISK.....	17
WITNESS HOST SIZING - FLASH DEVICE.....	17
CLUSTER COMPUTE RESOURCE UTILIZATION.....	18
NETWORKING DESIGN CONSIDERATIONS.....	19
<i>Connectivity.....</i>	<i>19</i>
<i>Type of networks.....</i>	<i>19</i>
<i>Considerations related to single default gateway on ESXi hosts.....</i>	<i>19</i>
<i>Caution when implementing static routes.....</i>	<i>20</i>
<i>Dedicated/Customer TCPIP stacks for VSAN Traffic.....</i>	<i>20</i>
<i>L2 design versus L3 design.....</i>	<i>21</i>
<i>Why not L3 between data sites?.....</i>	<i>22</i>
CONFIGURATION OF NETWORK FROM DATA SITES TO WITNESS HOST.....	23

<i>Option 1: Physical on-premises witness connected over L3 & static routes</i>	23
<i>Option 2: Virtual witness on-premises connected over L3 & static routes</i>	25
<i>Option 3: 2 Node configuration for Remote Office/Branch Office Deployment</i>	27
BANDWIDTH CALCULATION.....	29
<i>Requirements between Data Sites</i>	29
<i>Requirements when read locality is not available</i>	30
<i>Requirements between data sites and the witness site</i>	30
THE ROLE OF VIRTUAL SAN HEARTBEATS IN VIRTUAL SAN STRETCHED CLUSTER...	32
CLUSTER SETTINGS - VSPHERE HA	33
TURN ON VSPHERE HA.....	34
HOST MONITORING.....	34
ADMISSION CONTROL.....	35
HOST HARDWARE MONITORING - VM COMPONENT PROTECTION.....	36
DATASTORE FOR HEARTBEATING.....	37
<i>Why disable heartbeat datastores?</i>	37
VIRTUAL MACHINE RESPONSE FOR HOST ISOLATION.....	38
ADVANCED OPTIONS.....	39
<i>Network Isolation Response and Multiple Isolation Response Addresses</i>	39
CLUSTER SETTINGS - DRS	40
PARTIALLY AUTOMATED OR FULLY AUTOMATED DRS.....	41
VM/HOST GROUPS & RULES	42
HOST GROUPS.....	42
VM GROUPS.....	43
VM/HOST RULES.....	43
INSTALLATION	45
BEFORE YOU START.....	45
<i>What is a Preferred domain/preferred site?</i>	45
<i>What is read locality?</i>	45
<i>Witness host must not be part of the VSAN cluster</i>	46
VIRTUAL SAN HEALTH CHECK PLUGIN FOR STRETCHED CLUSTERS.....	47
<i>New Virtual SAN health checks for Stretched Cluster configurations</i>	48
USING A WITNESS APPLIANCE	49
PHYSICAL ESXi PREPARATION FOR WITNESS DEPLOYMENT.....	49
A NOTE ABOUT PROMISCUOUS MODE.....	49
SETUP STEP 1: DEPLOY THE WITNESS ESXi OVA.....	50
SETUP STEP 2: CONFIGURE WITNESS ESXi VM MANAGEMENT NETWORK.....	55
SETUP STEP 3: ADD WITNESS ESXi VM TO vCENTER SERVER.....	59
SETUP STEP 4: CONFIGURE VSAN NETWORK ON WITNESS HOST.....	64
SETUP STEP 5: IMPLEMENT STATIC ROUTES.....	68
CONFIGURING VIRTUAL SAN STRETCHED CLUSTER	70
CREATING A NEW VIRTUAL SAN STRETCHED CLUSTER.....	70
<i>Create Step 1: Create a Cluster</i>	70
<i>Create Step 2 Configure Virtual SAN as a stretched cluster</i>	71
<i>Create Step 3 Validate Network</i>	71
<i>Create Step 4 Claim Disks</i>	72
<i>Create Step 5 Create Fault Domains</i>	72
<i>Create Step 6 Select witness host</i>	73
<i>Create Step 7 Claim disks for witness host</i>	74

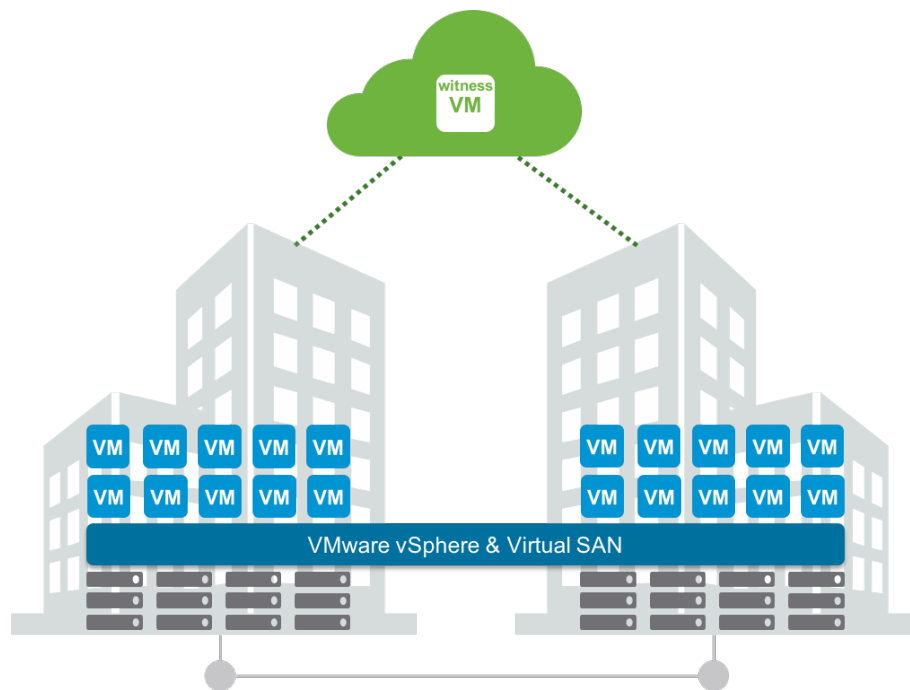
<i>Create Step 8 Complete</i>	74
CONVERTING AN EXISTING CLUSTER TO A STRETCHED CLUSTER.....	75
<i>Convert Step 1 Fault Domains & Stretched Cluster</i>	75
<i>Convert Step 2 Selecting hosts to participate</i>	75
<i>Convert Step 3 Configure fault domains</i>	76
<i>Convert Step 4 Select a witness host</i>	76
<i>Convert Step 5 Claim disks for witness host</i>	77
<i>Convert Step 6 Complete</i>	77
CONFIGURE STRETCHED CLUSTER SITE AFFINITY	79
<i>Configure Step 1 Create Host Groups</i>	79
<i>Configure Step 2: Create VM Groups</i>	81
<i>Configure Step 3: Create VM/Host Rules</i>	82
<i>Configure Step 4: Set vSphere HA rules</i>	83
VERIFYING VIRTUAL SAN STRETCHED CLUSTER COMPONENT LAYOUTS	84
UPGRADING A VIRTUAL SAN 6.1 STRETCHED CLUSTER TO VIRTUAL SAN 6.2.....	85
<i>Upgrading Step 1: Upgrade vCenter Server</i>	85
<i>Upgrading Step 2: Upgrade hosts in each site</i>	86
<i>Upgrading Step 3: Upgrade the witness appliance</i>	86
<i>Upgrading Step 4: Upgrade the on-disk format</i>	88
MANAGEMENT AND MAINTENANCE	89
MAINTENANCE MODE CONSIDERATION	89
<i>Maintenance mode on a site host</i>	89
<i>Maintenance mode on the witness host</i>	90
FAILURE SCENARIOS.....	91
HOW READ LOCALITY IS ESTABLISHED AFTER FAILOVER TO OTHER SITE?	92
SINGLE DATA HOST FAILURE - SECONDARY SITE.....	93
SINGLE DATA HOST FAILURE - PREFERRED SITE.....	95
SINGLE WITNESS HOST FAILURE - WITNESS SITE.....	97
NETWORK FAILURE - DATA SITE TO DATA SITE.....	99
<i>Data network test with multiple ESXi hosts per site</i>	101
<i>Data network test on host that contains virtual machine data only</i>	101
NETWORK FAILURE - DATA SITE TO WITNESS SITE	102
DISK FAILURE - DATA SITE HOST.....	103
DISK FAILURE - WITNESS HOST	103
VM PROVISIONING WHEN A SITES IS DOWN.....	103
REPLACING A FAILED WITNESS HOST	104
RECOVERING FROM A COMPLETE SITE FAILURE	106
APPENDIX A: ADDITIONAL RESOURCES.....	107
LOCATION OF THE WITNESS APPLIANCE OVA.....	107
APPENDIX B: CLI COMMANDS FOR VIRTUAL SAN STRETCHED CLUSTER	108
.....	108
ESXCLI	108
<i>esxcli vsan cluster preferredfaultdomain</i>	108
<i>esxcli vsan cluster unicastagent</i>	108
RVC - RUBY VSPHERE CONSOLE.....	109
<i>vsan.stretchedcluster.config_witness</i>	109
<i>vsan.stretchedcluster.remove_witness</i>	109
<i>vsan.stretchedcluster.witness_info</i>	109

Introduction

VMware Virtual SAN 6.1, shipping with vSphere 6.0 Update 1, introduced a new feature called VMware Virtual SAN Stretched Cluster. Virtual SAN Stretched Cluster is a specific configuration implemented in environments where disaster/downtime avoidance is a key requirement. This guide was developed to provide additional insight and information for installation, configuration and operation of a Virtual SAN Stretched Cluster infrastructure in conjunction with VMware vSphere. This guide will explain how vSphere handles specific failure scenarios and discuss various design considerations and operational procedures.

Virtual SAN Stretched Clusters with Witness Host refers to a deployment where a user sets up a Virtual SAN cluster with 2 active/active sites with an identical number of ESXi hosts distributed evenly between the two sites. The sites are connected via a high bandwidth/low latency link.

The third site hosting the Virtual SAN Witness Host is connected to both of the active/active data-sites. This connectivity can be via low bandwidth/high latency links.



Each site is configured as a Virtual SAN Fault Domain. The nomenclature used to describe a Virtual SAN 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 15+15+1 (31 nodes).

In Virtual SAN Stretched Clusters, there is only one witness host in any configuration.

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

Support Statements

vSphere versions

Virtual SAN Stretched Cluster configurations require vSphere 6.0 Update 1 (U1) or greater. This implies both vCenter Server 6.0 U1 and ESXi 6.0 U1. This version of vSphere includes Virtual SAN version 6.1. This is the minimum version required for Virtual SAN Stretched Cluster support.

vSphere & Virtual SAN

Virtual SAN version 6.1 introduced features including both All-Flash and Stretched Cluster functionality. There are no limitations on the edition of vSphere used for Virtual SAN. However, for Virtual SAN Stretched Cluster functionality, vSphere DRS is very desirable. DRS will provide initial placement assistance, and will also automatically migrate virtual machines to their correct site in accordance to Host/VM affinity rules. It can also help will locating virtual machines to their correct site when a site recovers after a failure. Otherwise the administrator will have to manually carry out these tasks. Note that DRS is only available in Enterprise edition and higher of vSphere.

Hybrid and All-Flash support

Virtual SAN Stretched Cluster is supported on both hybrid configurations (hosts with local storage comprised of both 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).

On-disk formats

VMware supports Virtual SAN Stretched Cluster with the v2 on-disk format only. The v1 on-disk format is based on VMFS and is the original on-disk format used for Virtual SAN. The v2 on-disk format is the version which comes by default with Virtual SAN version 6.x. Customers that upgraded from the original Virtual SAN 5.5 to Virtual SAN 6.0 may not have upgraded the on-disk format for v1 to v2, and are thus still using v1. VMware recommends upgrading the on-disk format to v2 for improved performance and scalability, as well as stretched cluster support. In Virtual SAN 6.2 clusters, the v3 on-disk format allows for additional features, discussed later, specific to 6.2.

Witness host as an ESXi VM

Both physical ESXi hosts and virtual ESXi hosts (nested ESXi) are supported for the witness host. VMware provides a Witness Appliance for those customers who wish to use the ESXi VM. A witness host/VM cannot be shared between multiple Virtual SAN Stretched Clusters.

Features supported on VSAN but not VSAN Stretched Clusters

The following are a list of products and features support on Virtual SAN but not on a stretched cluster implementation of Virtual SAN.

- SMP-FT, the new Fault Tolerant VM mechanism introduced in vSphere 6.0, is supported on standard VSAN 6.1 deployments, but it is not supported on stretched cluster VSAN deployments at this time. **The exception to this rule, is when using 2 Node configurations in the same physical location.*
- The maximum value for *NumberOfFailuresToTolerate* in a Virtual SAN Stretched Cluster configuration is 1. This is the limit due to the maximum number of Fault Domains being 3.
- In a Virtual SAN Stretched Cluster, there are only 3 Fault Domains. These are typically referred to as the Preferred, Secondary, and Witness Fault Domains. Standard Virtual SAN configurations can be comprised of up to 32 Fault Domains.
- The Erasure Coding feature introduced in Virtual SAN 6.2 requires 4 Fault Domains for RAID5 type protection and 6 Fault Domains for RAID6 type protection. Because Stretched Cluster configurations only have 3 Fault Domains, Erasure Coding is not supported on Stretched Clusters at this time.

Features supported on vMSC but not VSAN Stretched Clusters

The following are a list of products and features support on vSphere Metro Storage Cluster (vMSC) but not on a stretched cluster implementation of Virtual SAN.

- RR-FT, the original (and now deprecated) Fault Tolerant mechanism for virtual machines is supported on vSphere 5.5 for vMSC. It is not supported on stretched cluster Virtual SAN.
- Note that the new SMP-FT, introduced in vSphere 6.0 is not supported on either vMSC or stretched cluster VSAN, but does work on standard VSAN deployments.

New concepts in Virtual SAN - Stretched Cluster

Virtual SAN Stretched Clusters versus Fault Domains

A common question is how stretched cluster differs from Fault Domains, which is a Virtual SAN feature that was introduced with Virtual SAN version 6.0. Fault domains enable what might be termed “rack awareness” where the components of virtual machines could be distributed amongst multiple hosts in multiple racks, and should a rack failure event occur, the virtual machine would continue to be available. However, these racks would typically be hosted in the same data center, and if there was a data center wide event, fault domains would not be able to assist with virtual machines availability.

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

The witness host

The witness host is a dedicated ESXi host (or appliance) whose purpose is to host the witness component of virtual machines objects. The witness must have connection to both the master Virtual SAN node and the backup Virtual SAN node to join the cluster. In steady state operations, the master node resides in the “preferred site”; the backup node resides in the “secondary site”. Unless the witness host connects to both the master and the backup nodes, it will not join the Virtual SAN cluster.

Read locality in Virtual SAN Stretched Cluster

In traditional Virtual SAN 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 replica1 and 50% will come from replica2. In the case of a policy setting of `NumberOfFailuresToTolerate=2` in non-stretched Virtual SAN clusters, results in three copies of the data, 33% of the reads will come from replica1, 33% of the reads will come from replica2 and 33% will come from replica3.

In a Virtual SAN Stretched Cluster, we wish to avoid increased latency caused by reading across the inter-site link. To insure that 100% of reads, occur in the

site the VM resides on, the read locality mechanism was introduced. Read locality overrides the *NumberOfFailuresToTolerate=1* policy's behavior to distribute reads across the two data sites.

DOM, the Distributed Object Manager in Virtual SAN, takes care of this. DOM is responsible for the creation of virtual machine storage objects in the Virtual SAN cluster. It is also responsible for providing distributed data access paths to these objects. There is a single DOM owner per object. There are 3 roles within DOM; Client, Owner and Component Manager. The DOM Owner coordinates access to the object, including reads, locking and object configuration and reconfiguration. All objects changes and writes also go through the owner. The DOM owner of an object will now take into account which fault domain the owner runs in a Virtual SAN Stretched Cluster configuration, and will read from the replica that is in the same domain.

There is now another consideration with this read locality. One must avoid unnecessary vMotion of the virtual machine between sites. Since the read cache blocks are stored on one site, if the VM moves around freely and ends up on the remote site, the cache will be cold on that site after the move. Now there will be sub-optimal performance until the cache is warm again. To avoid this situation, soft affinity rules are used to 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.

Virtual SAN 6.2 introduced Client Cache, a mechanism that allocates 0.4% of host memory, up to 1GB, as an additional read cache tier. Virtual machines leverage the Client Cache of the host they are running on. Client Cache is not associated with Stretched Cluster read locality, and runs independently.

Requirements

In addition to Virtual SAN hosts, the following is a list of requirements for implementing Virtual SAN Stretched Cluster.

VMware vCenter server

A Virtual SAN Stretched Cluster configuration can be created and managed by a single instance of VMware vCenter Server. Both the Windows version and the Virtual Appliance version (Linux) are supported for configuration and management of a Virtual SAN Stretched Cluster.

A witness host

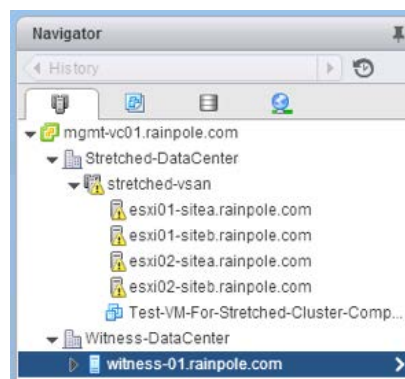
In a Virtual SAN Stretched Cluster, the witness components are only ever placed on the witness host. Either a physical ESXi host or a special witness appliance provided by VMware, can be used as the witness host.

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

It is important that witness host is not added to the VSAN cluster. The witness host is selected during the creation of a Virtual SAN Stretched Cluster.

The witness appliance will have a unique identifier in the vSphere web client UI to assist with identifying that a host is in fact 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 witness host is dedicated for each stretched cluster.



Networking and latency requirements

When Virtual SAN is deployed in a stretched cluster across multiple sites using fault domains, there are certain networking requirements that must be adhered to.

Layer 2 and Layer 3 support

Both Layer 2 (same subnet) and Layer 3 (routed) configurations are used in a recommended Virtual SAN Stretched Cluster deployment.

- **VMware recommends** that Virtual SAN communication between the data sites be over stretched L2.
- **VMware recommends** that Virtual SAN communication between the data sites and the witness site is routed over L3.

Note: A common question is whether L2 for Virtual SAN traffic across all sites is supported. There are some considerations with the use of a stretched L2 domain between the data sites and the witness site, and these are discussed in further detail in the design considerations section of this guide. Another common question is whether L3 for VSAN traffic across all sites is supported. While this can work, it is not the VMware recommended network topology for Virtual SAN Stretched Clusters at this time.

Virtual SAN traffic between data sites is **multicast**. Witness traffic between a data site and the witness site is **unicast**.

Supported geographical distances

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

Data site to data site network latency

Data site to data site network refers to the communication between non-witness sites, in other words, 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 **5msec** (< 2.5msec one-way).

Data site to data site bandwidth

Bandwidth between sites hosting virtual machine objects will be workload dependent. For most workloads, VMware recommends a minimum of 10Gbps

or greater bandwidth between sites. In use cases such as 2 Node configurations for Remote Office/Branch Office deployments, dedicated 1Gbps bandwidth can be sufficient with less than 10 Virtual Machines.

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

Data Site to witness network latency

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

In most Virtual SAN Stretched Cluster configurations, latency or RTT (Round Trip Time) between sites hosting VM objects and the witness nodes should not be greater than **200msec** (100msec one-way).

In typical 2 Node configurations, such as Remote Office/Branch Office deployments, this latency or RTT is supported up to **500msec** (250msec one-way).

The latency to the witness is dependent on the number of objects in the cluster. **VMware recommends** that on Virtual SAN Stretched Cluster configurations up to 10+10+1, a latency of less than or equal to 200 milliseconds is acceptable, although if possible, a latency of less than or equal to 100 milliseconds is preferred. For configurations that are greater than 10+10+1, **VMware recommends** a latency of less than or equal to 100 milliseconds is required.

Data Site to witness network bandwidth

Bandwidth between sites hosting VM objects and the witness nodes are dependent on the number of objects residing on Virtual SAN. It is important to size data site to witness bandwidth appropriately for both availability and growth. A standard rule of thumb is 2Mbps for every 1000 objects on Virtual SAN.

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

Inter-site MTU consistency

It is important to maintain a consistent MTU size between data nodes and the witness in a Stretched Cluster configuration. Ensuring that each VMkernel interface designated for Virtual SAN traffic, is set to the same MTU size will prevent traffic fragmentation. The Virtual SAN Health Check checks for a uniform MTU size across the Virtual SAN data network, and reports on any inconsistencies.

Configuration Minimums and Maximums

Virtual Machines per host

The maximum number of virtual machines per ESXi host is unaffected by the Virtual SAN Stretched Cluster configuration. The maximum is the same as for normal VSAN deployments.

VMware recommends that customers should run their hosts at 50% of maximum number of virtual machines supported in a standard Virtual SAN cluster to accommodate a full site failure. In the event of full site failures, the virtual machines on the failed site can be restarted on the hosts in the surviving site.

Hosts per cluster

The minimum number of hosts in a Virtual SAN Stretched Cluster is 3. In such a configuration, site 1 will contain a single ESXi host, site 2 will contain a single ESXi host and then there is a witness host at the third site, the witness site. The nomenclature for such a configuration is 1+1+1. This is commonly referred to as a 2 Node configuration.

The maximum number of hosts in a Virtual SAN Stretched Cluster is 31. Site 1 contains ESXi 15 hosts, site 2 contains 15 ESXi hosts, and the witness host on the third site makes 31. This is referred to as a 15+15+1 configuration.

Witness host

There is a maximum of 1 witness host per Virtual SAN Stretched Cluster. The witness host requirements are discussed in the design considerations section of this guide. VMware provides a fully supported witness virtual appliance, in Open Virtual Appliance (OVA) format, 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 on the third site.

Number Of Failures To Tolerate

Because Virtual SAN Stretched Cluster configurations effectively have 3 fault domains, the *NumberOfFailuresToTolerate* (FTT) policy setting, has a maximum of 1 for objects. Virtual SAN cannot comply with FTT values that are greater than 1 in a stretched cluster configuration.

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 important role in Virtual SAN Stretched Cluster. Similar to the *NumberOfFailuresToTolerate* (FTT) policy setting discussed previously, the maximum number of fault domains in a Virtual SAN Stretched Cluster is 3. The first FD is the “preferred” data site, the second FD is the “secondary” data site and the third FD is the witness host site.

Design Considerations

Witness host sizing - compute

When dealing with a physical server, the minimum ESXi host requirements will meet the needs of a witness host. The witness host must be capable of running the same version of ESXi as Virtual SAN data nodes.

When using a witness appliance (ESXi in a VM), the size is dependent on the configurations and this is decided during the deployment process. The witness appliance, irrespective of the configuration, uses at least two vCPUs. The physical host that the witness appliance runs on must be at least vSphere 5.5 or greater.

Witness host sizing - magnetic disk

The purpose of the witness host is to store witness components for virtual machine objects. Since a single magnetic disk supports approximately 21,000 components, and the maximum components supported on the witness host is 45,000, a minimum of 3 magnetic disks is required on the witness host if there is a need to support the maximum complement of components.

If using a physical ESXi host, a single physical disk can support a maximum of 21,000 components. Each witness component in a Virtual SAN stretch cluster requires 16MB storage. To support 21,000 components on a magnetic disk, **VMware recommends** a disk of approximately 350GB in size.

To accommodate the full 45,000 components on the witness host, **VMware recommends** 3 magnetic disks of approximately 350GB are needed, keeping the limit of 21,000 components per disk in mind.

If using the witness appliance instead of a physical ESXi host, there is no manual storage configuration required. Instead, the desired configuration size is chosen during deployment. Care will need to be taken that the underlying datastore for the VMDKs of the witness appliance supports the storage requirements. This will be highlighted in more detail during the installation section of the guide.

Witness host sizing - flash device

VMware recommends the flash device capacity (e.g. SSD) on the witness host should be approximately 10GB in size for the maximum number of 45,000 components is required. In the witness appliance, one of the VMDKs is tagged as a flash device. There is no requirement for an actual flash device.

Note that this witness host sizing is for component maximums. Smaller configurations that do not need the maximum number of components can run with fewer resources. Here are the three different sizes for the witness appliance.

Tiny (10 VMs or fewer)

- 2 vCPUs, 8 GB vRAM
- 8 GB ESXi Boot Disk, one 10 GB SSD, one 15 GB HDD
- Supports a maximum of 750 witness components

Normal (up to 500 VMs)

- 2 vCPUs, 16 GB vRAM
- 8 GB ESXi Boot Disk, one 10 GB SSD, one 350 GB HDD
- Supports a maximum of 22,000 witness components

Large (more than 500 VMs)

- 2 vCPUs, 32 GB vRAM
- 8 GB ESXi Boot Disk, one 10 GB SSD, three 350 GB HDDs
- Supports a maximum of 45,000 witness components

Note: When a physical host is used for the witness host, VMware will also support the tagging of magnetic disks as SSDs, implying that there is no need to purchase a flash device for physical witness hosts. This tagging can be done from the vSphere web client UI.

Cluster Compute resource utilization

For full availability, **VMware recommends** that customers should be running at 50% of resource consumption across the Virtual SAN Stretched Cluster. In the event of a complete site failure, all of the virtual machines could be run on the surviving site (aka fault domain)

VMware understands that some customers will want to run close to 80% and even 100% of resource utilization because they do not want to dedicate resources just to protect themselves against a full site failure since site failures are very rare. In these cases, customer should understand that not all virtual machines will be restarted on the surviving site.

Networking Design Considerations

A Virtual SAN Stretched Cluster requires 3 sites; the first site will maintain the first copy of the virtual machine data (data site 1), the second site will maintain the second copy of the virtual machine data (data site 2) and the third site will maintain the witness component(s). The three sites all need to communicate, both at the management network level and at the VSAN network level. There also need to be a common virtual machine network between the data sites. To summarize, the following are the network requirements for a Virtual SAN Stretched Cluster:

Connectivity

- Management network: connectivity to all 3 sites
- VM network: connectivity between the data sites (the witness will not run virtual machines that are deployed on the VSAN cluster)
- vMotion network: connectivity between the data sites (virtual machines will never be migrated from a data host to the witness host)
- Virtual SAN network: connectivity to all 3 sites

Type of networks

VMware recommends the following network types for Virtual SAN Stretched Cluster:

- Management network: L2 stretched or L3 (routed) between all sites. Either option should both work fine. The choice is left up to the customer.
- VM network: **VMware recommends** L2 stretched between data sites. In the event of a failure, the VMs will not require a new IP to work on the remote site
- vMotion network: L2 stretched or L3 (routed) between data sites should both work fine. The choice is left up to the customer.
- Virtual SAN network: **VMware recommends** L2 stretched between the two data sites and L3 (routed) network between the data sites and the witness site. L3 support for the Virtual SAN network was introduced in VSAN 6.0.

Considerations related to single default gateway on ESXi hosts

The major consideration with implementing this configuration is that each ESXi host comes with a default TCP/IP stack, and as a result, only has a single default gateway. The default route is typically associated with the management network TCP/IP stack. Now consider the situation where, for isolation and security reasons, the management network and the Virtual SAN network are completely isolated from one another. The management network

might be using vmk0 on physical NIC 0, and the VSAN network might be using vmk2 on physical NIC 1, i.e. completely distinct network adapters and two distinct TCPIP stacks. This implies that the Virtual SAN network has no default gateway.

Consider also that the Virtual SAN network is stretched over data site 1 and 2 on an L2 broadcast domain, e.g. 172.10.0.0 and the witness on site 3 the VSAN network is on another broadcast domain, e.g. 172.30.0.0. If the VMkernel adapters on the VSAN network on data site 1 or 2 tries to initiate a connection to the VSAN network on the witness site (site 3), and since there is only one default gateway associated with the management network, the connection will fail. This is because the traffic will be routed through the default gateway on the ESXi host, and thus the management network on the ESXi host, and there is no route from the management network to the VSAN network.

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 Virtual SAN network on a Virtual SAN Stretched Cluster, static routes could be added as follows, using the above example IP addresses:

1. Hosts on data site 1 have a static route added so that requests to reach the 172.30.0.0 witness network on site 3 are routed via the 172.10.0.0 interface
2. Hosts on data site 2 have a static route added so that requests to reach the 172.30.0.0 witness network on site 3 are routed via the 172.10.0.0 interface
3. The witness host on site 3 has a static route added so that requests to reach the 172.10.0.0 data site 1 and data site 2 network are routed via the 172.30.0.0 interface

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

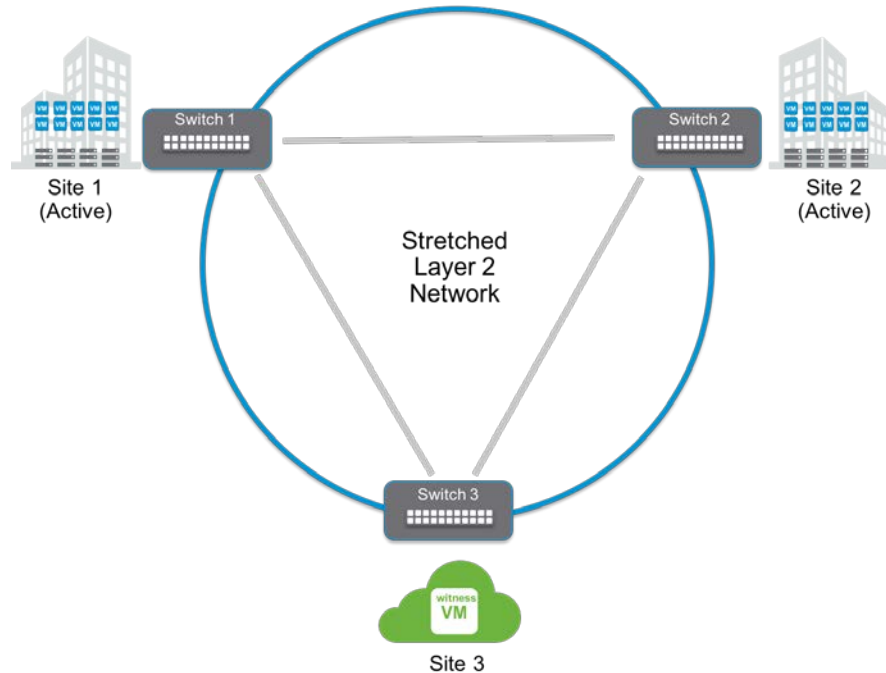
Using static routes requires administrator intervention. Any new ESXi hosts that are added to the cluster at either site 1 or site 2 needed to have static routes manually added before they can successfully communicate to the witness, and the other data site. Any replacement of the witness host will also require the static routes to be updated to facilitate communication to the data sites.

Dedicated/Customer TCPIP stacks for VSAN Traffic

At this time, the Virtual SAN traffic does not have its own dedicated TCPIP stack. Custom TCPIP stacks are also not applicable for Virtual SAN traffic.

L2 design versus L3 design

Consider a design where the Virtual SAN Stretched Cluster is configured in one large L2 design as follows, where Site 1 and Site 2 are where the virtual machines are deployed. The Witness site contains the witness host:

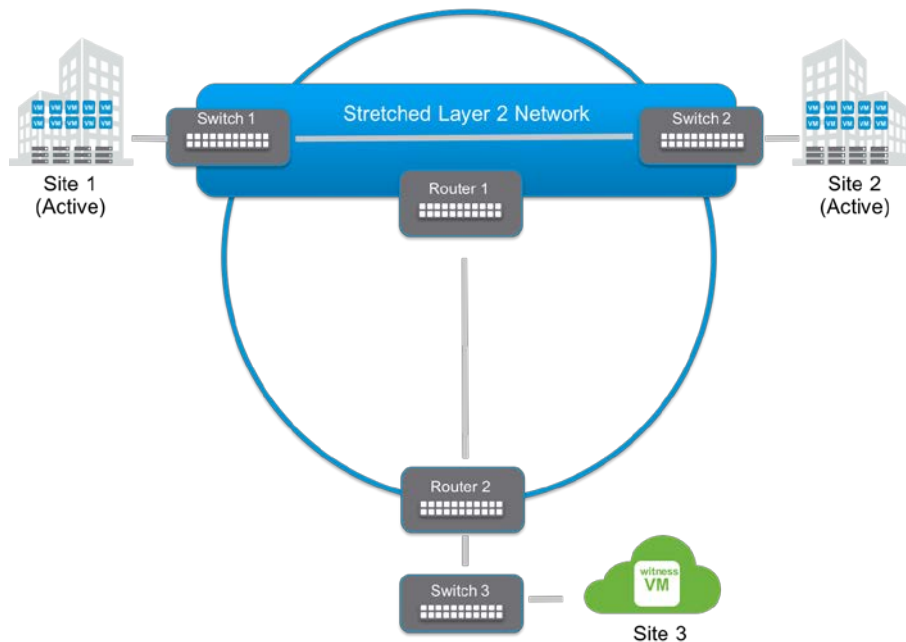


In the event of the link between Switch 1 and Switch 2 is broken (the link between the Site 1 and Site 2). Network traffic will now route from Site 1 to Site 2 via Site 3. Considering VMware will support a much lower bandwidth for the witness host, customers may see a decrease in performance if network 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 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 the majority of customers.

To avoid the situation previously 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.
- Between Site 1 and Witness Site 3, implement a L3 (routed) configuration.
- Between Site 2 and Witness Site 3, implement a L3 (routed) configuration.
- In the event of a failure on either of the data sites network, this configuration will prevent any traffic from Site 1 being routed to Site 2 via Witness Site 3, and thus avoid any performance degradation.



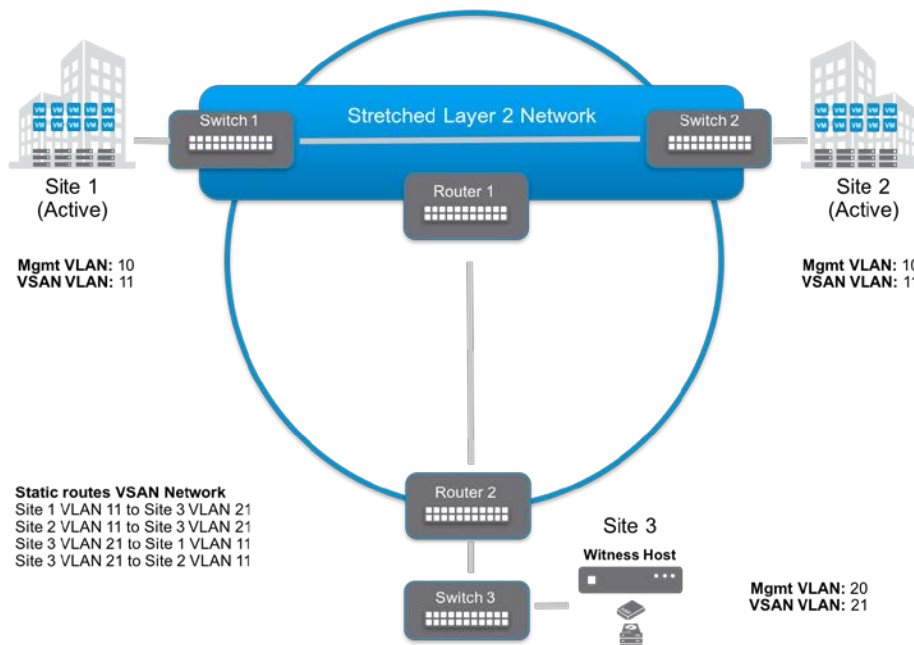
Why not L3 between data sites?

It is also important to consider that having different subnets at the data sites is going to be painful for any virtual machines that failover to the other site since there is no easy, automated way to re-IP the guest OS to the network on the other data site.

Configuration of network from data sites to witness host

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 on-premises witness 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 the Site 3. In order for the Virtual SAN 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 Virtual SAN traffic from Site 1 and Site 2 is able to 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 some 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 Virtual SAN 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 Virtual SAN interface must have a static route added which redirects Virtual SAN traffic destined for the data sites' hosts. Adding static routes is achieved using the `esxcfg-route -a` command on the ESXi hosts. This will have to be repeated on all ESXi hosts in the stretched cluster.

For this to work, the network switches need to be IP routing enabled between the Virtual SAN 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 Virtual SAN Stretched Cluster to work properly.

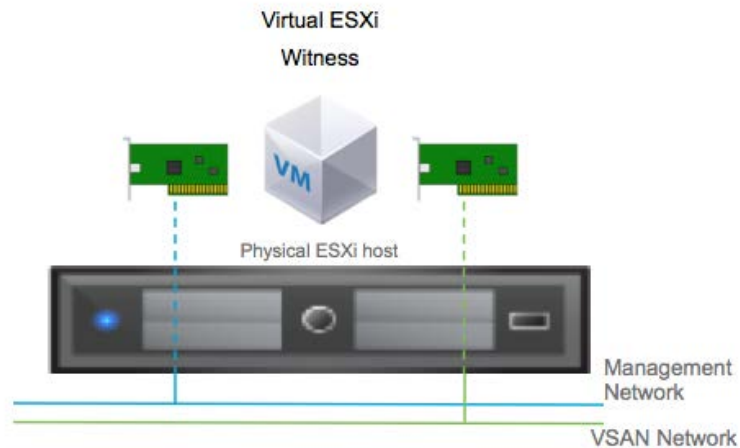
Note that we have not mentioned the ESXi management network here. The vCenter server will still be required to manage both the ESXi hosts at the data sites and the ESXi witness. In many cases, this is not an issue for customer. However, in the case of stretched clusters, it might be necessary to add 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 (without NAT'ing), 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 Virtual SAN Stretched Cluster. This is because there will never be a migration or deployment of virtual machines to the Virtual SAN witness. Its purpose is to maintain witness objects only, and does not require either of these networks for this task.

Option 2: Virtual witness on-premises 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 pre-configured. 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.

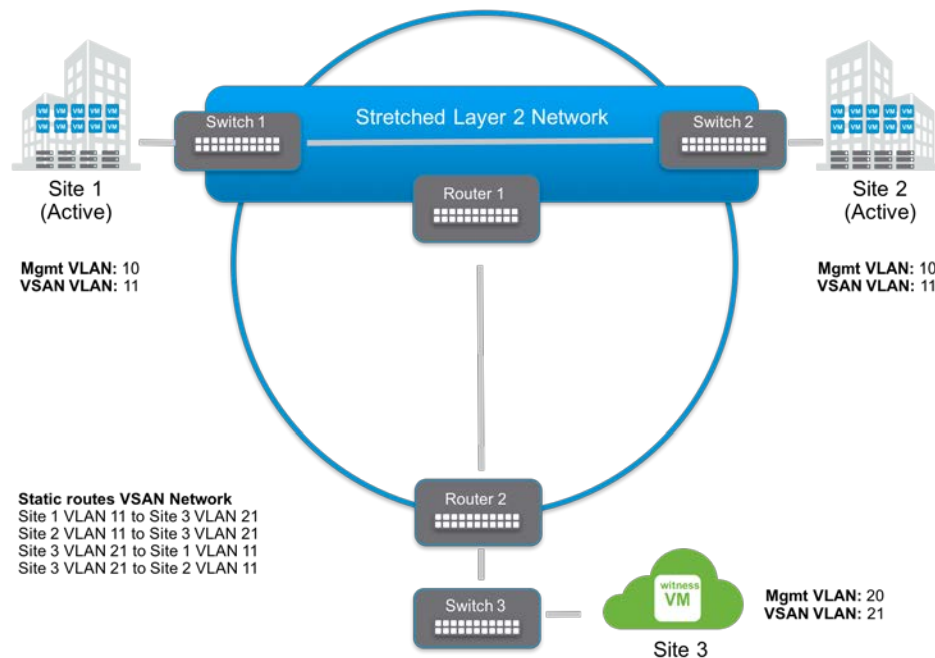


Once the virtual ESXi witness has been successfully deployed, the static route configuration must be configured.

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, 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 Virtual SAN 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 Virtual SAN 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 Virtual SAN Stretched Cluster will form.

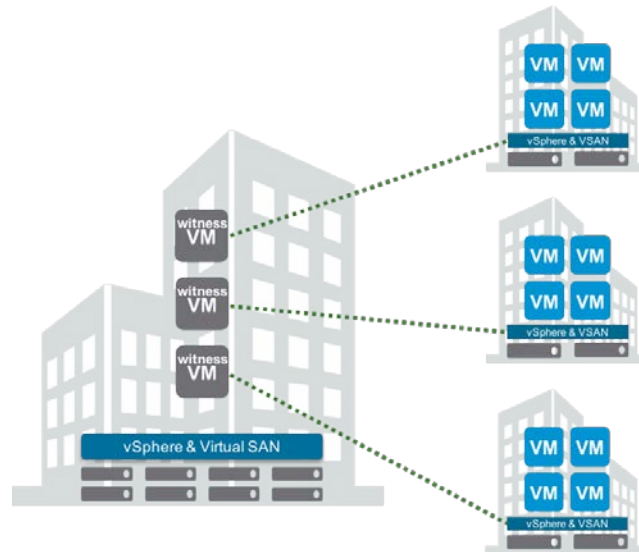


Note that once again we have not mentioned the management network here. As mentioned before, vCenter needs to manage the remote ESXi witness and the hosts on the data sites. If necessary, a static route should be added to the vCenter server to reach the management network of the witness ESXi host, and similarly a static route should be added to the ESXi witness to reach the vCenter server.

Also note that, as before, 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 Virtual SAN Stretched Cluster. This is because there will never be a migration or deployment of virtual machines to the VSAN witness. Its purpose is to maintain witness objects only, and does not require either of these networks for this task.

Option 3: 2 Node configuration for Remote Office/Branch Office Deployment

In the use case of Remote Office/Branch Office (ROBO) deployments, it is common to have 2 Node configurations at one or more remote offices. This deployment model can be very cost competitive when a running a limited number of virtual machines no longer require 3 nodes for Virtual SAN.

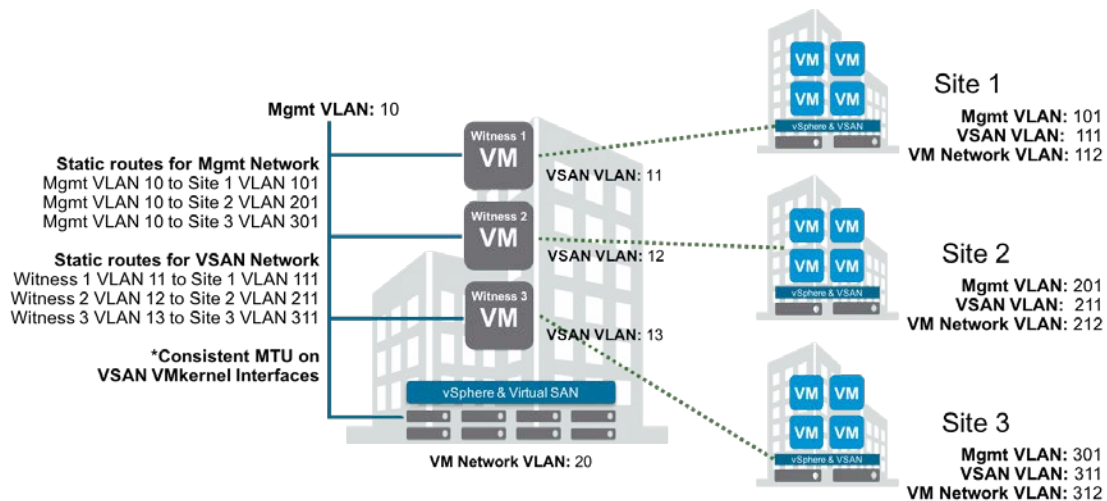


Virtual SAN 2 Node configurations are Virtual SAN Stretched Clusters comprised of two data nodes and one witness node. This is a 1+1+1 Stretched Cluster configuration. Each data node behaves as a data site, and the two nodes are typically in the same location. The witness VM could reside at the primary datacenter or another location.

Management traffic for the data nodes is typically automatically routed to the vCenter server at the central datacenter. Routing for the VSAN network, as shown in previous scenarios, will require static routes between the VSAN interfaces on each data node and the witness VM running in the central datacenter.

Because they reside in the same physical location, networking between data nodes is consistent with that of a traditional Virtual SAN cluster. Data nodes still require a static route to the Witness VM residing in the central datacenter. The witness VM's secondary interface, designated for Virtual SAN traffic will also require a static route to each of data node's VSAN traffic enabled VMkernel interface.

Adding static routes is achieved using the `esxcfg-route -a` command on the ESXi hosts and witness VM.



In the illustration above, the central datacenter management network is on VLAN 10. For vCenter to manage each of the 2 node (ROBO) deployments, there must be a route to each host's management network. This could be on an isolated management VLAN, but it is not required. Depending on the network configuration, vCenter itself may require static routes to each of the remote ESXi host management VMkernel interfaces. All the normal requirements for vCenter to connect to ESXi hosts should be satisfied.

The management VMkernel for the witness VM, in the central datacenter, can easily reside on the same management VLAN in the central datacenter, not requiring any static routing.

The VSAN network in each site must also have routing to the respective witness VM VSAN interface. Because the VMkernel interface with VSAN traffic enabled uses the same gateway, static routes will be required to and from the data nodes to the witness VMs. Remember the witness VM will never run an VM workloads, and therefore the only traffic requirements are for management and VSAN witness traffic, because its purpose is to maintain witness objects only.

For remote site VMs to communicate with central datacenter VMs, appropriate routing for the VM Network will also be required.

Bandwidth calculation

As stated in the requirements section, the bandwidth requirement between the two main sites is dependent on workload and in particular the number of write operations per ESXi host. Other factors 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, may also need to be factored in.

Requirements between Data Sites.

Reads are not included in the calculation as we are assuming read locality, which means that 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 is comprised of overhead for Virtual SAN 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 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 Virtual SAN 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 Virtual SAN 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 Virtual SAN 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 MB/s, or 2.24Gbps bandwidth)

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

$$B = 280 \text{ Mbps} * 1.4 * 1.25 = 3,920 \text{ Mbps or } 3.92\text{Gbps}$$

Using the above formula, a Virtual SAN Stretched Cluster with a dedicated 10Gbps inter-site link, can accommodate approximately 170,000 4KB write IOPS. Customers will need to 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.

While it might be possible to use 1Gbps connectivity for very small Virtual SAN Stretched Cluster implementations, the majority of implementations will require 10Gbps connectivity between sites. Therefore, **VMware recommends a minimum of 10Gbps network connectivity between sites for optimal performance and for possible future expansion of the cluster.**

Requirements when read locality is not available.

Note that the previous calculations are only for regular Stretched Cluster traffic with read locality. If there is a device failure, read operations also have to 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 on top of 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, or 320 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 bandwidth between data sites. Because hosts designated as a witness do not maintain any VM data, but rather only component metadata, the requirements are much smaller.

Virtual Machines on Virtual SAN are comprised of many objects, which can potentially be split into multiple components, depending on factors like policy and size. The number of components on Virtual SAN have a direct impact on 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 that occur when the Preferred Site goes offline, and the Secondary Site takes ownership of all of the components.

When the primary site goes offline, the secondary site becomes the master. The Witness sends updates to the new master, followed by the new master 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, as well ownership of all of the components within 5 seconds.

Witness to Site Examples

Workload 1

With a VM being comprised of

- 3 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 Virtual SAN, 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 rule of thumb can be stated that for every 1,000 components, 2 Mbps is appropriate.

Workload 2

With a VM being comprised of

- 3 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 Virtual SAN, the resulting calculation is:

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

$$B + 10\% = 32.78 \text{ Mbps} + 3.28 \text{ Mbps} = 36.05 \text{ Mbps}$$

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 36Mbps.

The role of Virtual SAN heartbeats in Virtual SAN Stretched Cluster

As mentioned previously, when VSAN is deployed in a stretched cluster configuration, the VSAN master node is placed on the preferred site and the VSAN backup node is placed on the secondary site. So long as there are nodes (ESXi hosts) available in the “preferred” site, then a master is always selected from one of the nodes on this site. Similarly, for the “secondary” site, so long as there are nodes available on the secondary site.

The VSAN master node and the VSAN backup node send heartbeats every second. If communication is lost for 5 consecutive heartbeats (5 seconds) between the master and the backup due to an issue with the backup node, the master 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 there is a complete site failure, the master selects a backup node from the “preferred” site.

A similar scenario arises when the master has a failure.

When a node rejoins an empty site after a complete site failure, either the master (in the case of the node joining the primary site) or the backup (in the case where the node is joining the secondary site) will migrate to that site.

If communication is lost for 5 consecutive heartbeats (5 seconds) between the master and the witness, the witness is deemed to have failed. If the witness has suffered a permanent failure, a new witness host can be configured and added to the cluster.

Cluster Settings – vSphere HA

Certain **vSphere HA** behaviors have been modified especially for Virtual SAN. It checks the state of the virtual machines on a per virtual machine basis. vSphere HA can make a decision 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 Virtual SAN Stretched Cluster, VMware recommends the following:

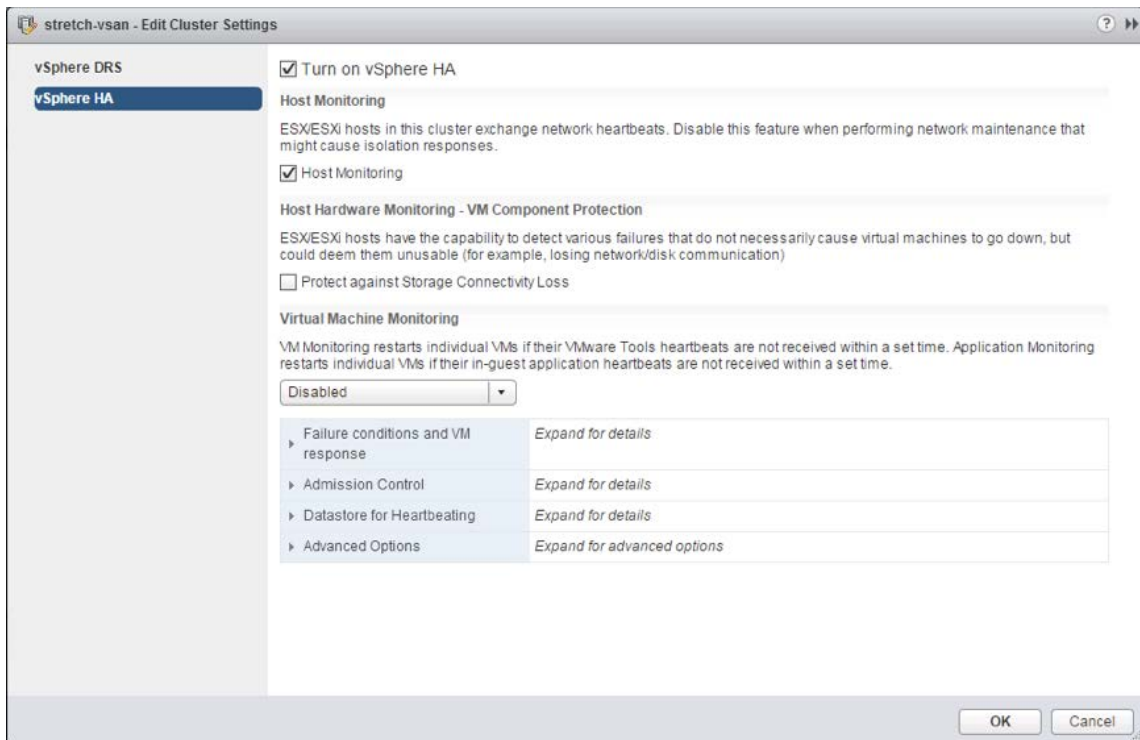
vSphere HA	Turn on
Host Monitoring	Enabled
Host Hardware Monitoring – VM Component Protection: “Protect against Storage Connectivity Loss”	Disabled (default)
Virtual Machine Monitoring	Customer Preference – Disabled by default
Admission Control	Set to 50%
Host Isolation Response	Power off and restart 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

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 check box.



Host Monitoring

Host monitoring should be enabled on Virtual SAN 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.

Host Monitoring

ESX/ESXi hosts in this cluster exchange network heartbeats. Disable this feature when performing network maintenance that might cause isolation responses.

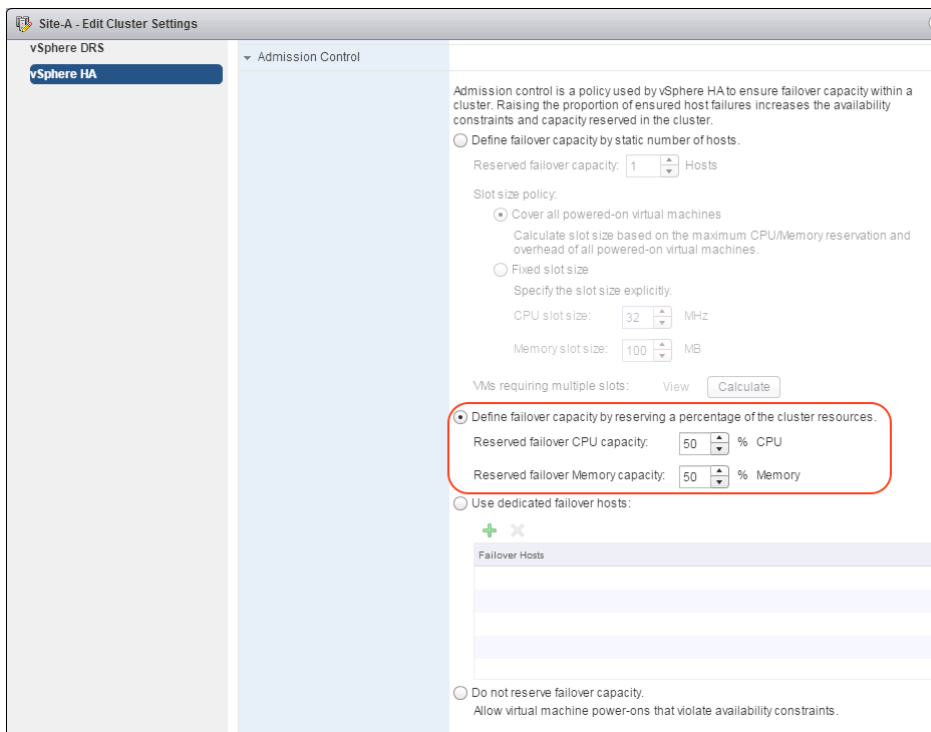
Host Monitoring

Admission Control

Admission control ensures that HA has sufficient resources available to restart virtual machines after a failure. As a full site failure is one scenario that needs to be taken into account 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 full site failure. Since ESXi hosts will be equally divided across both sites in a Virtual SAN Stretched Cluster, and to ensure that all workloads can be restarted by vSphere HA, **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 6.0 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%.



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.

Host Hardware Monitoring – VM Component Protection

vSphere 6.0 introduces 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 PDL, permanent device loss condition, is a condition that is communicated by the storage controller to ESXi host via a SCSI sense code. This condition indicates that a disk device has become unavailable and is likely permanently unavailable. When it is not possible for the storage controller to communicate back the status to the ESXi host, then 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, in the event of an APD/PDL occurring 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.

Note that 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 Virtual SAN, and thus cannot guarantee that the virtual machine will be able to restart elsewhere after termination. For example, there may be resources available to restart the virtual machine, but accessibility to the virtual machine by the remaining hosts in the cluster is not known to HA. For traditional datastores, this is not a problem, 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 or not.

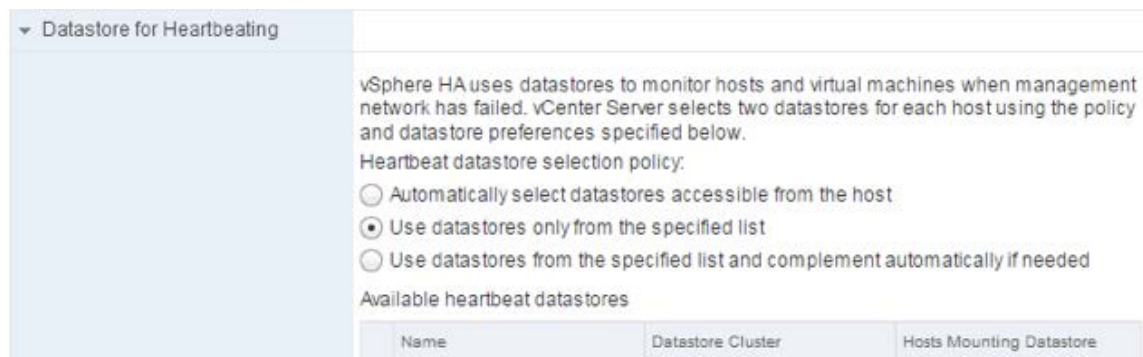
At the moment, it is not possible for vSphere HA to understand the complete inaccessibility vs. partial inaccessibility on a per virtual machine basis on Virtual SAN; hence the lack of VMCP support by HA for VSAN.

VMware recommends leaving VM Component Protection (VMCP) **disabled**.

Datstore 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. For Virtual SAN configurations, including stretched cluster configurations, this functionality should be disabled. If there are other datastores available to the ESXi hosts (i.e. NFS or VMFS datastores), then heartbeat datastores should be disabled. If there are no additional datastores available to the ESXi hosts other than the VSAN Datastore, then this step isn't necessary.

To disable datastore heartbeating, under HA settings, open the Datastore for Heartbeating section. Select the option "Use datastore from only the specified list", and ensure that there are no datastores 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`.

Why disable heartbeat datastores?

If you have a heartbeat datastore and only the VSAN traffic network fails, vSphere HA does not restart the virtual machines on another host in the cluster. When you restore the link, the virtual machines will continue to run. If virtual machine availability is your utmost concern, keeping in mind that a virtual machine restart is necessary in the event of a host isolation event, then you should not setup a heartbeat datastore. Any time the VSAN network causes a host to get isolated, vSphere HA will power on the virtual machine on another host in the cluster.

Of course, with a restart the in-memory state of the apps is lost, but the virtual machine has minimal downtime. If you do not want a virtual machine to fail over when there is a VSAN traffic network glitch, then a heartbeat datastore

should be configured. Of course, you will need other non-VSAN datastores to achieve this.

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 is able to reach the isolation response IP address. There is a chance that communication could be completely lost to the host in question, so requests to shutdown the virtual machines may not succeed. Therefore, **VMware recommends** that the Response for Host Isolation is to *Power off and restart VMs*.

The screenshot shows the 'standard - Edit Cluster Settings' window with the 'vSphere HA' tab selected. The 'Failure conditions and VM response' section is expanded, showing a table of failure conditions and their corresponding responses. The 'Response for Host Isolation' is highlighted with a red circle and set to 'Power off and restart VMs'.

Failure	Response	Details
Host failure	Restart VMs	Restart VMs using VM restart priority ordering.
Host Isolation	Power off and restart VMs	VMs on isolated hosts will be powered off and restarted on available hosts.
Datastore with Permanent Device Loss	Disabled	Datastore protection for All Paths Down and Permanent Device Loss is disabled.
Datastore with All Paths Down	Disabled	Datastore protection for All Paths Down and Permanent Device Loss is disabled.
Guest not heartbeating	Disabled	VM and application monitoring disabled.

VM restart priority: Medium

Response for Host Isolation: Power off and restart VMs

Response for Datastore with Permanent Device Loss (PDL): Disabled

Response for Datastore with All Paths Down (APD): Disabled

Delay for VM failover for APD: 3 minutes

Response for APD recovery after APD timeout: Disabled

Advanced Options

When vSphere HA is enabled on a VSAN Cluster, uses a heart beat mechanisms to validate the state of an ESXi host. Network heart beating is the primary mechanism for HA to validate availability of the hosts.

If a host is not receiving any heartbeats, it uses a failsafe mechanism to detect if it is merely isolated from its HA master 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 to traditional vSphere environments where the management network is used for vSphere HA communication. However, even in VSAN environments, vSphere HA continues to use the default gateway on the management network for isolation detection responses. This should be changed so that the isolation response IP address is on the VSAN network.

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

Network Isolation Response and Multiple Isolation Response Addresses

In a Virtual SAN Stretched Cluster, one of the isolation addresses should reside in the site 1 datacenter and the other should reside in the site 2 datacenter. This would enable vSphere HA to validate complete network isolation in the case of a connection failure between sites.

VMware recommends enabling host isolation response and specifying an isolation response addresses that is 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 Virtual SAN Stretched Cluster site and another isolation response IP address from the secondary Virtual SAN Stretched Cluster 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 first 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 second site.

For further details on how to configure this setting, information can be found in [KB Article 1002117](#).

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 stretched environments.

If administrators wish to enable DRS on Virtual SAN Stretched Cluster, there is a requirement to have a vSphere Enterprise license edition or higher.

There is also a requirement to create VM to Host affinity rules mapping VM to Host groups. These specify which virtual machines and hosts reside in the preferred site and which reside in the secondary 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. Note that if DRS is not enabled on the cluster, then VM to Host affinity “should” rules are not honored. These soft (should) rules are DRS centric and are honored/rectified/warned only when DRS is enabled on the cluster.

Another consideration is that without DRS, there will be considerable management overhead for administrators, as they will have to initially place virtual machines on the correct hosts in order for them to power up without violating the host affinity rules. If the virtual machine is initially placed on the incorrect host, administrators will need to manually migrate them to the correct site before they can be powered on.

Another consideration is related to full site failures. On a site failure, vSphere HA will restart all virtual machines on the remaining site. When the failed site recovers, administrators will have to identify the virtual machines that should reside on the recovered site, and manually move each virtual machine back to the recovered site manually. DRS, with affinity rules, can make this operation easier.

With vSphere DRS enabled on the cluster, the virtual machines can simply be deployed to the cluster, and then the virtual machine is powered on, DRS will move the virtual machines to the correct hosts to conform to the Host/VM groups and rules settings. Determining which virtual machines should be migrated back to a recovered site is also easier with DRS.

Another area where DRS can help administrators is by automatically migrating virtual machines to the correct site in the event of a failure, and the failed site recovers. DRS, and VM/Host affinity rules, can make this happen automatically without administrator intervention.

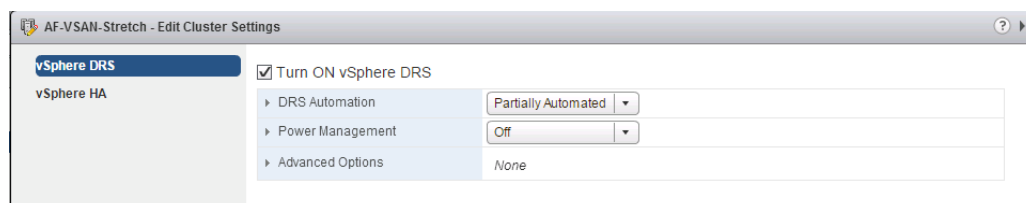
VMware recommends enabling vSphere DRS on Virtual SAN Stretched Clusters where the vSphere edition allows it.

Partially Automated or Fully Automated DRS

Customers can decide whether to place DRS in *partially automated* mode or fully automated mode. With partially automated mode, DRS will handle the initial placement of virtual machines. However any further migration recommendations will be surfaced up 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.

With *fully automated* mode, DRS will take care of the initial placement and on-going load balancing of virtual machines. DRS should still adhere to the Host/VM groups and rules, and should never balance virtual machines across different sites. This is important as virtual machines on Virtual SAN Stretched Cluster will use read locality, which implies that they will cache locally. If the virtual machine is migrated by DRS to the other site, the cache will need to be warmed on the remote site before the virtual machine reaches it previous levels of performance.

One significant consideration with 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. Now 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 now back in the cluster. If in fully automated mode, the affinity rules are checked, and obviously a lot of them 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, but can now wait until VSAN has fully resynced the virtual machine components. DRS can then be changed back to fully automated mode, which will allow virtual machine migrations to take place to conform to the VM/Host affinity rules.



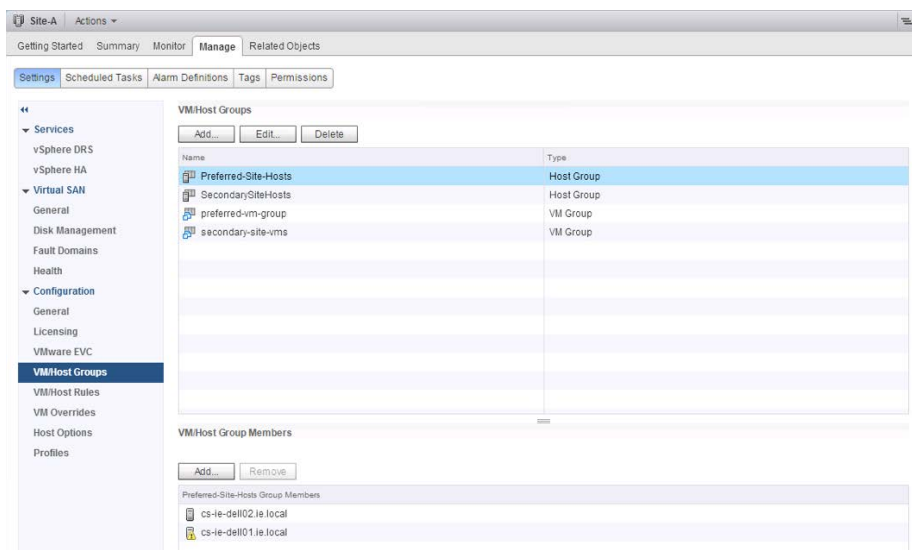
VM/Host Groups & Rules

VMware recommends enabling vSphere DRS to allow for the creation of Host-VM affinity rules to do initial placement of VMs and to avoid unnecessary vMotion of VMs between sites, and impacting read locality. Because the 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. If virtual machines are allowed to move freely across sites, it may end up on the remote site. Since Virtual SAN Stretched Cluster implements read locality, the cache on the remote site will be cold. This will impact performance until the cache on the remote site has been warmed.

Note that Virtual SAN Stretched Cluster has its own notion of a **preferred site**. This is setup at the configuration point, 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 two data sites *and* the witness agent can talk to both sites. In that case, the witness agent needs to decide which side’s cluster it will stick with. It does so with what has been specified as the “preferred” site.

Host groups

When configuring DRS with a Virtual SAN Stretched Cluster, **VMware recommends** creating two VM-Host affinity groups. An administrator could give these host groups the names of *preferred* and *secondary* to match the nomenclature used by VSAN. The hosts from site 1 should be placed in the *preferred* host group, and the hosts from site 2 should be placed in the *secondary* host group.



VM Groups

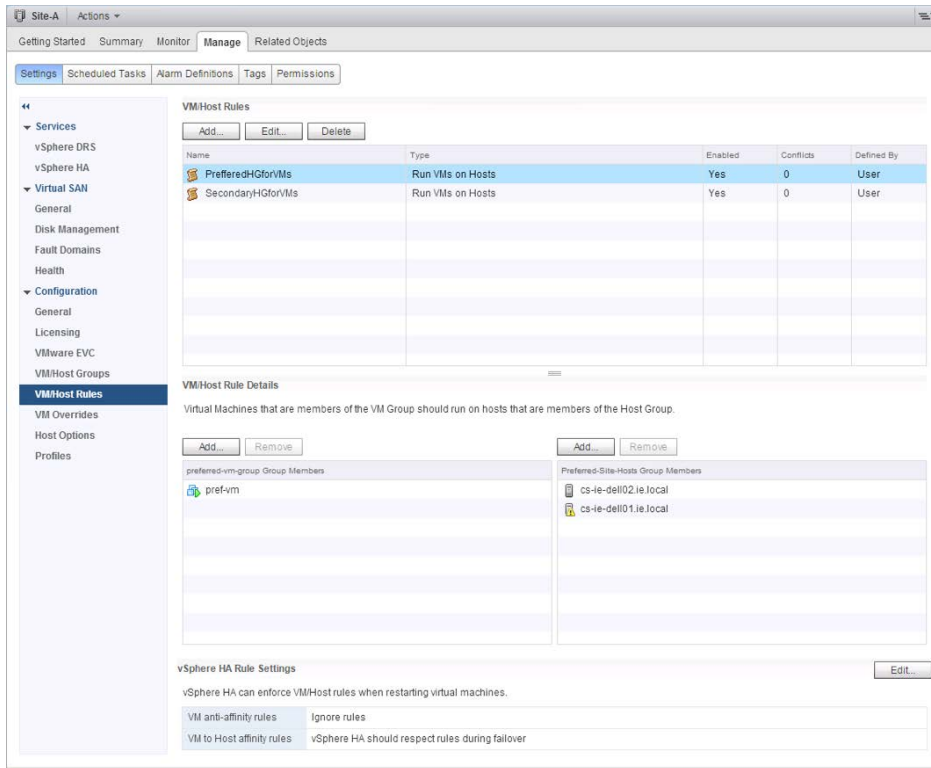
Two VM groups should also be created; one to hold the virtual machines placed on site 1 and the other to hold the virtual machines placed on site 2. Whenever a virtual machine is created and before it is powered on, assuming a *NumberOfFailuresToTolerate* policy setting of 1, the virtual machine should be added to the correct host affinity group. This will then ensure that a virtual always remains on the same site, reading from the same replica, unless a site critical event occurs necessitating the VM being failed over to the secondary site.

Note that to correctly use VM groups, first off all create the VM, but do power it on. Next, edit the VM groups and add the new VM to the desired group. Once added, and saved, the virtual machine can now be powered on. With DRS enabled on the cluster, the virtual machine will be checked to see if it is on the correct site according to the VM/Host Rules (discussed next) and if not, it is automatically migrated to the appropriate site, either “preferred” or “secondary”.

VM/Host Rules

When deploying virtual machines on a Virtual SAN Stretched Cluster, for the majority of cases, we wish the virtual machine to reside on the set of hosts in the selected host group. However, in the event of a full site failure, we wish the virtual machines to be restarted on the surviving site.

To achieve this, **VMware recommends** implementing “should respect rules” in the VM/Host Rules configuration section. These rules may be violated by vSphere HA 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 in this case, 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.



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 the event of a complete site failure. The “should rules” are implemented by clicking on 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”.

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.

Installation

The installation of Virtual SAN 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 delving into the installation of a Virtual SAN Stretched Cluster, there are a number of important features to highlight that are specific to stretch cluster environments.

What is a Preferred domain/preferred site?

Preferred domain/preferred site is simply a directive for Virtual SAN. The “preferred” site is the site that Virtual SAN wishes to remain running when there is a failure and the sites can no longer communicate. One might say that the “preferred site” is the site 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, while 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 Virtual SAN 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 compute resides. This is not the regular VSAN algorithm, which reads in a round-robin fashion across all replica copies of the data.

This new algorithm for Virtual SAN Stretched Clusters will reduce 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 Virtual SAN Stretched Clusters, 50% of the read requests will be sent to the remote site. This is a significant consideration for sizing of the network bandwidth. Please refer to the sizing of the network bandwidth between the two main sites for more details.

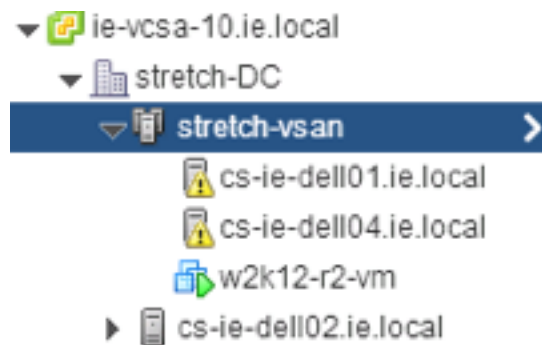
The advanced parameter `VSAN.DOMOwnerForceWarmCache` can be enabled or disabled to change the behavior of read locality. This advanced parameter

is hidden and is not visible in the Advanced System Settings vSphere web client. It is only available the CLI.

Caution: Read locality is enabled by default when Virtual SAN 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 Virtual SAN stretched cluster, only data hosts must be in the cluster object in vCenter. The witness host must remain outside of the cluster, and must not be added to the cluster at any point. Thus for a 1+1+1 configuration, where there is one host at each site and one physical ESXi witness host, the configuration will look similar to the following:



Note that the witness host is not shaded in blue in this case. The witness host only appears shaded in blue when a witness appliance (OVA) is deployed. Physical hosts that are used as witness hosts are not shaded in blue.

Virtual SAN Health Check Plugin for Stretched Clusters

Virtual SAN 6.1, shipped with vSphere 6.0U1, has a health check feature built in. This functionality was first available for Virtual SAN 6.0. The updated 6.1 version of the health check for Virtual SAN has enhancements specifically for Virtual SAN stretched cluster.

Once the ESXi hosts have been upgraded or installed with ESXi version 6.0U1, there are no additional requirements for enabling the VSAN health check. Note that ESXi version 6.0U1 is a requirement for Virtual SAN Stretched Cluster.

Similarly, once the vCenter Server has been upgraded to version 6.0U1, the VSAN Health Check plugin components are also upgraded automatically, provided vSphere DRS is licensed, and DRS Automation is set to Fully Automated. If vSphere DRS is not licensed, or not set to Fully Automated, then hosts will have to be evacuated and the Health Check vSphere Installable Bundle (vib) will have to be installed manually.

Please refer to the 6.1 Health Check Guide for additional information. The location is available in the appendix of this guide.

New Virtual SAN health checks for Stretched Cluster configurations

As mentioned, there are new health checks for Virtual SAN Stretched Cluster. Select the Cluster object in the vCenter inventory, click on Monitor > Virtual SAN > Health. Ensure the stretched cluster health checks pass when the cluster is configured.

Note that the stretched cluster checks will not be visible until the stretch cluster configuration is completed.

Virtual SAN Health (Last checked: Today at 12:38)	
Test Result	Test Name
✓ Passed	▶ Data health
✓ Passed	▶ Limits health
✓ Passed	▶ Network health
✓ Passed	▶ Physical disk health
✓ Passed	▼ Stretched cluster health
✓ Passed	Cluster with multiple unicast agents
✓ Passed	Fault domain number check
✓ Passed	Host without configured unicast agent
✓ Passed	Some hosts do not support stretched cluster
✓ Passed	Stretched cluster with no disk mapping witness host
✓ Passed	Stretched cluster without a witness host
✓ Passed	Witness host inside one of the fault domains
✓ Passed	Witness host part of cluster
✓ Passed	Witness host with invalid preferred fault domain
✓ Passed	Witness host with non-existing fault domain

Using a witness appliance

Virtual SAN stretched cluster supports the use of a ESXi virtual machine as a witness host. This is available as an OVA (Open Virtual Appliance) from VMware. However this witness ESXi virtual machine needs to reside on a physical ESXi host, and that requires some special networking configuration for the witness ESXi virtual machine.

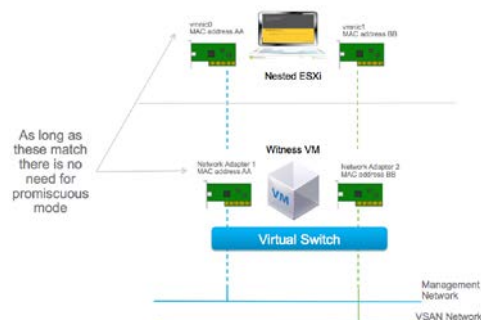
Physical ESXi preparation for witness deployment

The witness ESXi virtual machine contains two network adapters. One of the network adapters is used to connect to the ESXi/vCenter management network whilst the other network adapter is used to connect to the VSAN network. Therefore, there is a requirement to have two virtual machine network created on the physical ESXi host so that the witness ESXi virtual machine can communicate to the rest of the cluster. One of the virtual machine networks should be able to reach the management network and the other virtual machine network should be able to reach the VSAN network.

A note about promiscuous mode

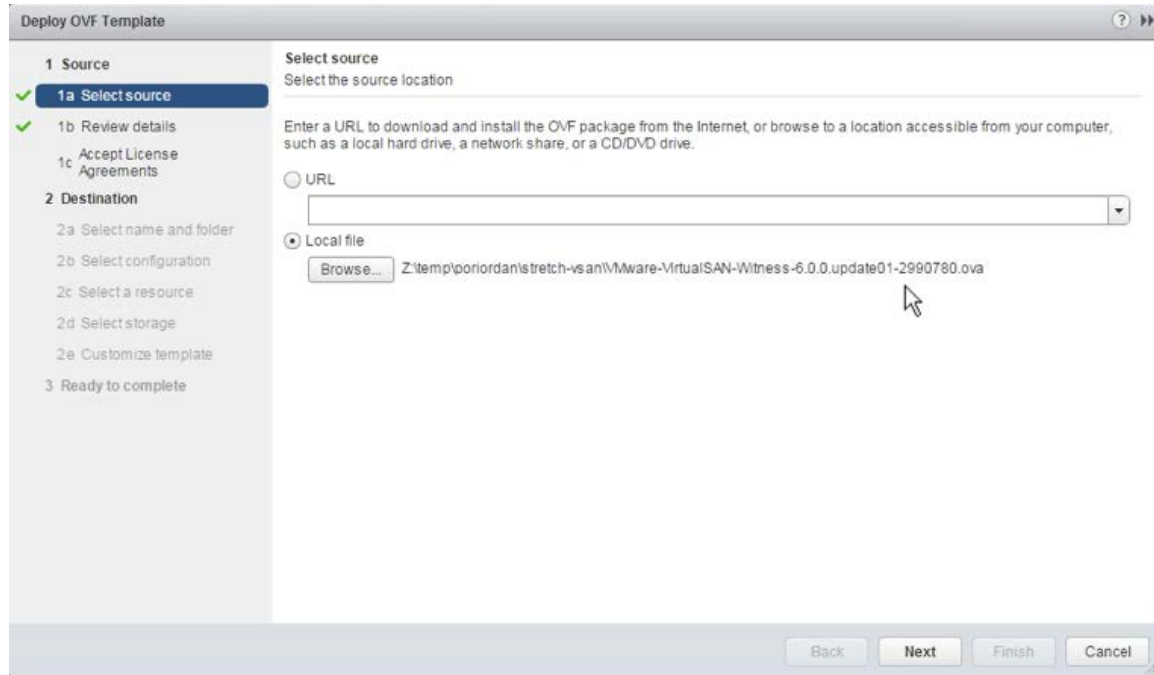
In many nested ESXi environments, there is a recommendation to enable promiscuous mode to allow all Ethernet frames to pass to all VMs that are attached to the port group, even if it is not intended for that particular VM. The reason promiscuous mode is enabled in many nested environments is to prevent a virtual switch from dropping packets for (nested) vmnics that it does not know about on nested ESXi hosts.

If the MAC address of the virtual machine network adapter matches the MAC address of the nested ESXi vmnic, no packets are dropped. The witness ESXi virtual machine OVA has been configured to have the MAC addresses match, then promiscuous mode would not be needed.

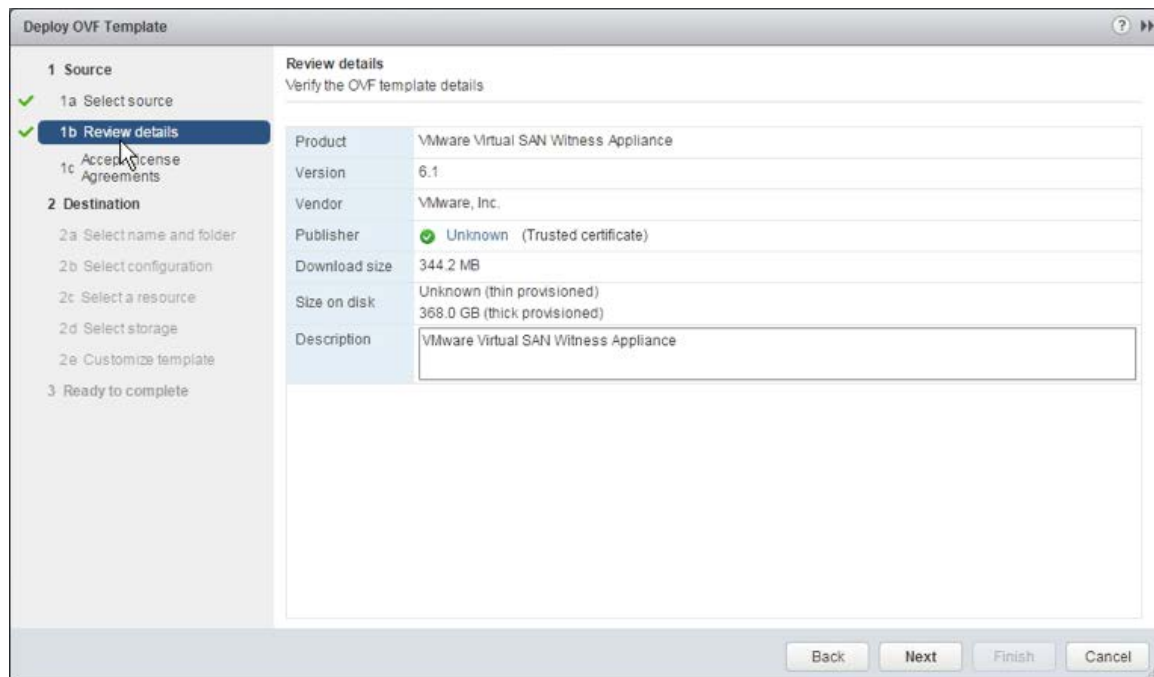


Setup Step 1: Deploy the Witness ESXi OVA

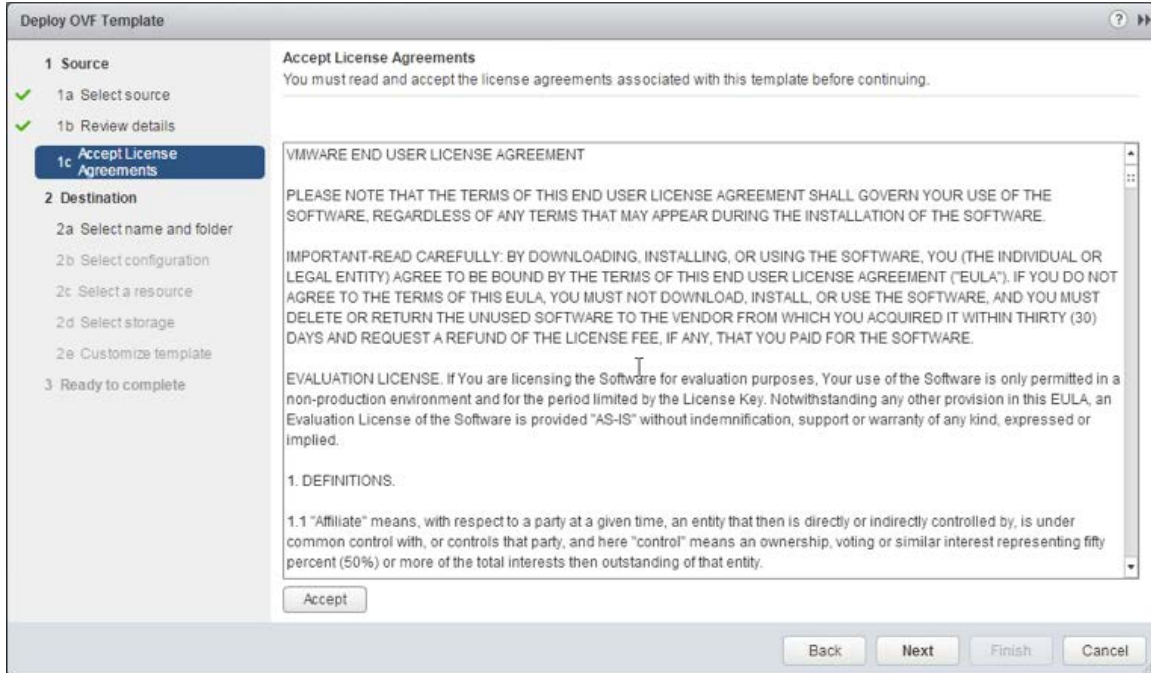
The first step is to download and deploy the witness ESXi OVA, or deploy it directly via a URL, as shown below. In this example it has been downloaded:



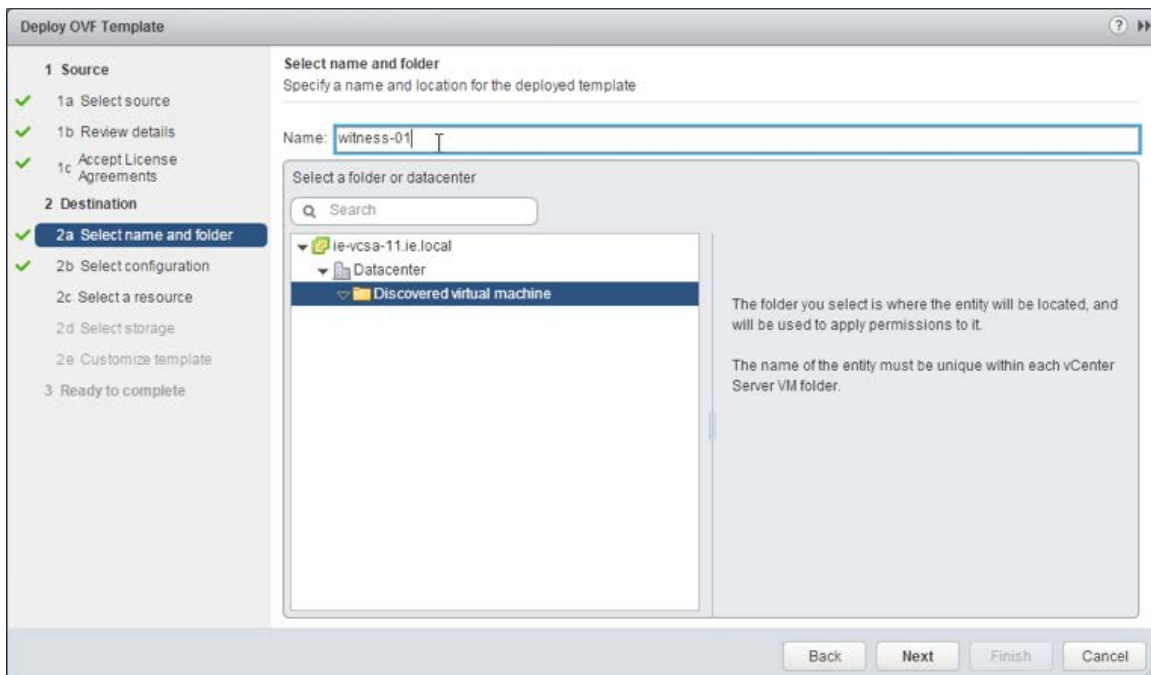
Examine the details. Note that it states that this is the VMware Virtual SAN Witness Appliance, version 6.1.



Accept the EULA as shown below:

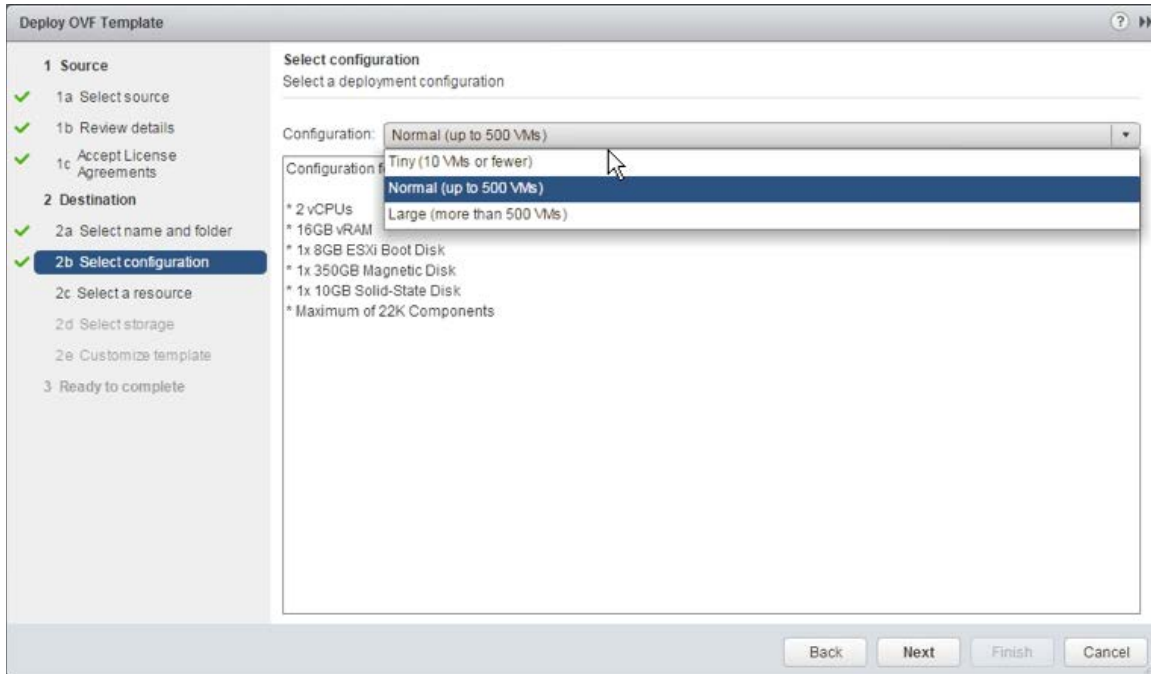


Give the witness a name (e.g. witness-01), and select a folder to deploy it to.

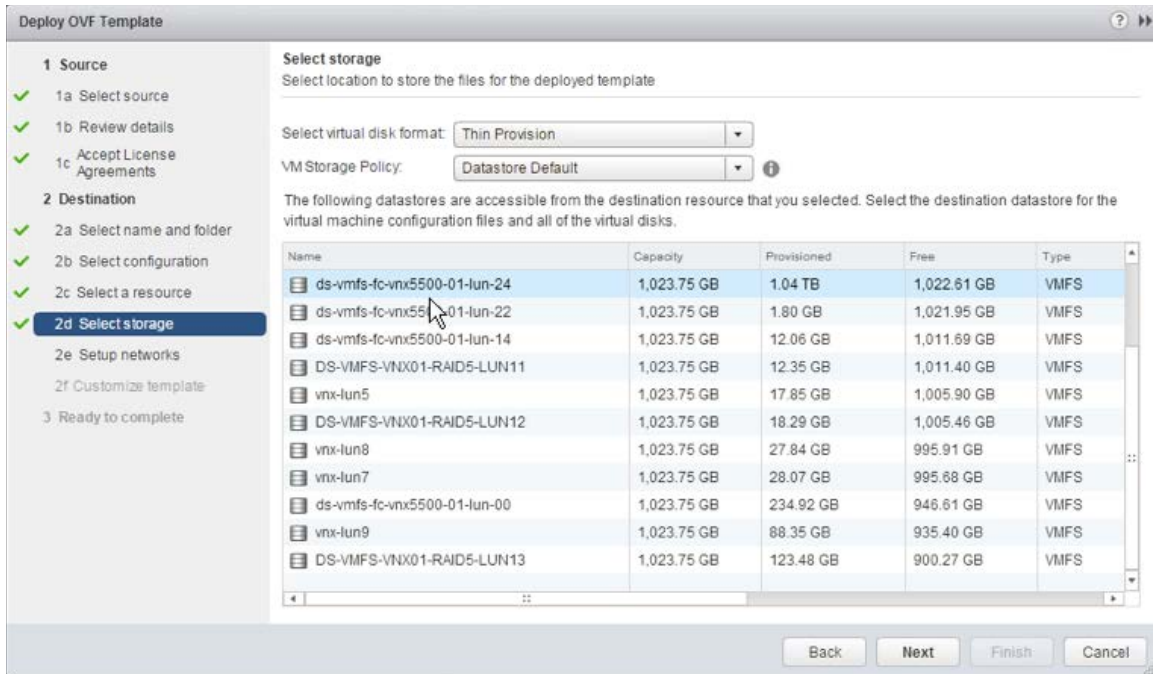


At this point a decision needs to be made regarding the expected size of the stretched cluster configuration. There are three options offered. If you expect the number of VMs deployed on the Virtual SAN Stretched Cluster to be 10 or fewer, select the **Tiny** configuration. If you expect to deploy more than 10 VMs,

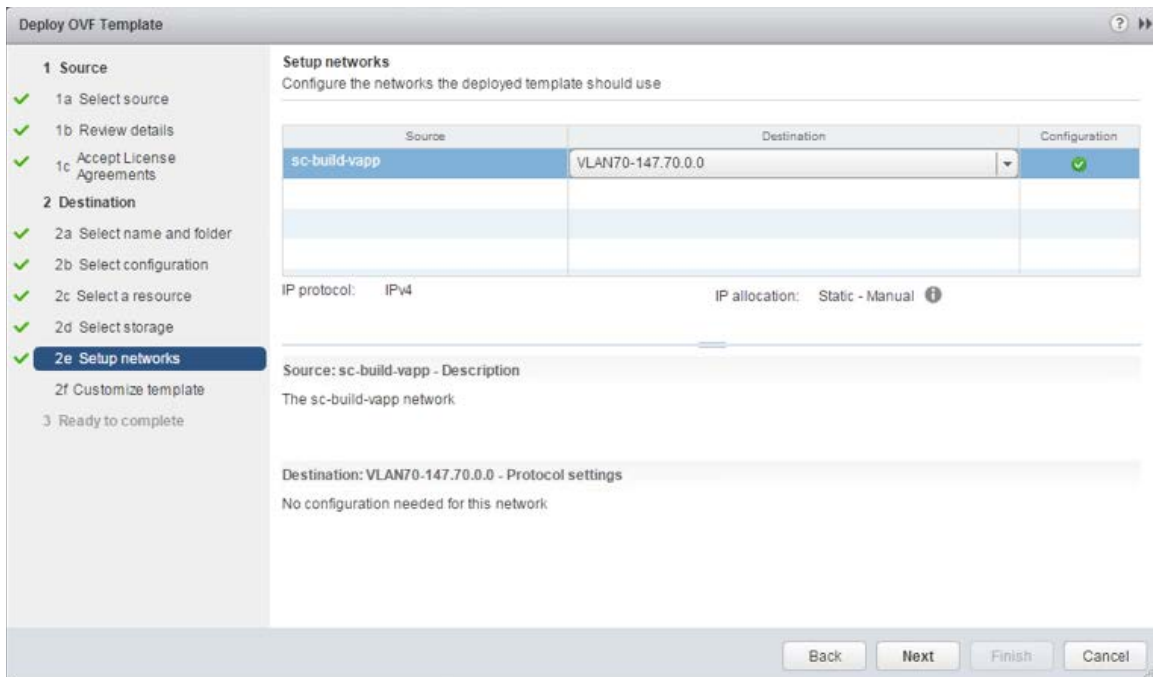
but less than 500 VMs, then the **Normal** (default option) should be chosen. For more than 500 VMs, choose the **Large** option. On selecting a particular configuration, the resources consumed by the appliance and displayed in the wizard (CPU, Memory and Disk):



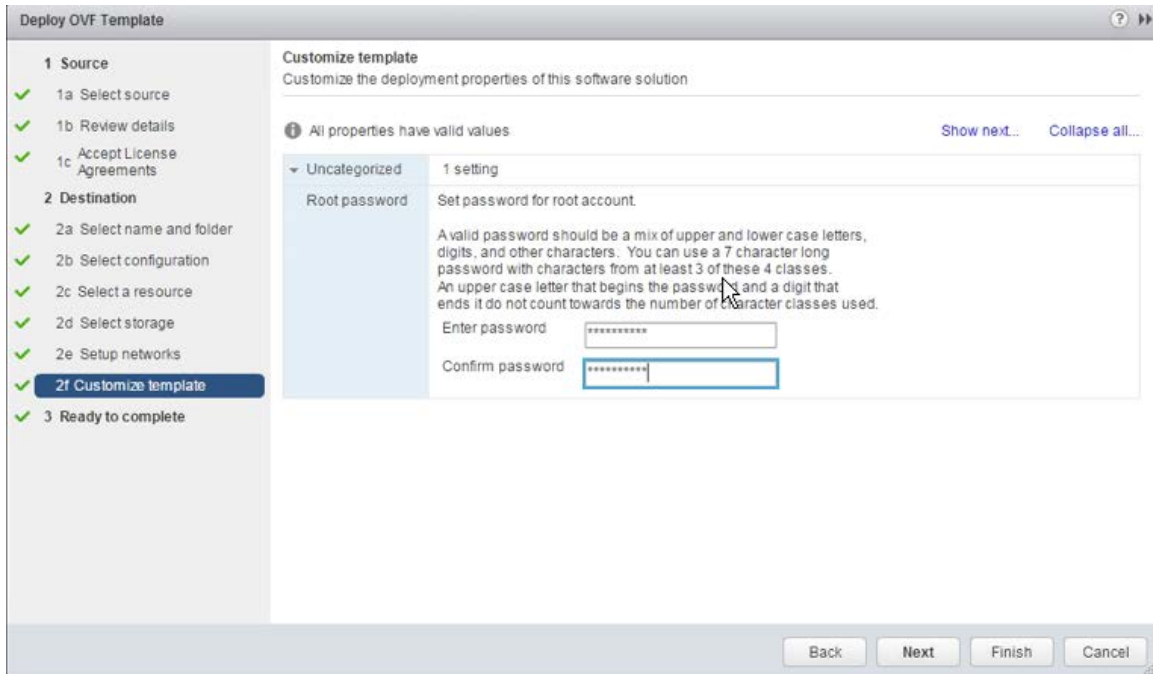
Select a datastore for the witness ESXi VM. This will be one of the datastore available to the underlying physical host. You should consider when the witness is deployed as thick or thin, as thin VMs may grow over time, so ensure there is enough capacity on the selected datastore.



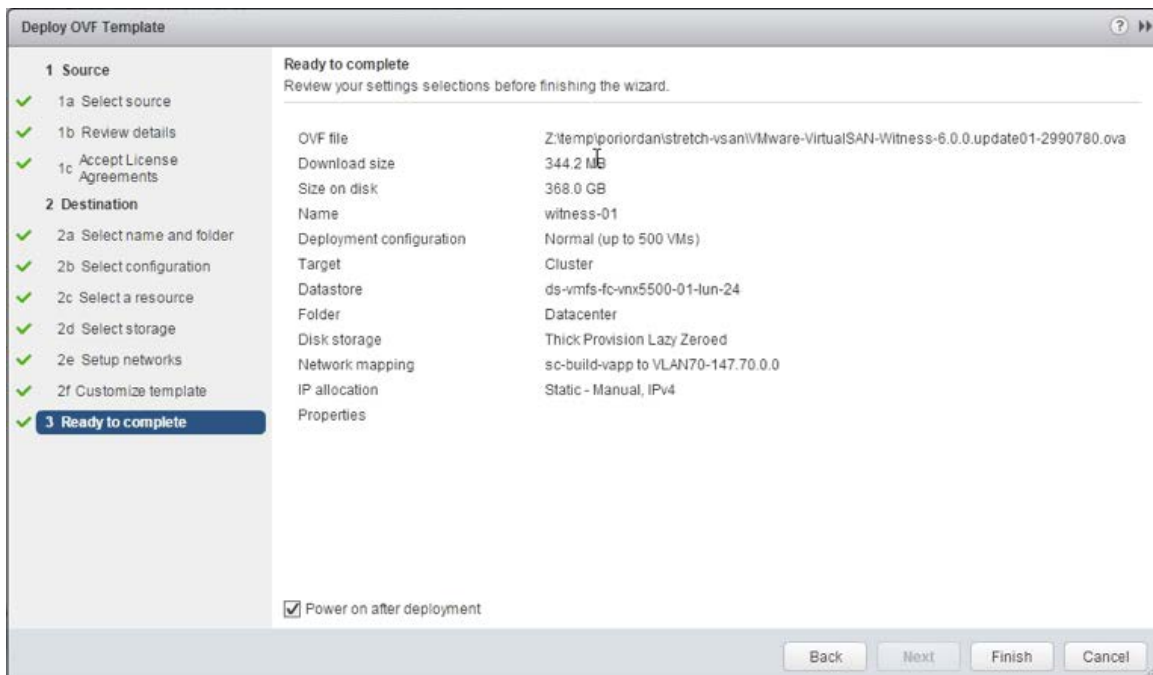
Select a network for the management network. This gets associated with both network interfaces (management and VSAN) at deployment, so later on the VSAN network configuration will need updating.



Give a root password for the witness ESXi host:



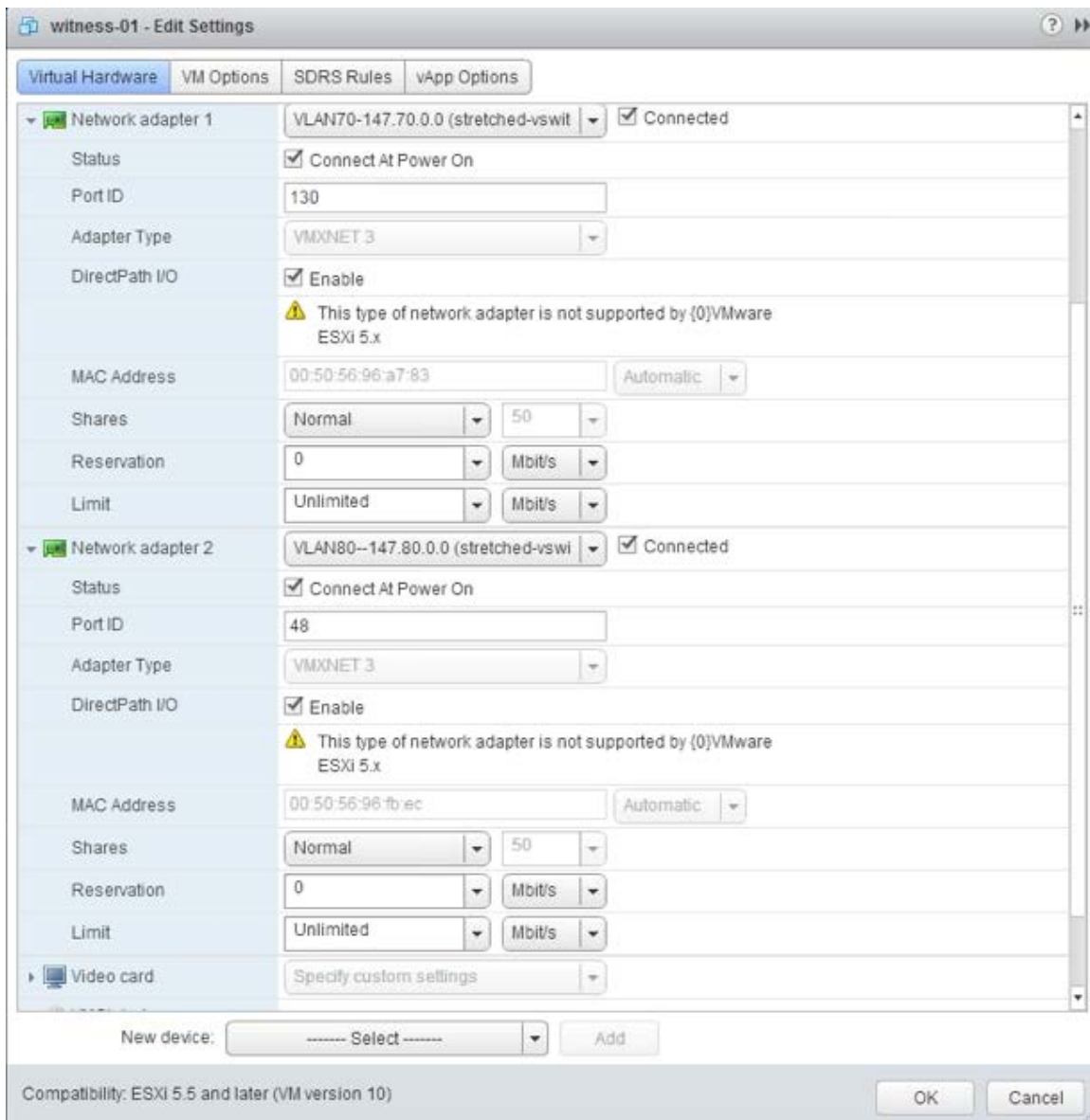
At this point, the witness appliance (ESXi VM) is ready to be deployed. You can choose to power it on after deployment by selecting the checkbox below, or power it on manually via the vSphere web client UI later:



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

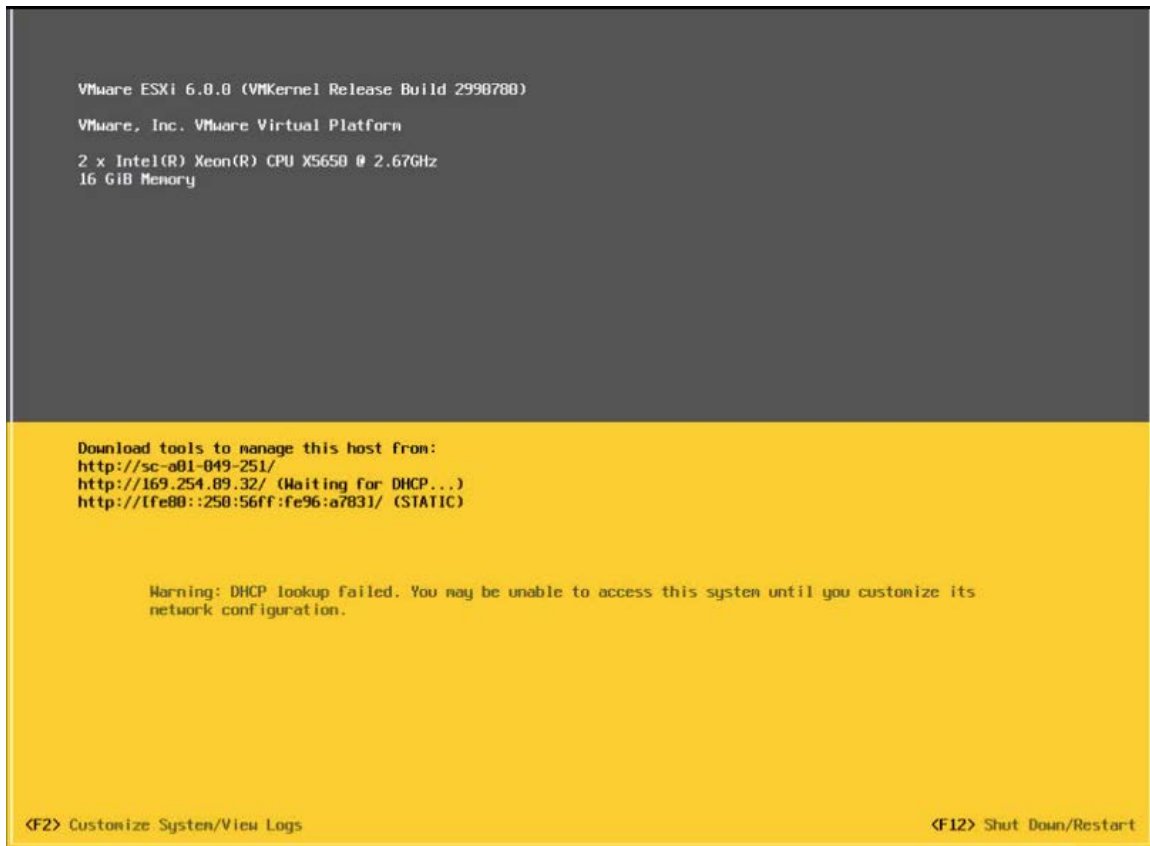
Setup Step 2: Configure Witness ESXi VM management network

Once the witness ESXi virtual machine has been deployed, select it in the vSphere web client UI, and edit the settings. As mentioned previously, there are two virtual networks adapters. Both adapters will be assigned to the management network that was selected during the OVA deploy. At this time, there is no way to select alternate networks, such as the VSAN network) during the deployment. Administrators will have to edit the network for network adapter 2 to ensure that it is attached to the correct VSAN network.



At this point, the console of the witness ESXi virtual machine 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, it is very likely that the landing page of the DCUI will look something similar to the following:



```
VMware ESXi 6.0.0 (VMKernel Release Build 2998788)
VMware, Inc. VMware Virtual Platform
2 x Intel(R) Xeon(R) CPU X5650 @ 2.67GHz
16 GiB Memory

Download tools to manage this host from:
http://sc-001-049-251/
http://169.254.89.32/ (Waiting for DHCP...)
http://1fe80:250:56ff:fe96:a7831/ (STATIC)

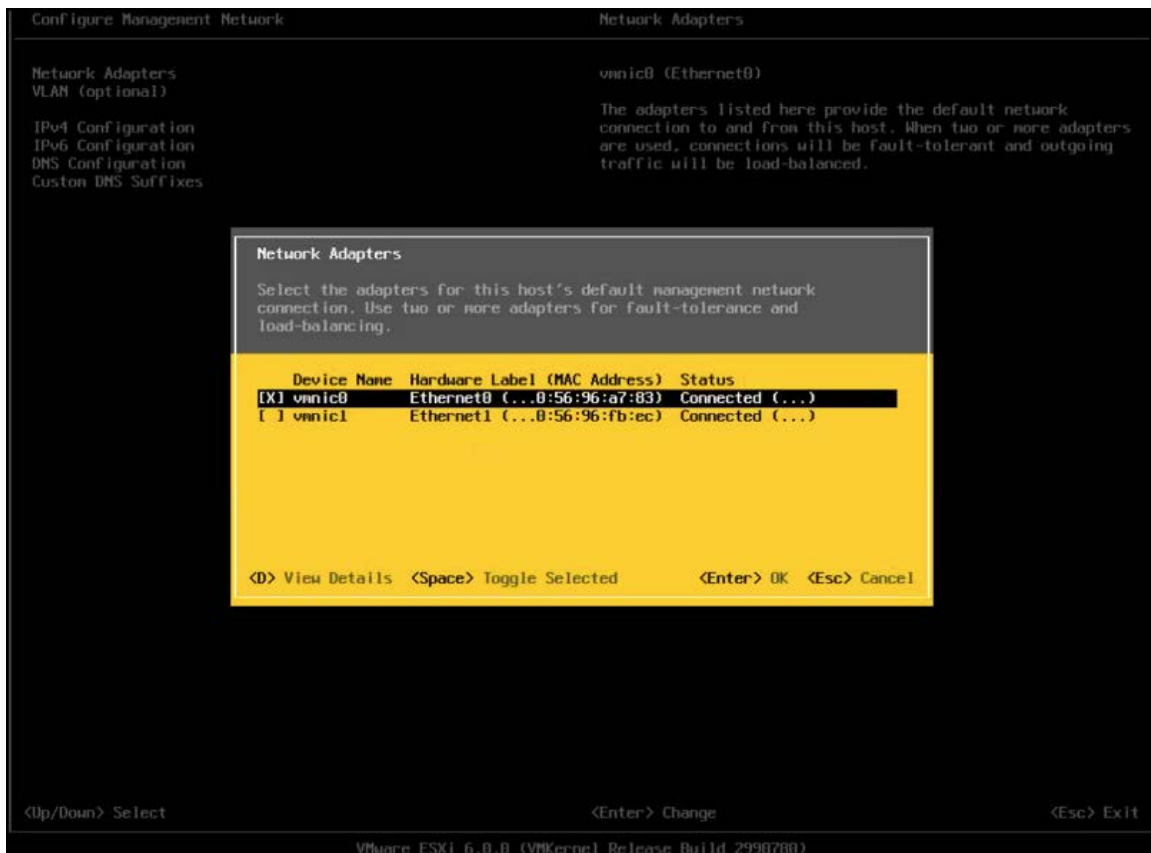
Warning: DHCP lookup failed. You may be unable to access this system until you customize its
network configuration.

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

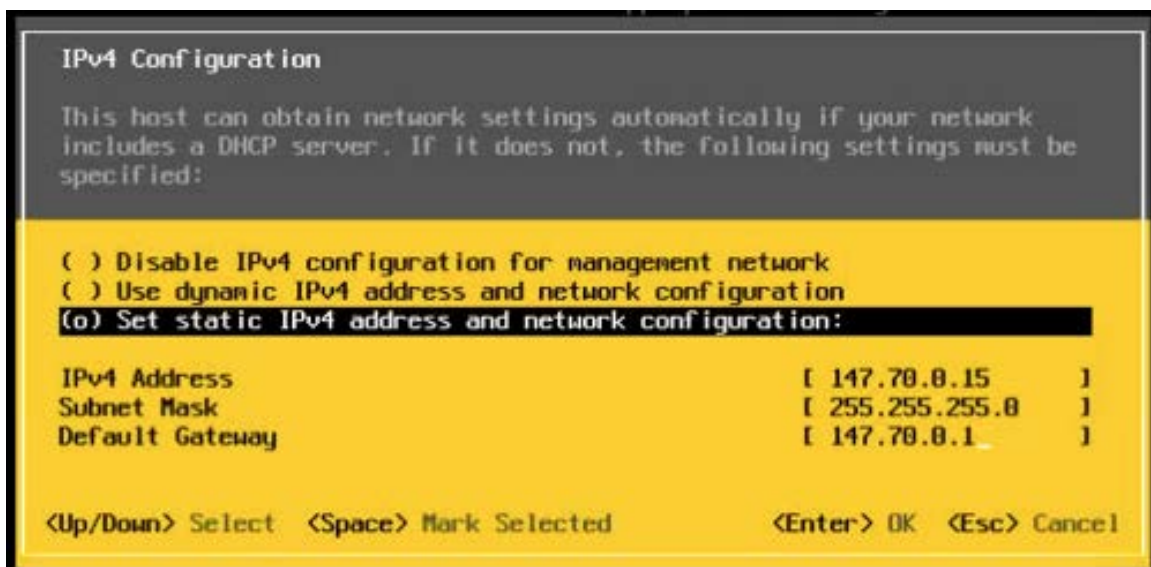
Use the <F2> key to customize the system. The root login and password will need to be provided at this point. This is the root password that was added during the OVA deployment earlier.

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 these match, there is no need to use promiscuous mode on the network, as discussed earlier.

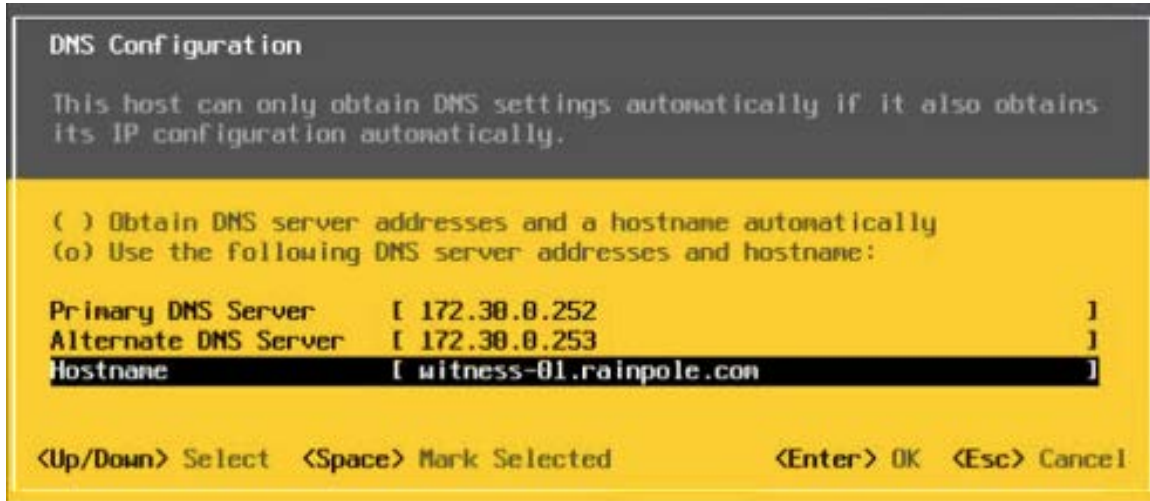
Select vmnic0, and if you wish to view further information, select the key <D> to see more details.



Navigate to the IPv4 Configuration section. This will be using DHCP by default. Select the static option as shown below and add the appropriate IP address, subnet mask and default gateway for this witness ESXi's management network.



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.



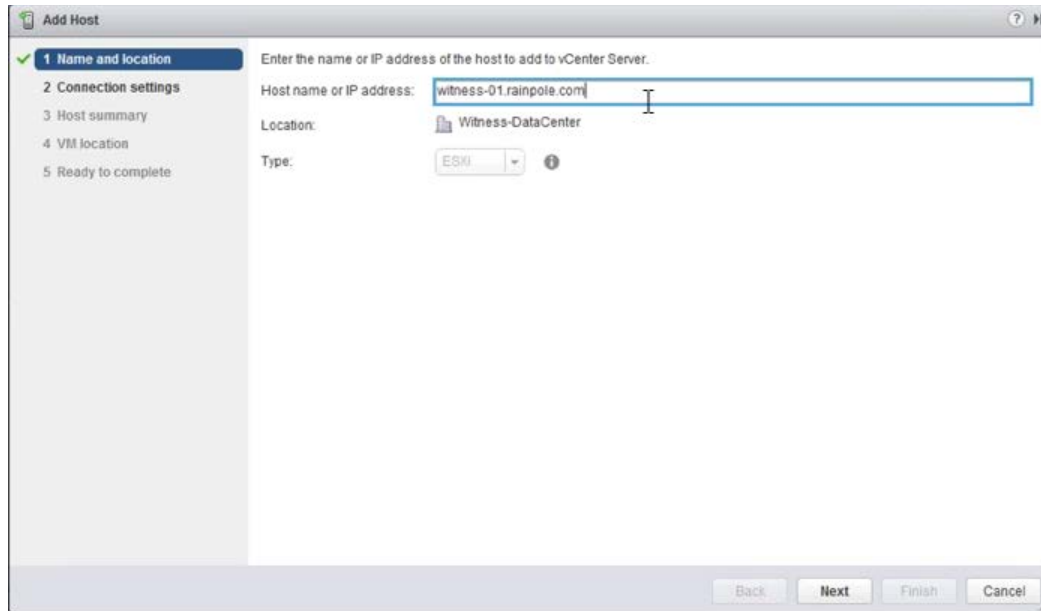
One final recommendation is to do a test of the management network. One can also try adding the IP address of the vCenter server at this point just to make sure that it is also reachable.



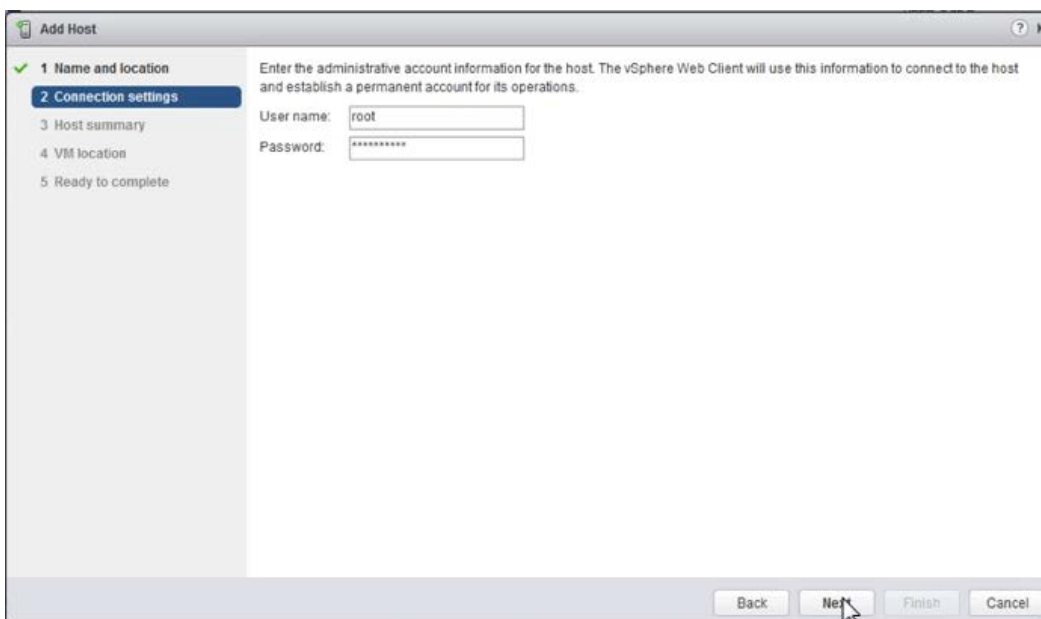
When all the tests have passed, and the FQDN is resolvable, administrators can move onto the next step of the configuration, which is adding the witness ESXi to the vCenter server.

Setup Step 3: Add Witness ESXi VM to vCenter Server

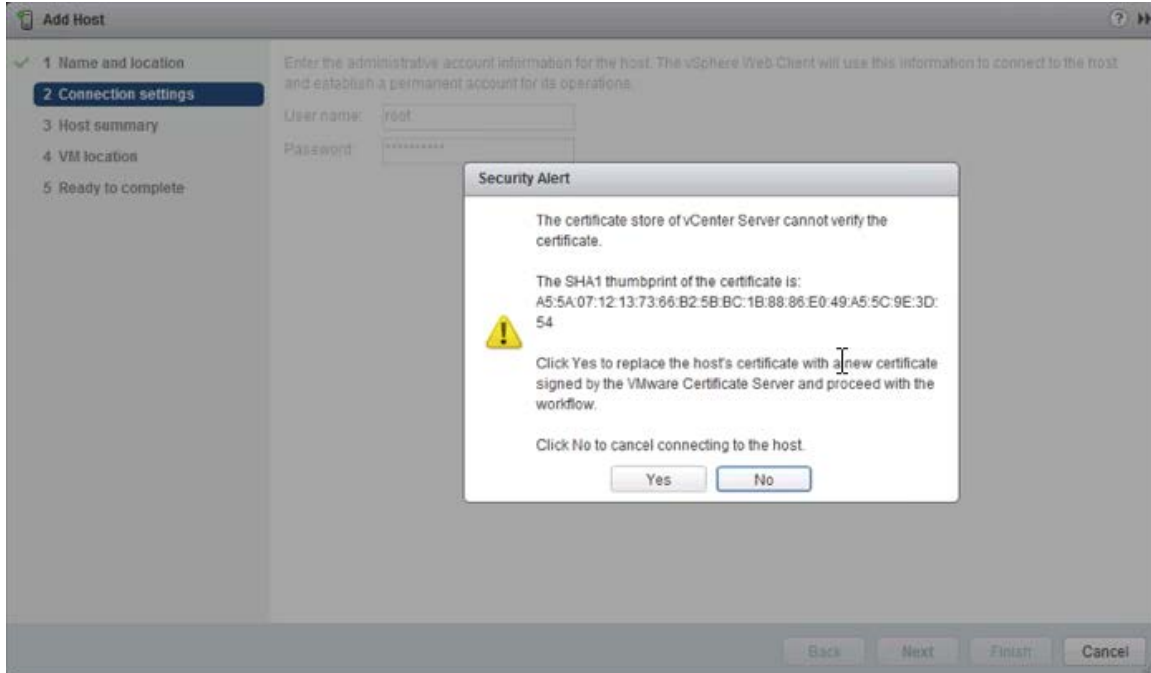
There is no difference to adding the witness ESXi VM to vCenter server when compared to adding physical ESXi hosts. However there are some interesting items to highlight during the process. First step is to provide the name of the witness to. In this example, vCenter server is managing multiple data centers, so we are adding the host to the witness data center.



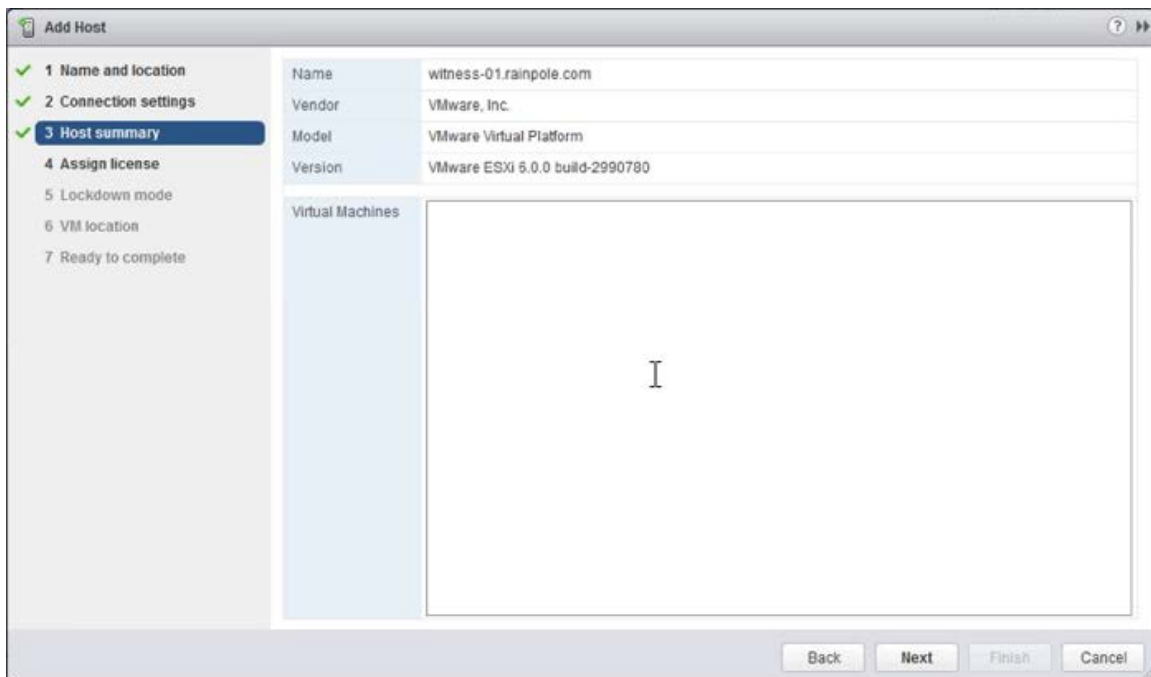
Provide the appropriate credentials, in this example root user and password:



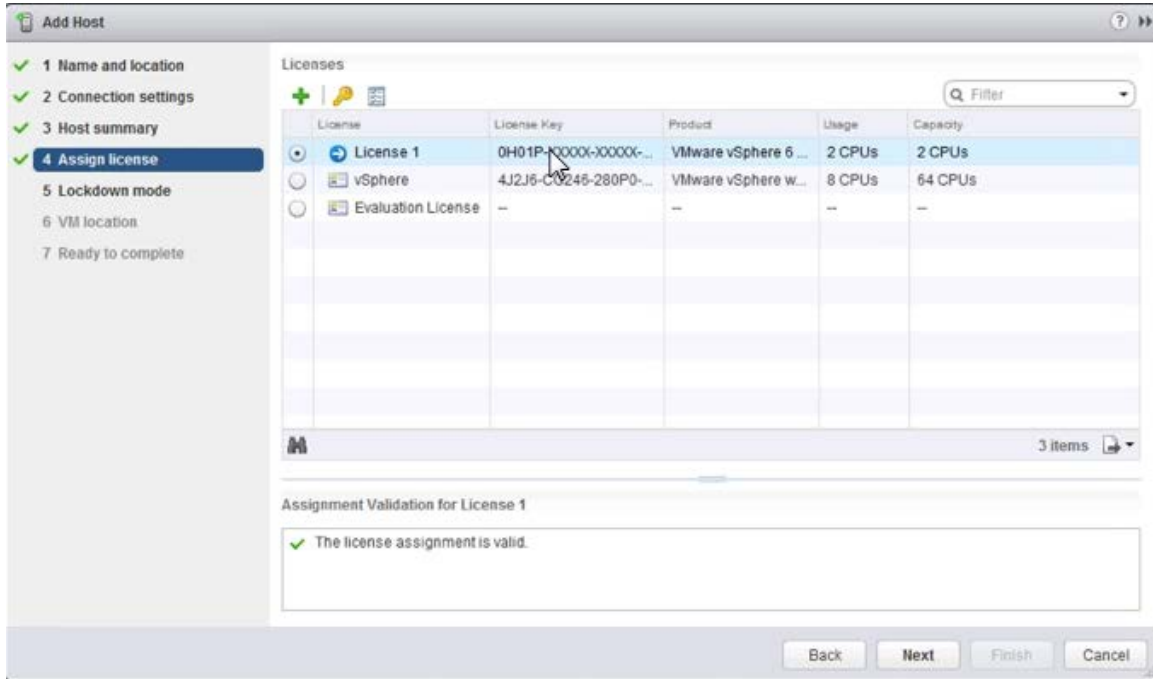
Acknowledge the certificate warning:



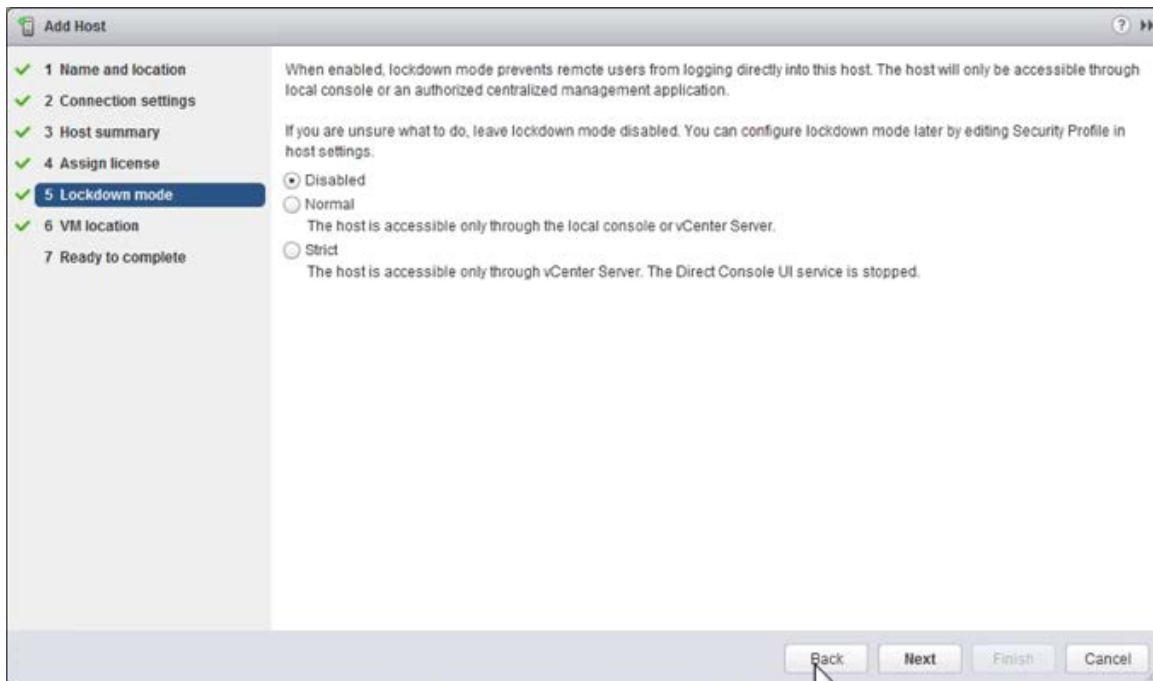
There should be no virtual machines on the witness appliance. Note that it can never run VMs in a Virtual SAN Stretched Cluster configuration. Note also the mode: VMware Virtual Platform. Note also that builds number may differ to the one shown here.



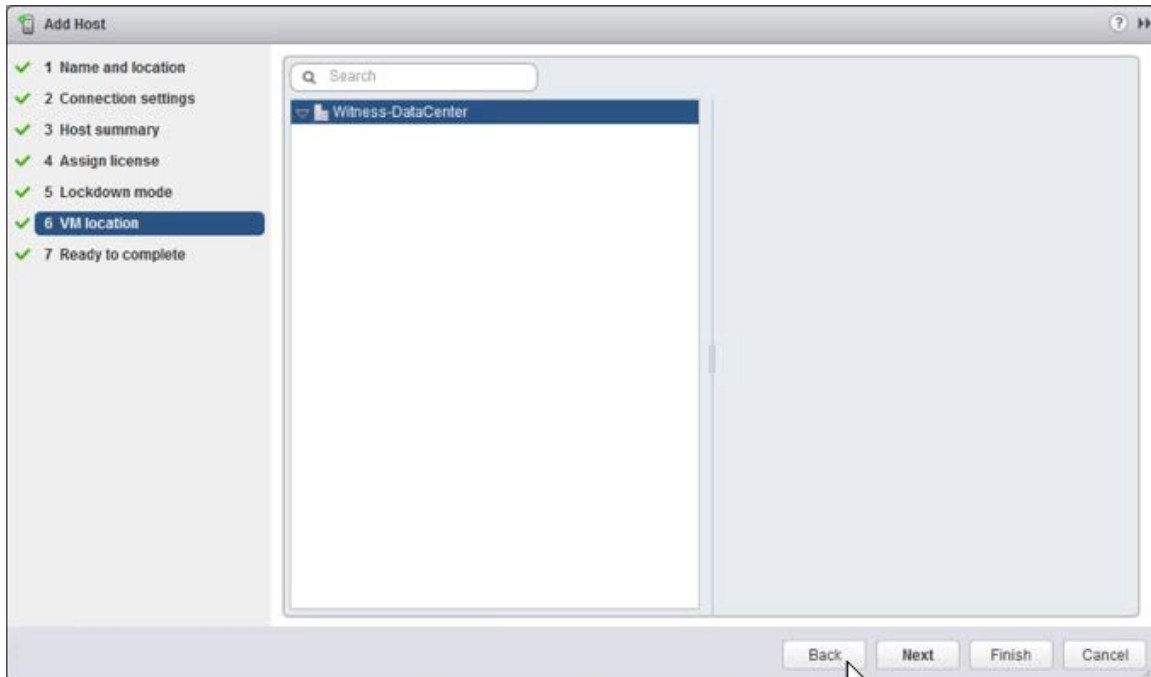
The witness appliance also comes with its own license. You do not need to consume vSphere licenses for the witness appliance:



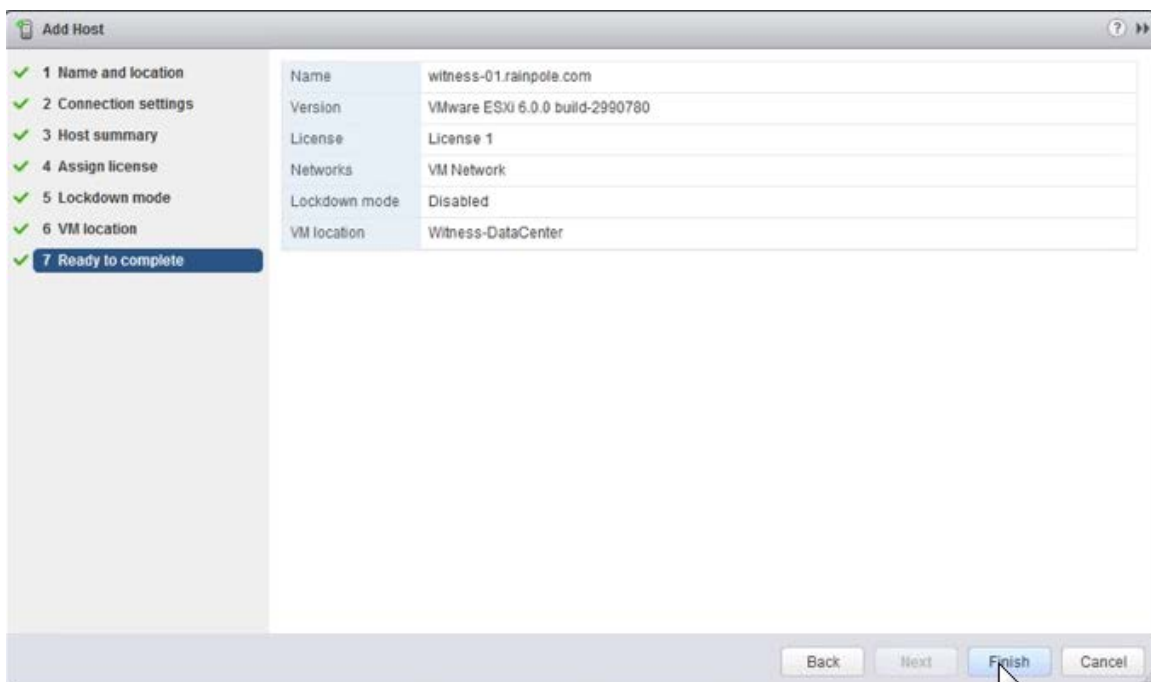
Lockdown mode is disabled by default. Depending on the policies in use at a customer's site, the administrator may choose a different mode to the default:



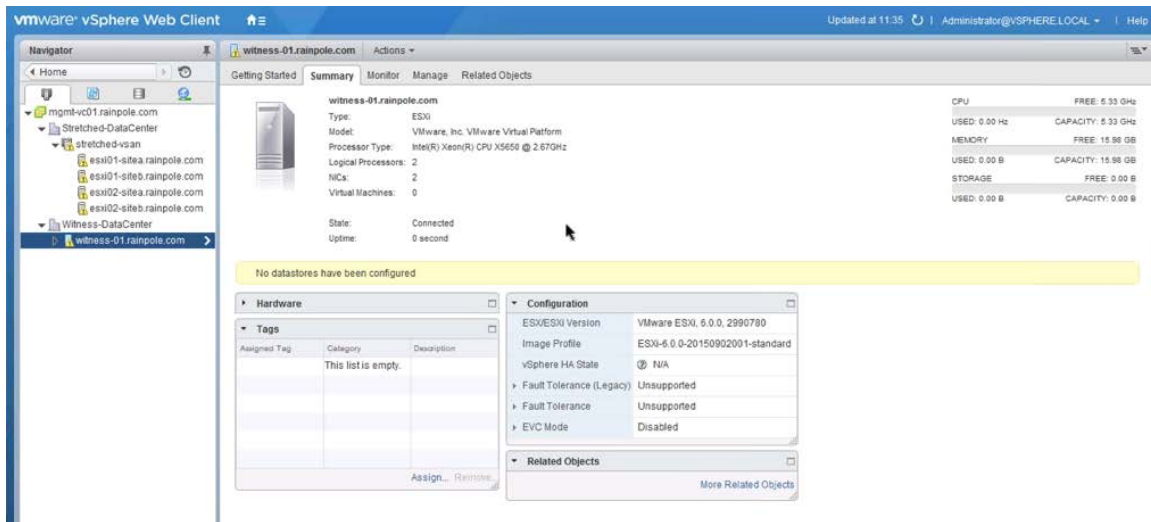
The next step is to choose a location for VMs. This will not matter for the witness appliance, as it will never host virtual machines of its own:



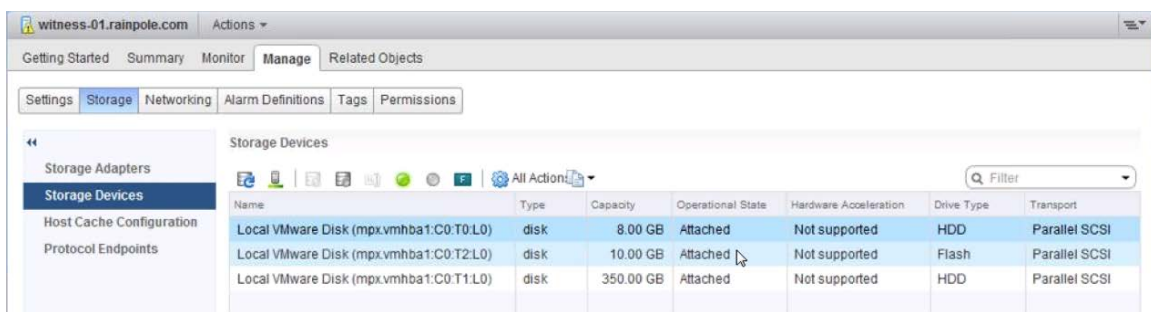
Click finish when ready to complete the addition of the witness to the vCenter server:



One final item of note is the appearance of the witness appliance 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 screen shot below, but should be clearly visible in your infrastructure. (**Note:** the “No datastores have been configured” message is because the nested ESXi host has no VMFS datastore. This can be ignored, or if necessary a small 2GB disk can be added to the host and a VMFS volume can be built on it to remove the message completely).



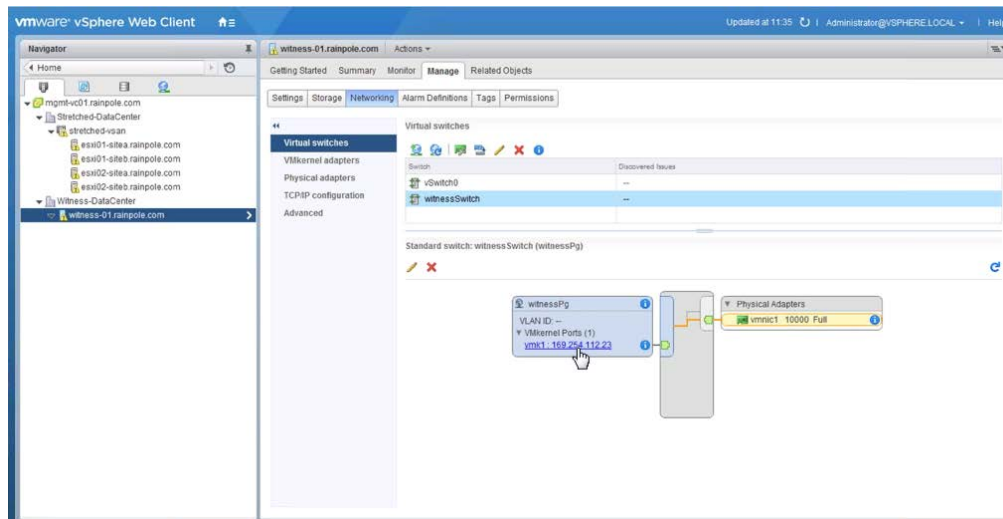
One final recommendation is to verify that the settings of the witness appliance matches the Tiny, Normal or Large configuration selected during deployment. For example, the Normal deployment should have an 8GB HDD for boot, a 10GB Flash that will be configured later on as a cache device and another 350 HDD that will also be configured later on as a capacity device.



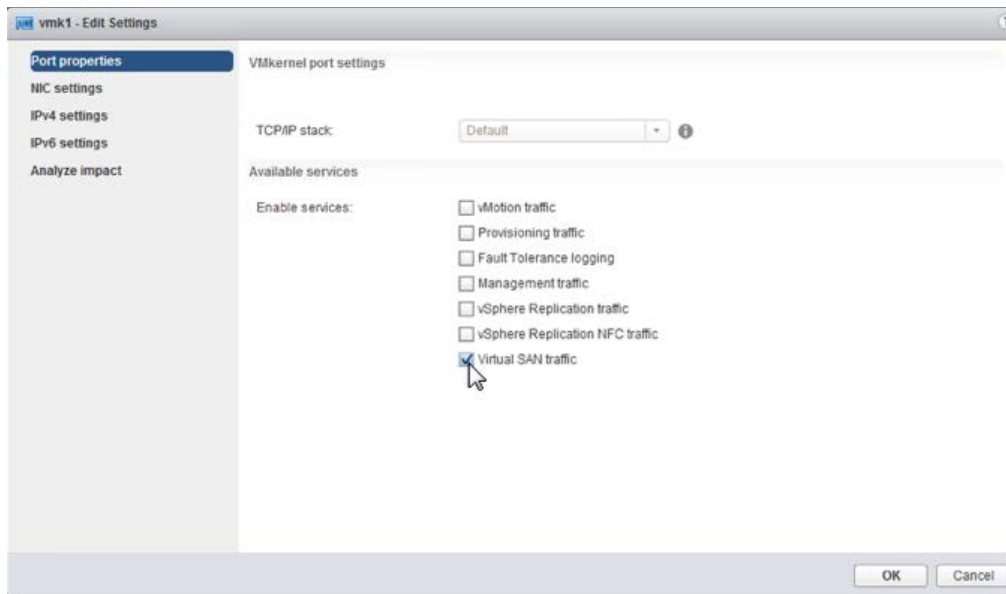
Once confirmed, you can proceed to the next step of configuring the VSAN network for the witness appliance.

Setup Step 4: Configure VSAN network on Witness host

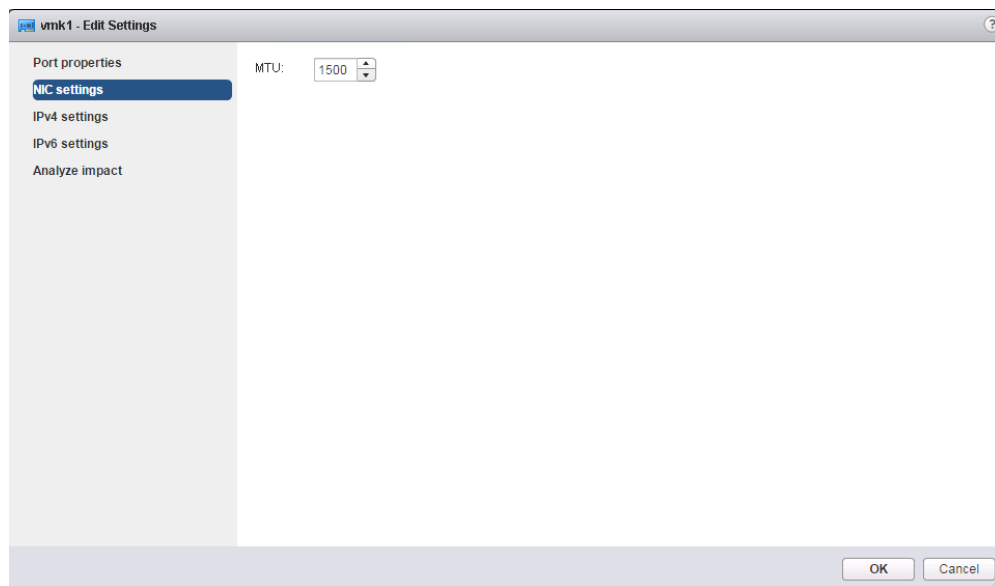
The next step is to configure the VSAN network correctly on the witness ESXi VM. When the witness is selected, navigate to Manage > Networking > Virtual Switches as shown below. The witness has a portgroup pre-defined called *witnessPg*. Here the VMkernel port to be used for VSAN traffic is visible. If there is no DHCP server on the VSAN network (which is likely), then the VMkernel adapter will not have a valid IP address, nor will it be tagged for VSAN traffic.



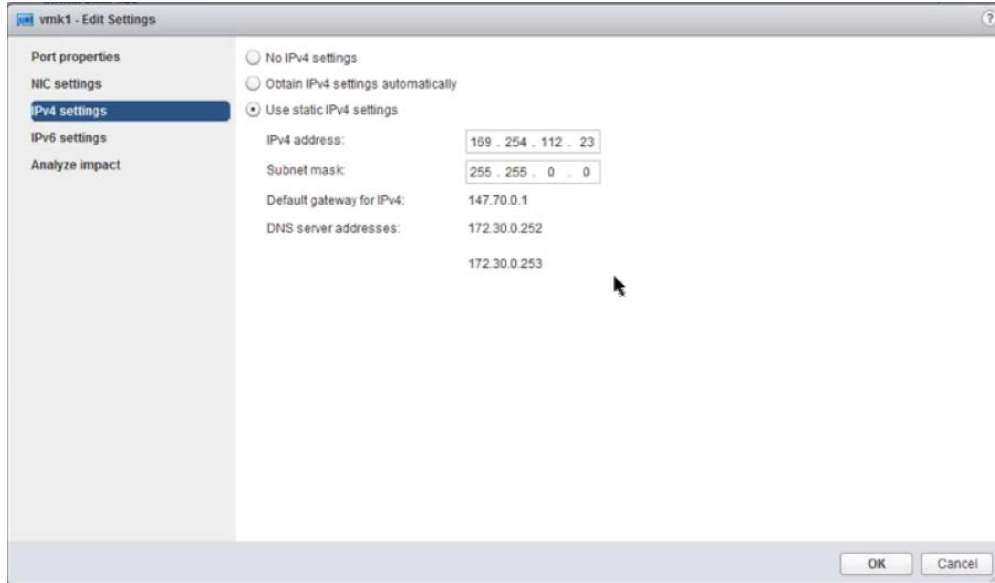
Select the *witnessPg* portgroup (which has a VMkernel adapter), and then select the option to edit it. Tag the VMkernel port for VSAN traffic, as shown below:



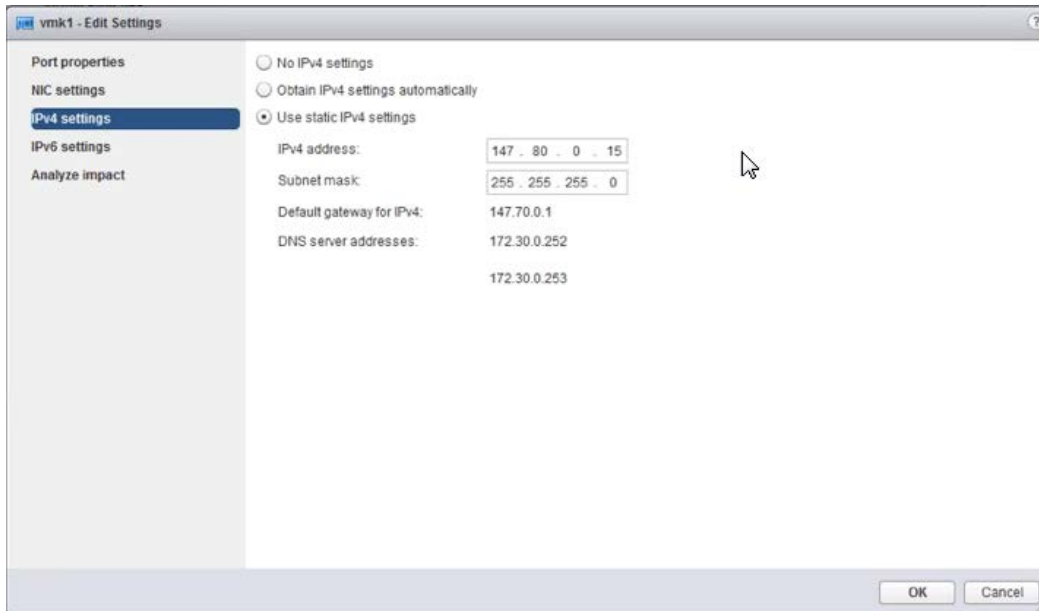
In the NIC settings, ensure the MTU is set to the same value as the Stretched Cluster hosts' VSAN VMkernel interface.



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



Once the VMkernel has been tagged for VSAN traffic, and has a valid IP, click OK.



Setup Step 5: Implement Static Routes

The final step before we can configure the Virtual SAN Stretched Cluster is to ensure that the hosts residing in the data sites can reach the witness host's VSAN network, and vice-versa. In the screen shots below, there are SSH sessions opened to a host in data site 1, a host in data site 2 and the witness host. Due to the reasons outlined earlier, with ESXi hosts having a single default TCPIP stack, and thus a single default gateway, there is no route to the VSAN networks from these hosts. Pings to the remote VSAN networks fail.

The image shows three PuTTY terminal windows. The top-left window is for an ESXi host in data site 1 (esxi01-sitea). It shows the current IP route table, which lacks routes for the witness host's VSAN networks (147.80.0.0/24 and 172.3.0.0/24). It then shows the configuration of static routes for these networks and a failed ping test to the witness host. The top-right window is for an ESXi host in data site 2 (esxi02-siteb), showing a similar lack of routes and a failed ping test. The bottom window is for the witness host (witness-01), showing its current route table with a default gateway and a failed ping test to one of the data site hosts.

```

[esxi01-sitea.rainpole.com - PuTTY]
[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
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:~] esxcli network ip route ipv4 add -n 147.80.0.0/24 vmk0 147.80.0.1
[root@esxi01-sitea:~] esxcli network ip route ipv4 add -n 172.3.0.0/24 vmk0 172.3.0.1
[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:~]

[esxi02-siteb.rainpole.com - PuTTY]
[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
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:~] esxcli network ip route ipv4 add -n 147.80.0.0/24 vmk0 147.80.0.1
[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:~]

[witness-01.rainpole.com - PuTTY]
VM Network      0      0      vmnic0
Management Network 0      1      vmnic0
Switch Name     Num Ports Used Ports Configured Ports MTU      Uplinks
witnessSwitch  1536   4      1024          1500     vmnic1

PortGroup Name  VLAN ID Used Ports Uplinks
witnessPg       0      1          vmnic1

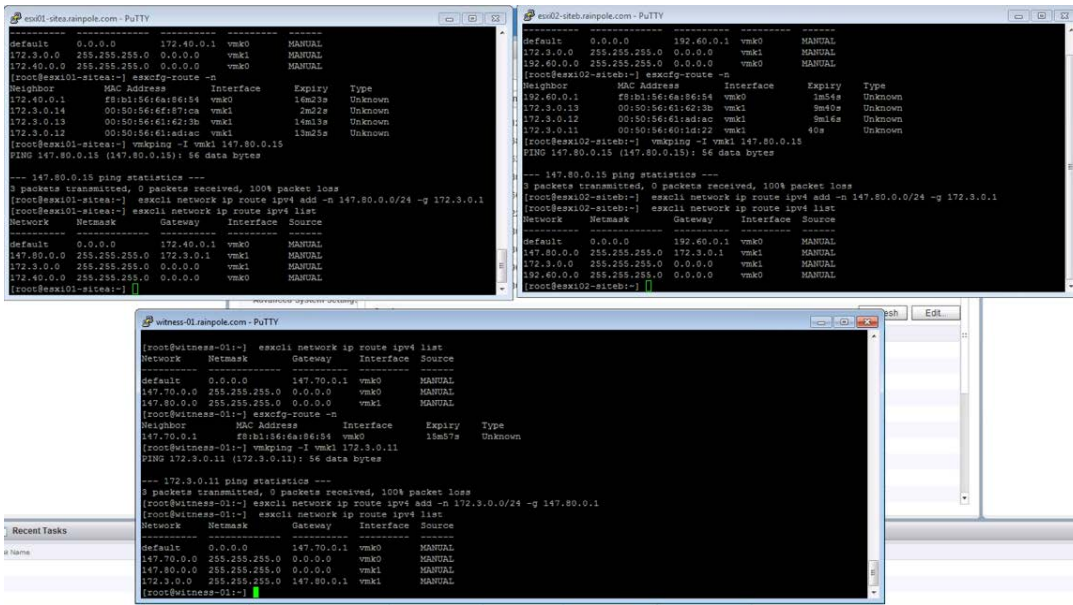
[root@witness-01:~] esxcli network ip route ipv4 list
Network      Netmask      Gateway      Interface    Source
-----
default      0.0.0.0      147.70.0.1   vmk0         MANUAL
147.70.0.0    255.255.255.0 0.0.0.0     vmk0         MANUAL
147.80.0.0    255.255.255.0 0.0.0.0     vmk1         MANUAL
[root@witness-01:~] esxcli network ip route ipv4 add -n 172.3.0.0/24 vmk1 172.3.0.1
[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:~]

```

To address this, administrators must implement static routes. Static routes, as highlighted previously, tell the TCPIP stack to use a different path to reach a particular network. Now we can tell the TCPIP stack on the data hosts to use a different network path (instead of the default gateway) to reach the VSAN network on the witness host. Similarly, we can tell the witness host to use an alternate path to reach the VSAN network on the data hosts rather than via the default gateway.

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.

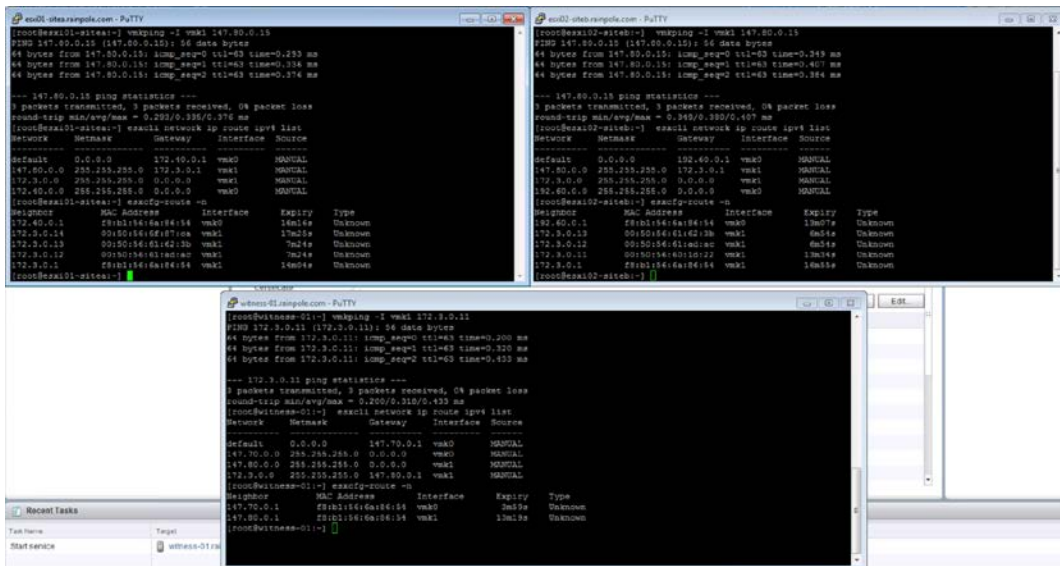
Virtual SAN Stretched Cluster Guide



The `esxcli` commands used to add a static route is:

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

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.



The final test is a ping test to ensure the remote VSAN network can now be reached, in both directions. Now the Virtual SAN Stretched Cluster can now be configured.

Configuring Virtual SAN Stretched Cluster

There are two methods to configure a Virtual SAN Stretched Cluster. A new cluster can be stretched or an existing cluster can be converted to a stretched cluster.

Creating a new Virtual SAN stretched cluster

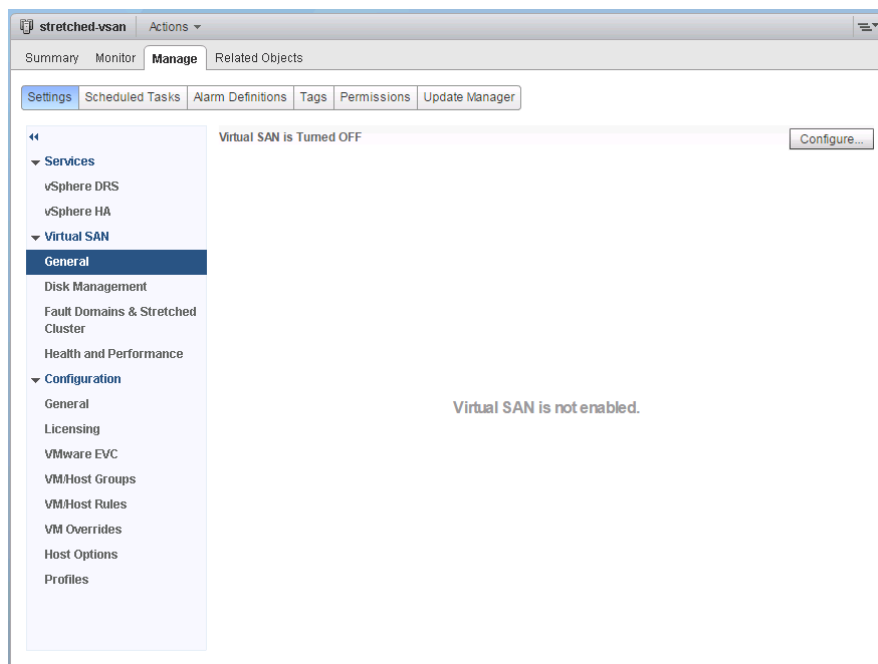
Creating a Virtual SAN stretched cluster from a group of hosts that doesn't already have Virtual SAN configured is relatively simple. A new Virtual SAN cluster wizard makes the process very easy.

Create Step 1: Create a Cluster

The following steps should be followed to install a new Virtual SAN stretched cluster. This example is a 3+3+1 deployment, meaning three ESXi hosts at the preferred site, three ESXi hosts at the secondary site and 1 witness host.

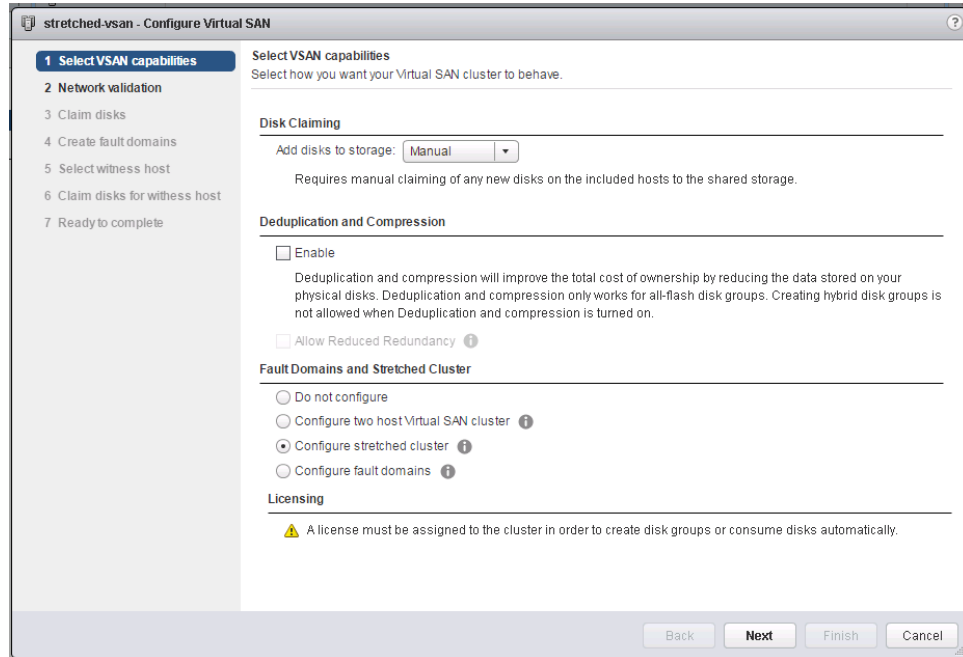
In this example, there are 6 nodes available: esx01-sitea, esx02-sitea, esx03-sitea, esx01-siteb, esx02-siteb, and esx03-siteb. All six hosts reside in a vSphere cluster called stretched-vsan. The seventh host witness-01, which is the witness host, is in its own datacenter and is not added to the cluster.

To setup Virtual SAN and configure stretch cluster navigate to the Manage > Virtual SAN > General. Click Configure to begin the Virtual SAN wizard.



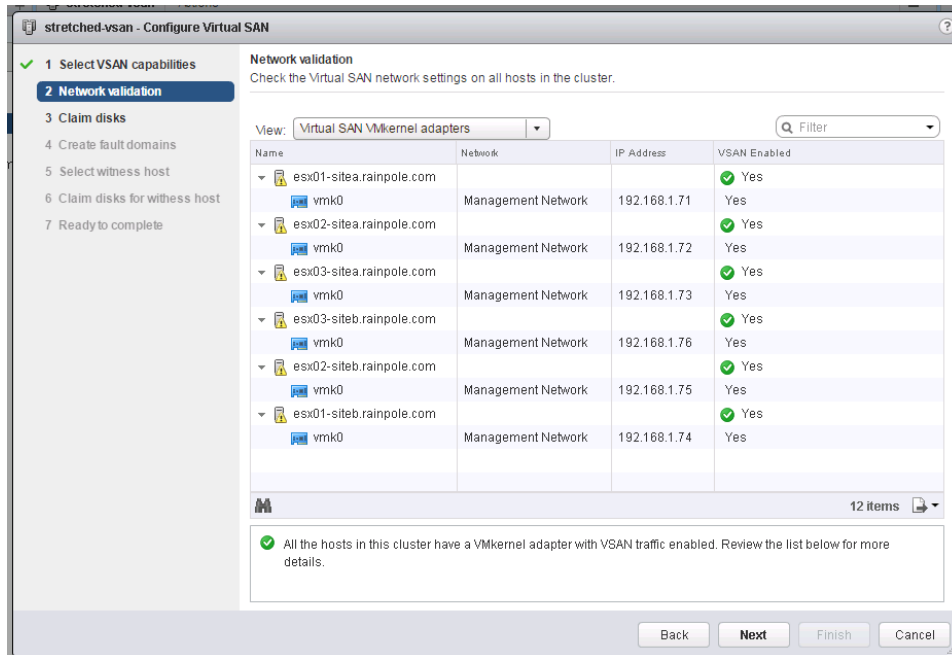
Create Step 2 Configure Virtual SAN as a stretched cluster

The initial wizard allows for choosing various options like disk claiming method, enabling Deduplication and Compression (All-Flash architectures only), as well as configuring fault domains or stretched cluster. Select **Configure stretched cluster**.



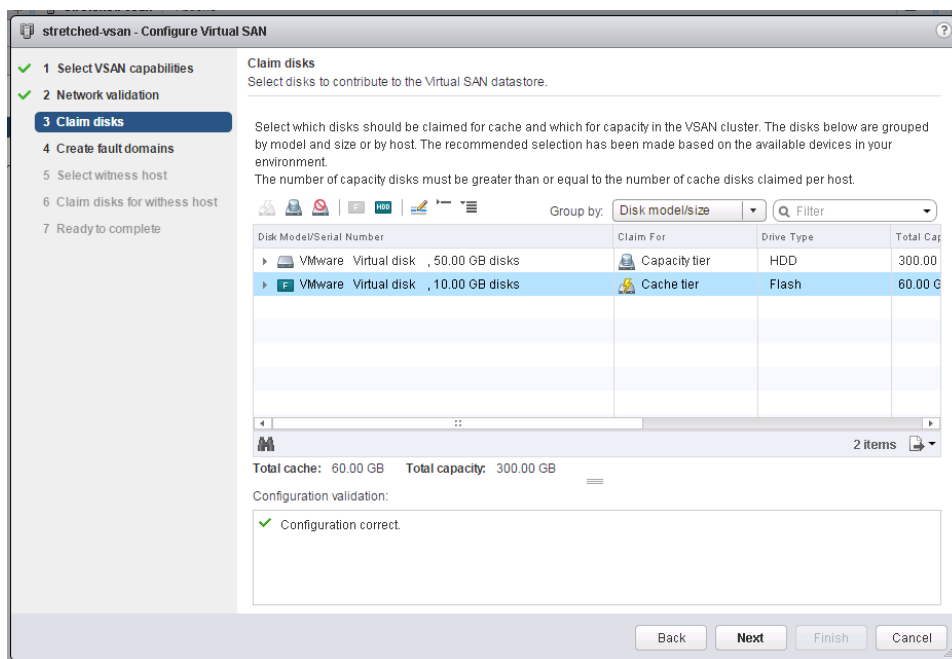
Create Step 3 Validate Network

Network validation will confirm that each host has a VMkernel interface with Virtual SAN traffic enabled. Select **Next**.



Create Step 4 Claim Disks

If the Claim Disks was set to Manual, disks should be selected for their appropriate role in the Virtual SAN cluster.

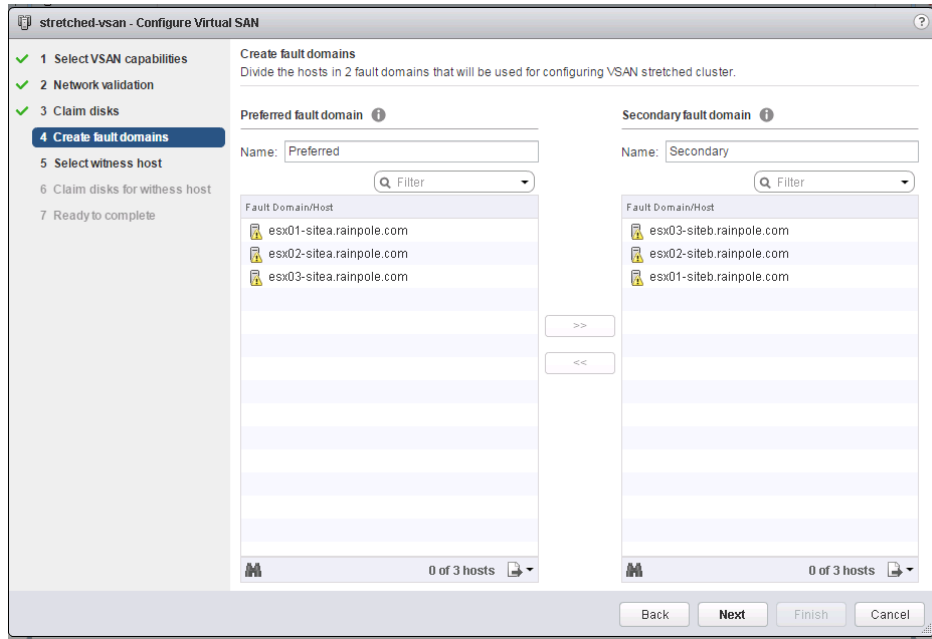


Select Next.

Create Step 5 Create Fault Domains

The Create fault domain wizard will allow hosts to be selected for either of the two sides of the stretched cluster. The default naming of these two fault

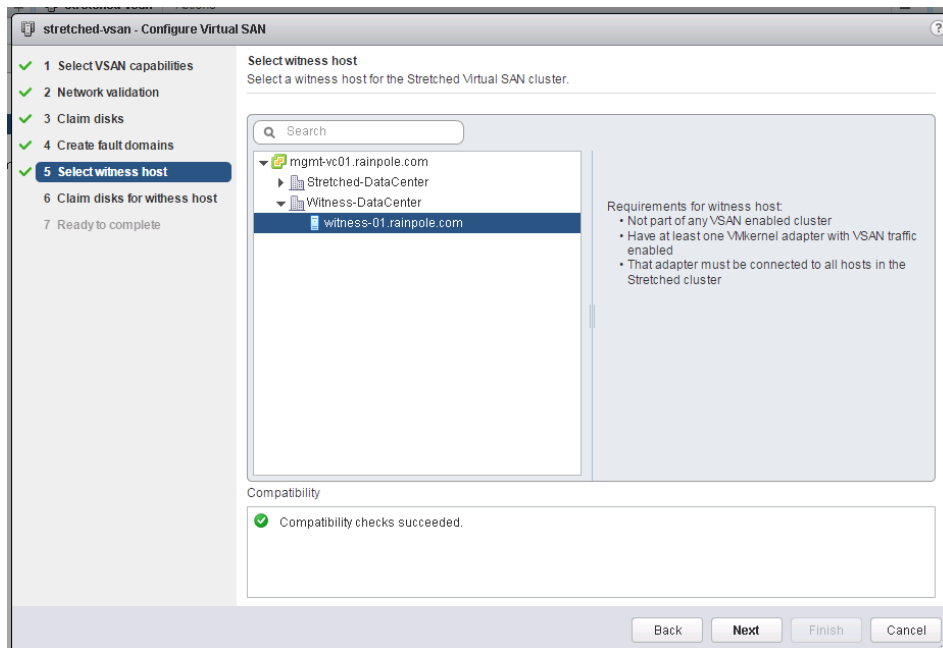
domains are Preferred and Secondary. These two fault domains will behave as two distinct sites.



Select Next.

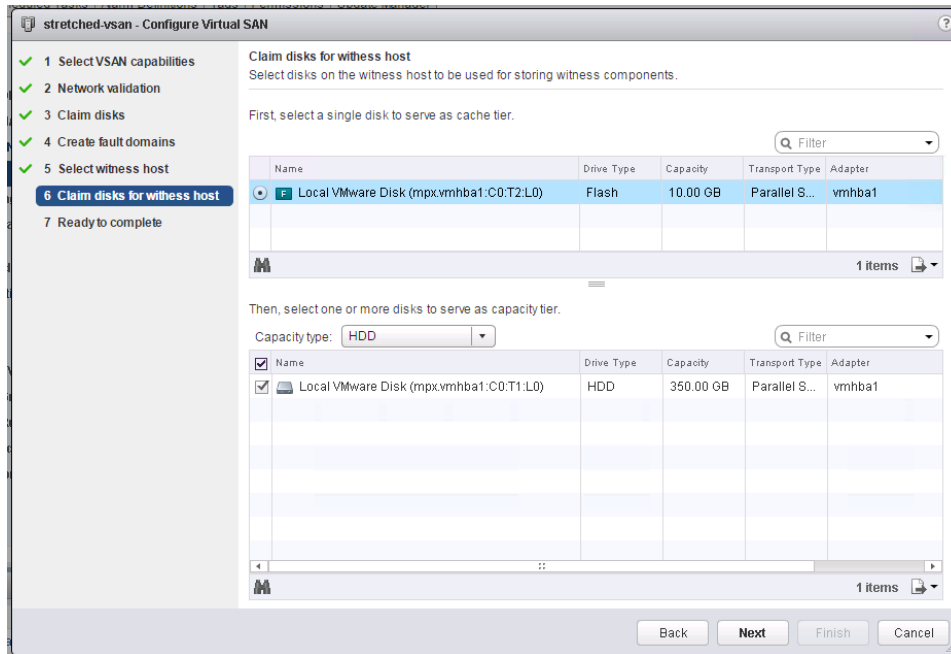
Create Step 6 Select witness host

The witness host detailed earlier must be selected to act as the witness to the two fault domains.



Create Step 7 Claim disks for witness host

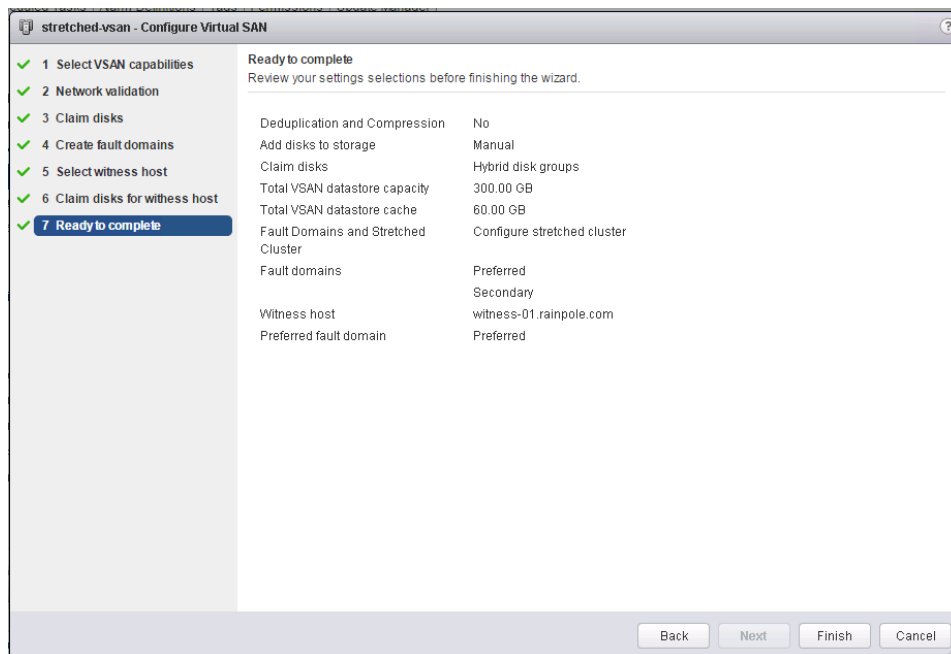
Just like physical Virtual SAN hosts, the witness needs a cache tier and a capacity tier. **Note: The witness does not actually require SSD backing and may reside on a traditional mechanical drive.*



Select Next.

Create Step 8 Complete

Review the Virtual SAN Stretched Cluster configuration for accuracy and select Finish.



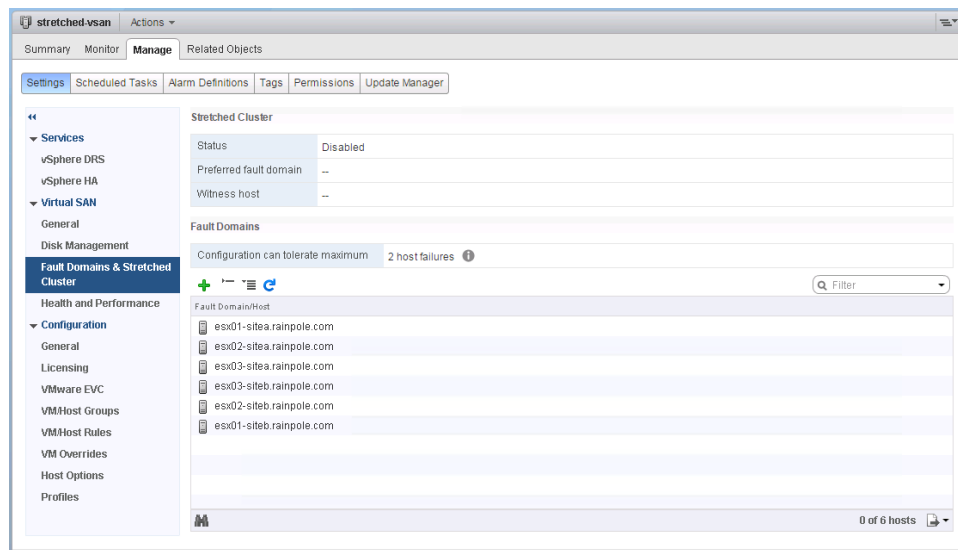
Converting an Existing Cluster to a Stretched Cluster

The following steps should be followed to convert an existing Virtual SAN cluster to a stretched cluster. This example is a 3+3+1 deployment, meaning three ESXi hosts at the preferred site, three ESXi hosts at the secondary site and 1 witness host.

Consider that all hosts are properly configured and Virtual SAN is already up and running.

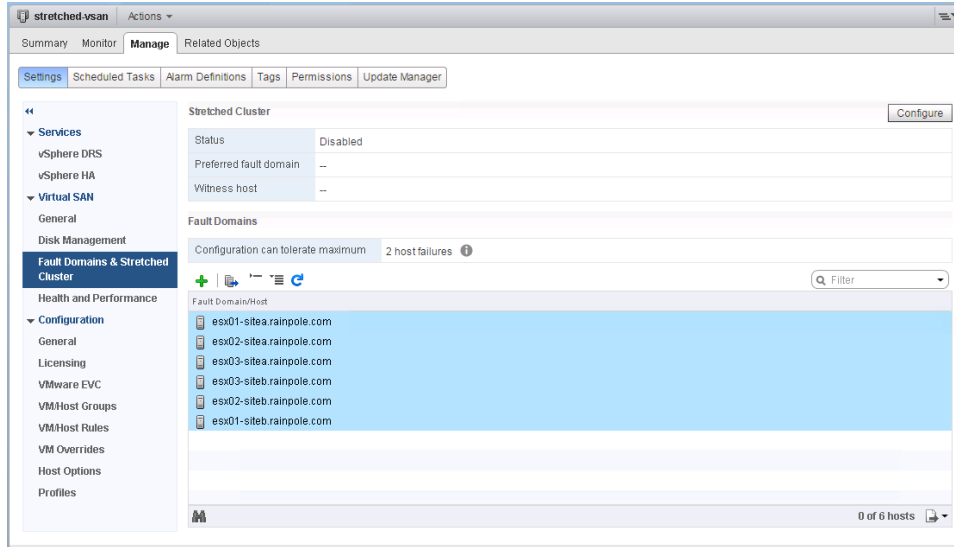
Convert Step 1 Fault Domains & Stretched Cluster

Configuring fault domains and stretched cluster settings is handled through the Fault Domains & Stretched Cluster menu item. Select > Manage > Settings > Fault Domains & Stretched Cluster.



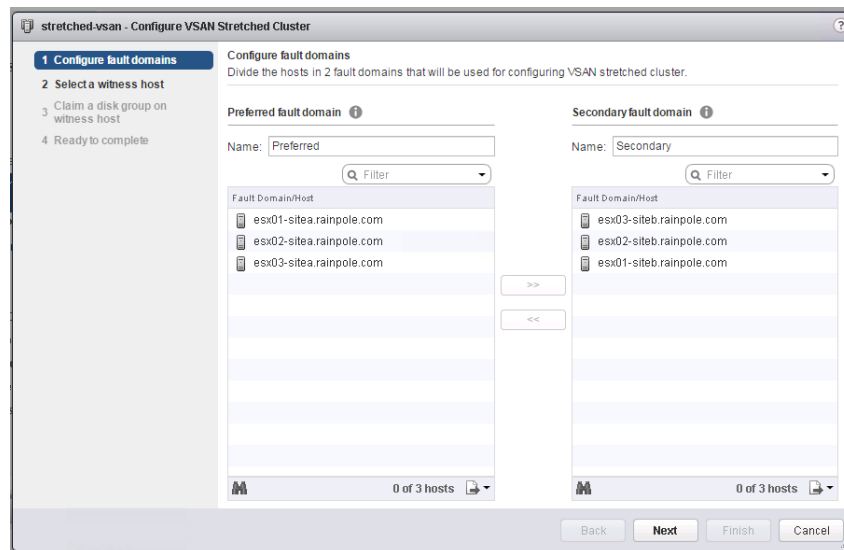
Convert Step 2 Selecting hosts to participate

Select each of the hosts in the cluster and select **Configure**.



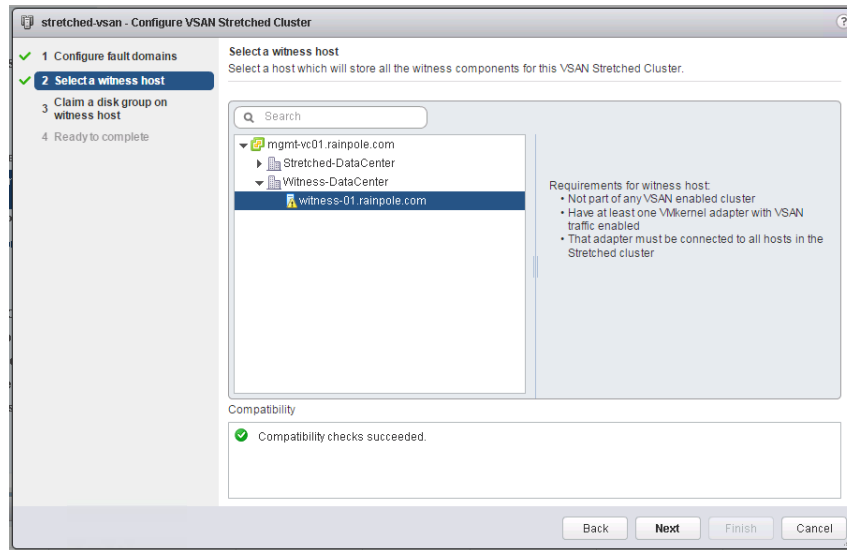
Convert Step 3 Configure fault domains

The Create fault domain wizard will allow hosts to be selected for either of the two sides of the stretched cluster. The default naming of these two fault domains are Preferred and Secondary. These two fault domains will behave as two distinct sites.



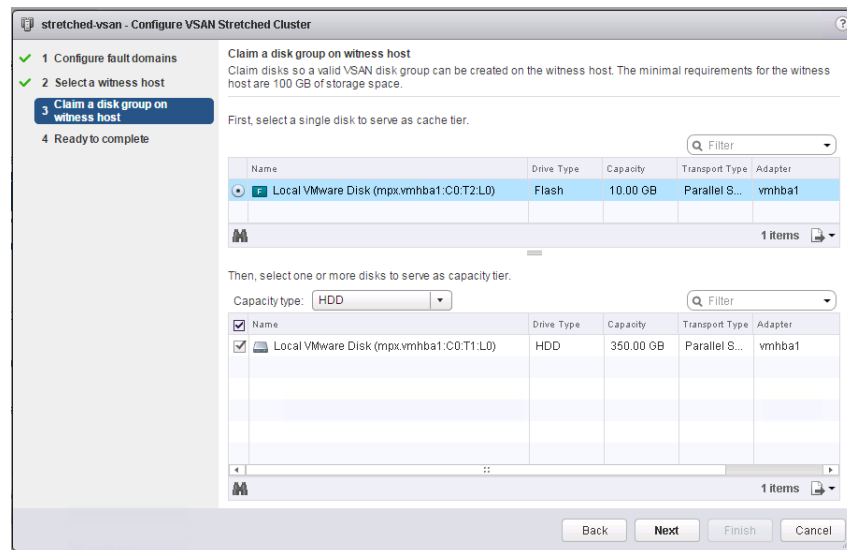
Convert Step 4 Select a witness host

The witness host detailed earlier must be selected to act as the witness to the two fault domains.



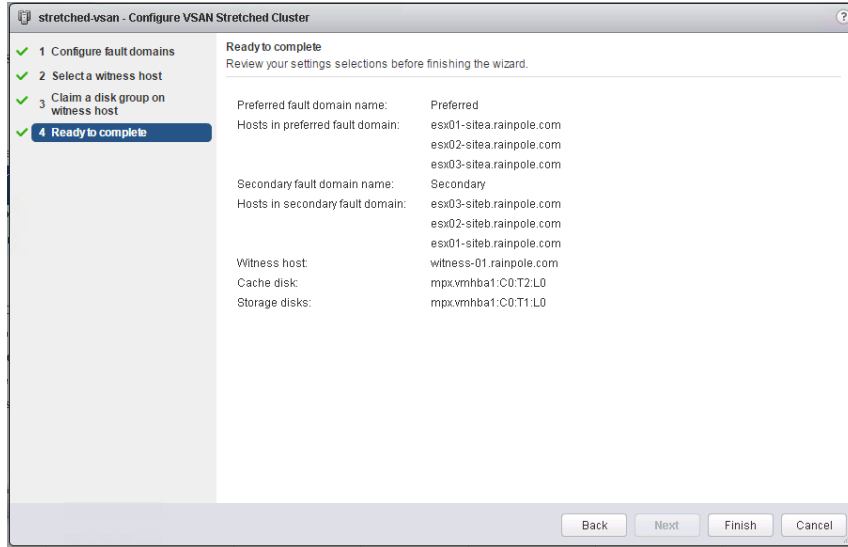
Convert Step 5 Claim disks for witness host

Just like physical Virtual SAN hosts, the witness needs a cache tier and a capacity tier. **Note: The witness does not actually require SSD backing and may reside on a traditional mechanical drive.*



Convert Step 6 Complete

Review the Virtual SAN Stretched Cluster configuration for accuracy and select Finish.

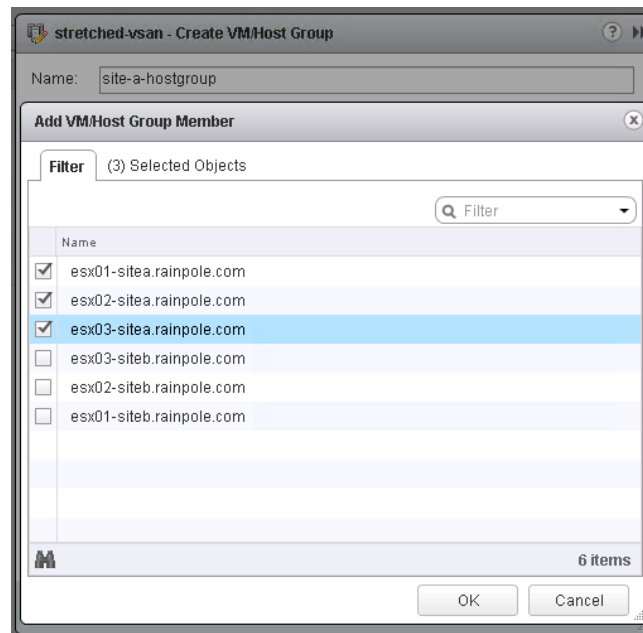


Configure stretched cluster site affinity

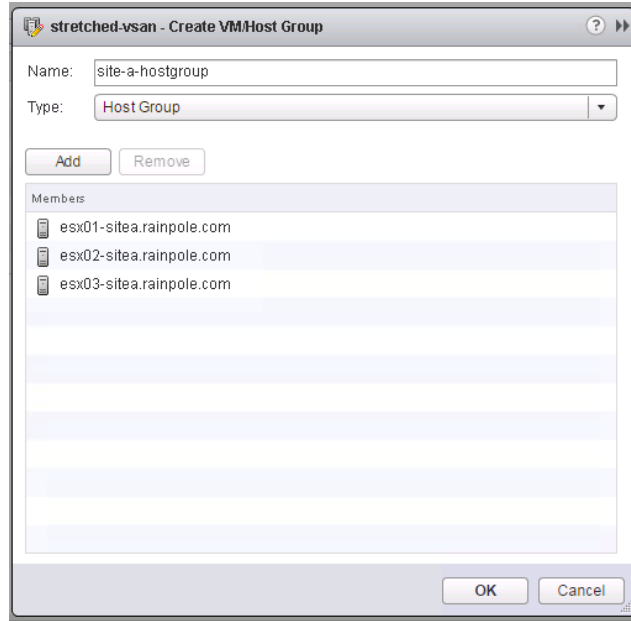
Configure Step 1 Create Host Groups

At this point, there needs to be a way of specifying which site a VM should be deployed to. 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 groups will contain the ESXi hosts from the preferred site whilst the second host group will contain the ESXi host from the secondary 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.

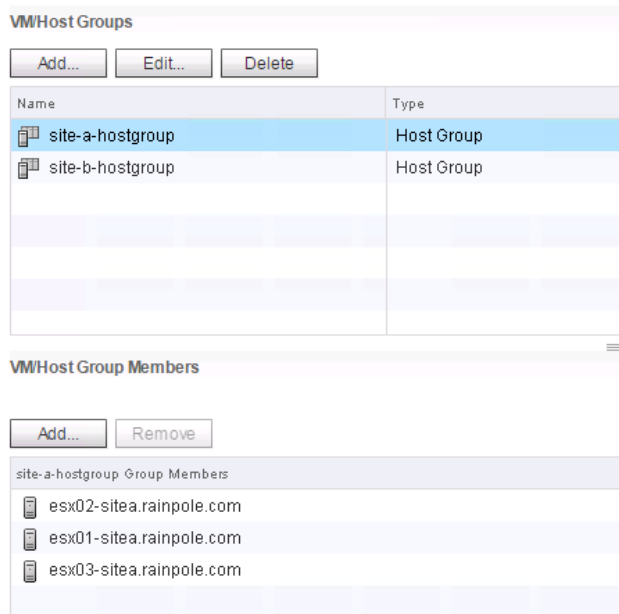
Navigate to cluster > Manage > VM/Host Groups. Select the option to “add” a group. Give the group a name, and ensure the group type is “Host Group” as opposed to “VM Group”. Next, click on the “Add” button to select the hosts should be in the host group. Select the hosts from site A.



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



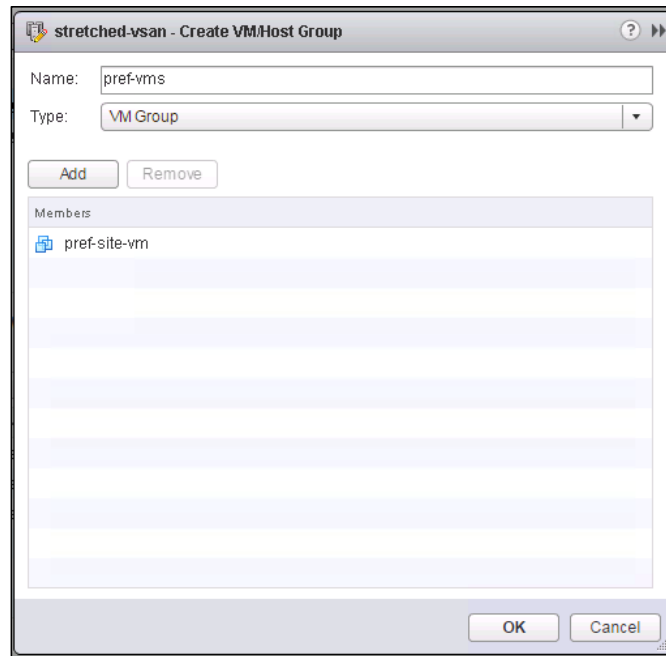
This step will need to be repeated for the secondary site. Create a host group for the secondary site and add the ESXi hosts from the secondary site to the host group.



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

Configure Step 2: Create VM Groups

Once the host groups are created, the initial set of virtual machines should now be created. Do not power on the virtual machines just yet. Once the virtual machines are in the inventory, you can now proceed with the creation of the VM Groups. First create the VM Group for the preferred site. Select the virtual machines that you want for the preferred site.



In the same way that a second host group had to be created previously for the secondary site, a secondary VM Group must be created for the virtual machines that should reside on the secondary site.

Configure Step 3: Create VM/Host Rules

Now that the host groups and VM groups are created, it is time to associate VM groups 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 sec-vms VM group with the host group called site-b-hostgroup, which will run the virtual machines in that group on the hosts in the secondary site.

The screenshot shows a dialog box titled "stretched-vsan - Create VM/Host Rule". It contains the following fields and options:

- Name:** sec-vm-hosts
- Enable rule.
- Type:** Virtual Machines to Hosts
- Description:** Virtual machines that are members of the Cluster VM Group sec-vms should run on host group site-b-hostgroup.
- VM Group:** sec-vms
- Host Group:** site-b-hostgroup
- Rule Type:** Should run on hosts in group (highlighted with a red box)

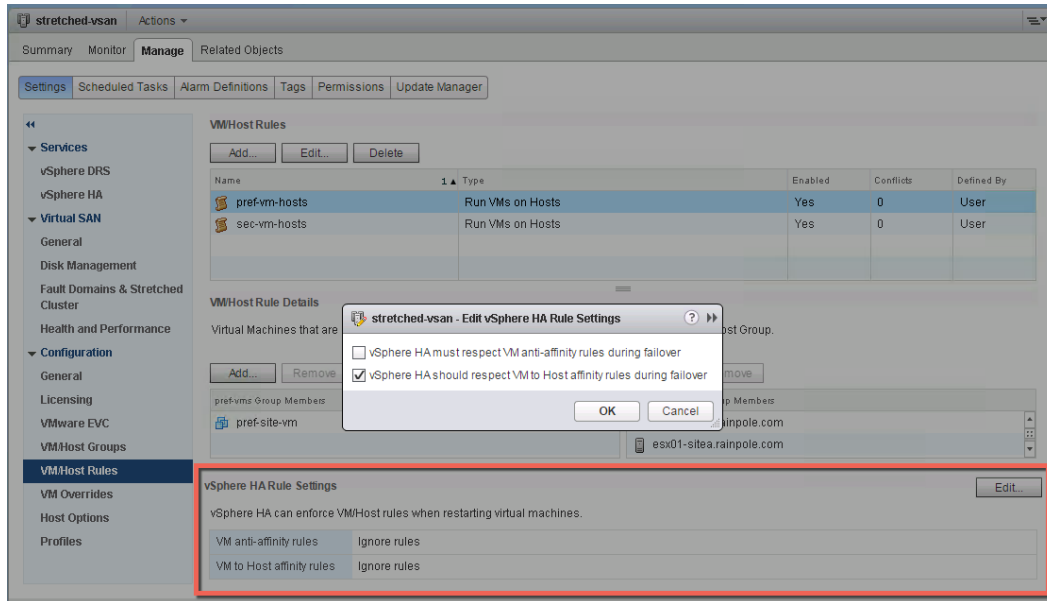
At the bottom right, there are "OK" and "Cancel" buttons.

One item highlighted above is that this is a “should” rule. We use a “should” rule as it allows vSphere HA to start the virtual machines on the other side of the stretched cluster in the event of a site failure.

Another VM/Host rule must be created for the primary site. Again this should be a “should” rule. Please note that 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.

Configure Step 4: Set vSphere HA rules

There is one final setting that needs to be placed on the VM/Host Rules. This setting once again defines how vSphere HA will behave when there is a complete site failure. In the screenshot below, there is a section in the VM/Host rules called vSphere HA Rule Settings. One of the settings is for VM to Host Affinity rules. A final step is to edit this from the default of “ignore” and change it to “vSphere HA should respect VM/Host affinity rules” as shown below:



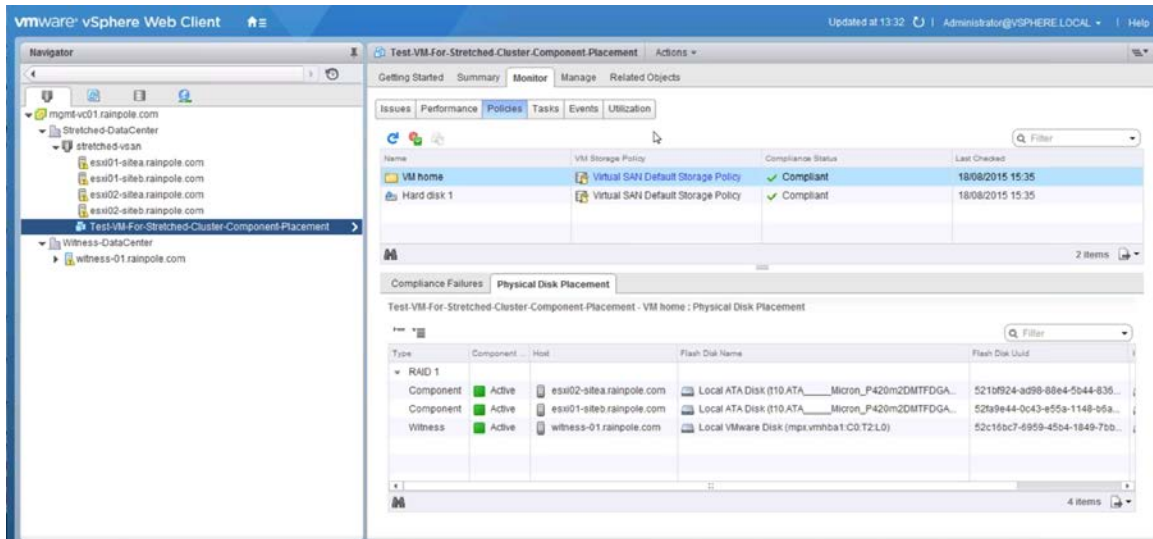
This setting can be interpreted as follows:

- If there are multiple hosts on either sites, and one hosts fails, vSphere HA will try to restart the VM on the remaining hosts on that site, maintained read affinity.
- If there is a complete site failure, then vSphere HA will try to restart the virtual machines on the hosts on the other site. If the “must respect” option shown above is selected, then vSphere HA would be unable to restart the virtual machines on the other site as it would break the rule. Using a “should” rule allows it to do just that.

Verifying Virtual SAN Stretched Cluster component layouts

That completes the setup of the Virtual SAN Stretched Cluster. The final steps are to power up the virtual machines created earlier, and examine the component layout. When *NumberOfFailuresToTolerate* = 1 is chosen, a copy of the data should go to both sites, and the witness should be placed on the witness host.

In the example below, *esx01-sitea* and *esx02-sitea* resides on site 1, whilst *esx01-siteb* and *esx02-siteb* resides on site 2. The host *witness-01* is the witness. The layout shows that the VM has been deployed correctly.



As we can clearly see, one copy of the data resides on storage in site1, a second copy of the data resides on storage in site2 and the witness component resides on the witness host and storage on the witness site. Everything is working as expected.

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

The witness configuration is not persisted. When recreating a stretched cluster VSAN, the witness will need to be re-configured. If you are using the same witness disk as before, the disk group will need to be deleted. This can only be done by opening an SSH session to the ESXi host, logging in as a privileged user and removing the disk group with the *esxcli vsan storage remove* command.

The fault domains are persisted, but VSAN does not know which FD is the preferred one. Therefore, under Fault Domains, the secondary FD will need to be moved to the secondary column as per of the reconfiguration.

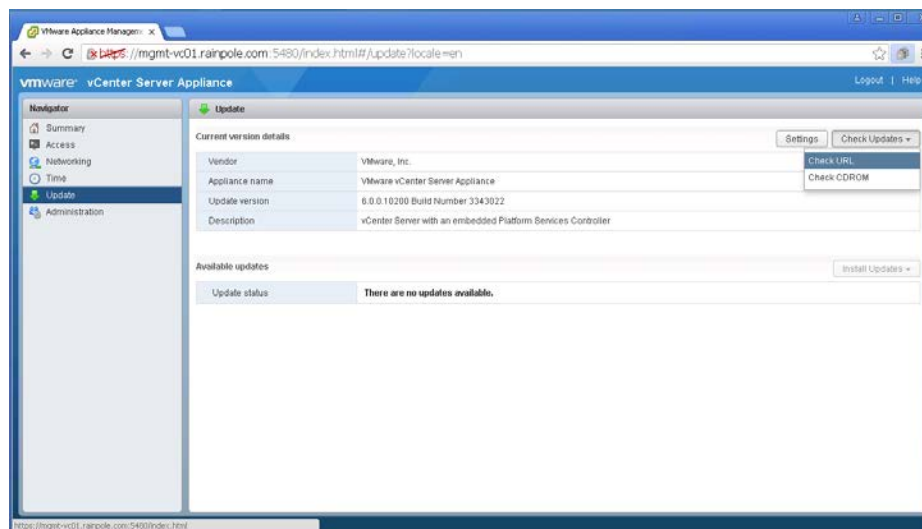
Upgrading a Virtual SAN 6.1 Stretched Cluster to Virtual SAN 6.2

Upgrading a Virtual SAN 6.1 Stretched Cluster is very easy. It is important though to follow a sequence of steps to ensure the upgrade goes smoothly.

Upgrading Step 1: Upgrade vCenter Server

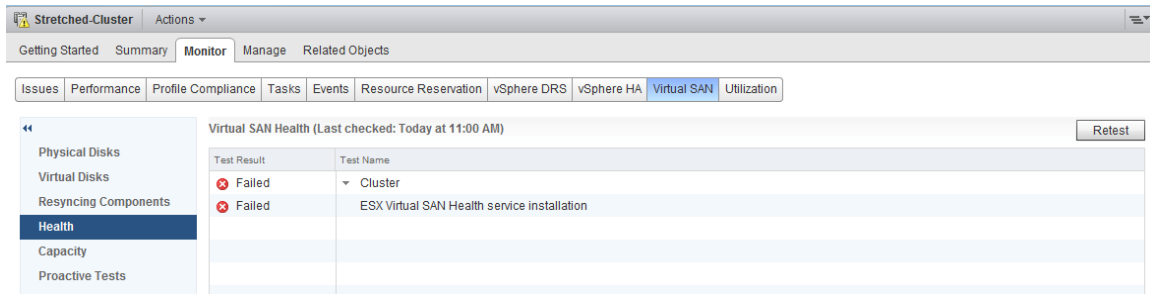
As with any vSphere upgrades, it is typically recommended to upgrade vCenter Server first. While vCenter Server for Windows installations are supported, the steps below describe the process when using the vCenter Server Appliance.

Log into the VAMI interface and select Update from the Navigator pane to begin the upgrade process.



*Refer to the documentation for vCenter Server for Windows to properly upgrade from vCenter Server 6.0 Update 1 to vCenter Server 6.0 Update 2.

After the upgrade has completed, the VCSA will have to be rebooted for the updates to be applied. It is important to remember that that Virtual SAN Health Check will not be available until after hosts have been upgraded to ESXi 6.0 Update 2.



Upgrading Step 2: 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.

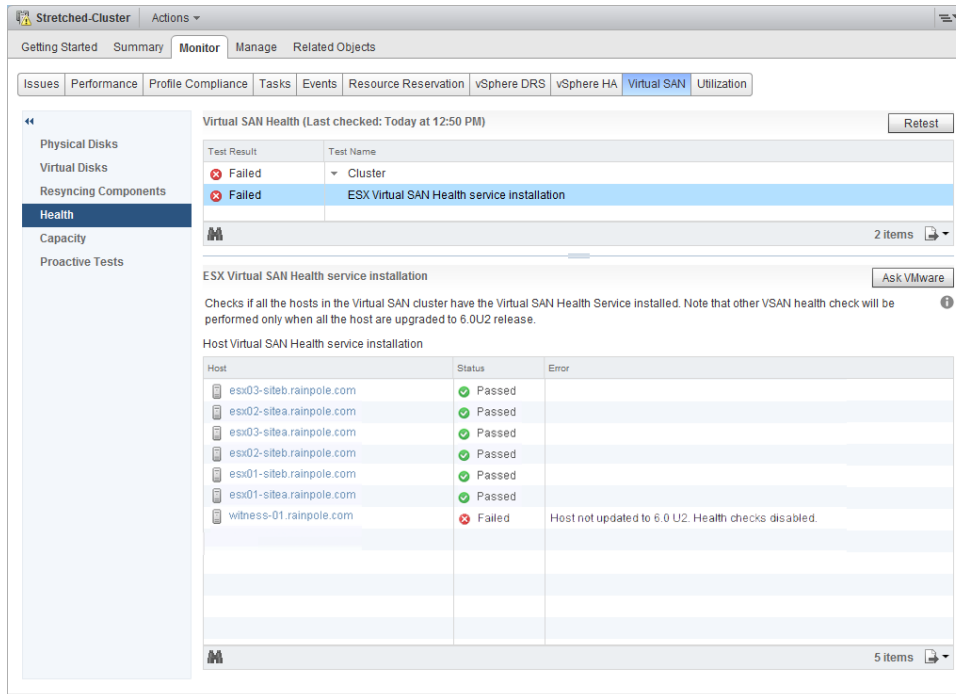
As with any upgrade, hosts will be required to be put in maintenance mode, remediated, upgraded, and rebooted. It is important to consider the amount of available capacity at each site. In sites that have sufficient available capacity, it would be desirable to choose the “full data migration” Virtual SAN 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.

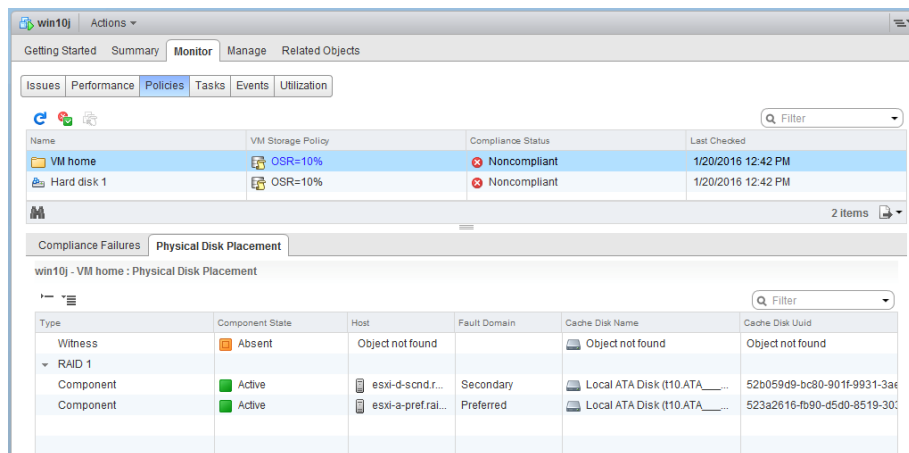
It is recommended to sequentially upgrade hosts at each site first, followed by sequentially upgrading the hosts at the alternate site. This method will introduce the least amount of additional storage traffic. While it is feasible to upgrade multiple hosts simultaneously across sites when there is additional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

Upgrading Step 3: Upgrade the witness appliance

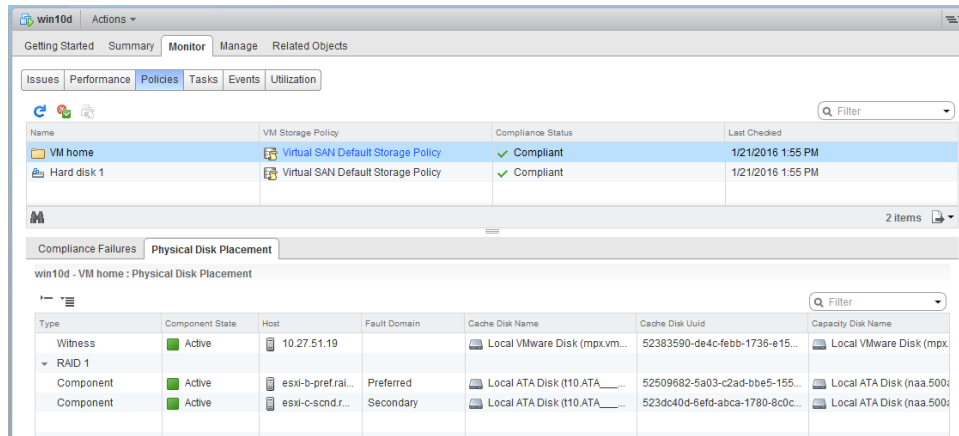
After both data sites have been upgraded, the witness will also need to be upgraded. The Health Check will show that the witness has not been upgraded to vSphere 6.0 Update 2.



Upgrading the witness is done in the same way that physical ESXi hosts are updated. It is important to remember that before the witness is upgraded, the witness components will no longer be available and objects will be noncompliant. Objects will report that the witness component is not found.

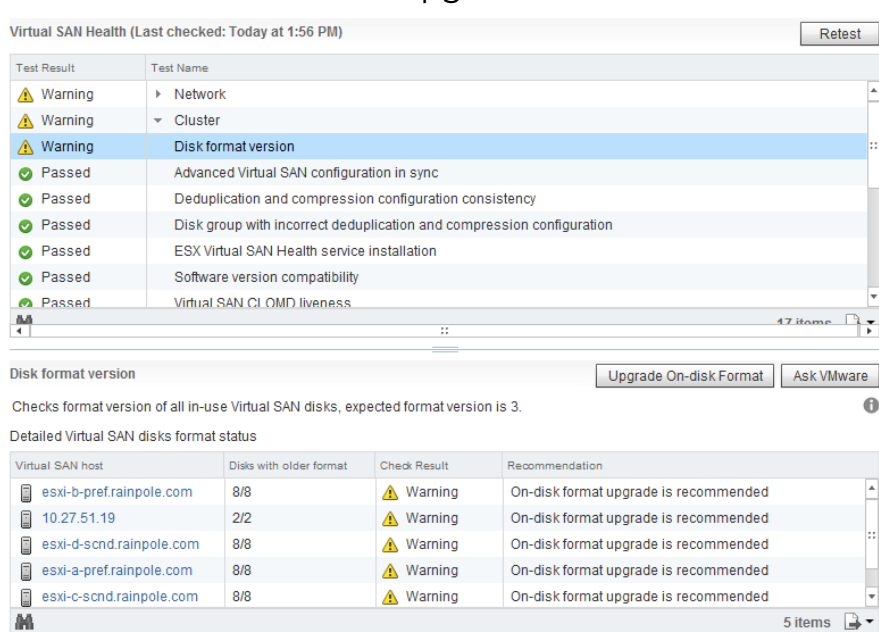


After the upgrade is complete, the witness component will return, and will reside on the witness.

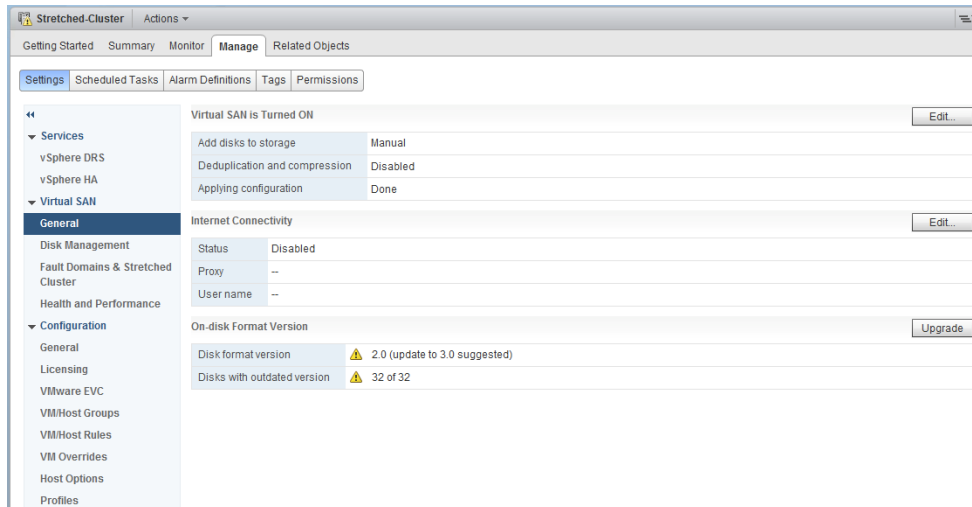


Upgrading Step 4: Upgrade the on-disk format

The final step in the upgrade process, will be to upgrade the on-disk format. To use the new features of Virtual SAN 6.2 like deduplication and compression or checksum, the on-disk format must be upgraded to version 3.0. The Health Check will assume that vSphere 6.0 Update 2 hosts would prefer a version 3.0 on-disk format, and as a result, it will throw an error until the format is upgraded.



To upgrade the on-disk format from version 2.0 to version 3.0, select Manage > then General under Virtual SAN. Then select Upgrade under the On-disk Format Version section.



The on-disk format upgrade will perform a rolling upgrade across the hosts in the cluster within each site. Each host's disk groups will be removed and recreated with the new on-disk format. The amount of time required to complete the on-disk upgrade will vary based on the cluster hardware configuration, how much data is on the cluster, and whatever over disk operations are occurring on the cluster. The on-disk rolling upgrade process does not interrupt virtual machine disk access and is performed automatically across the cluster.

The witness components residing on the witness appliance will be deleted and recreated. This process is relatively quick given the size of witness objects.

Once the on-disk format is complete, the cluster has been upgraded to Virtual SAN 6.2.

Management and Maintenance

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

Maintenance Mode Consideration

When it comes to maintenance mode in the Virtual SAN Stretched Cluster configuration, there are two scenarios to consider; maintenance mode on a site host and maintenance mode on the witness host.

Maintenance mode on a site host

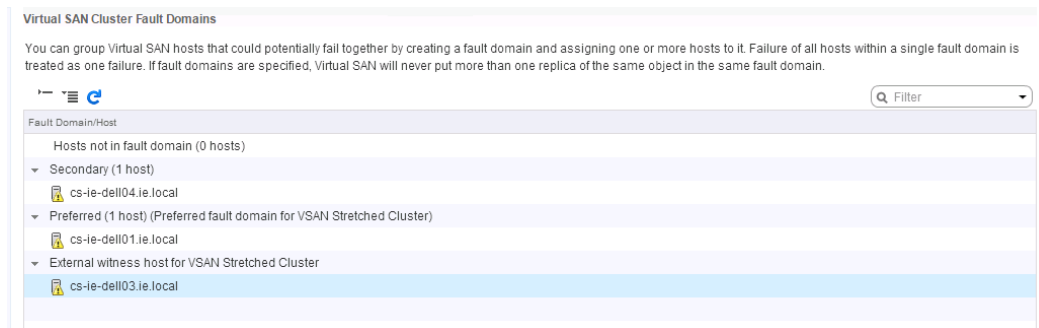
Maintenance mode in Virtual SAN Stretched Clusters is site specific. All maintenance modes (Ensure Accessibility, Full data migration and No data migration) are all supported. However, in order to do a Full data migration, you will need to ensure that there are enough resources in the same site to facilitate the rebuilding of components on the remaining node on that site.

Maintenance mode on the witness host

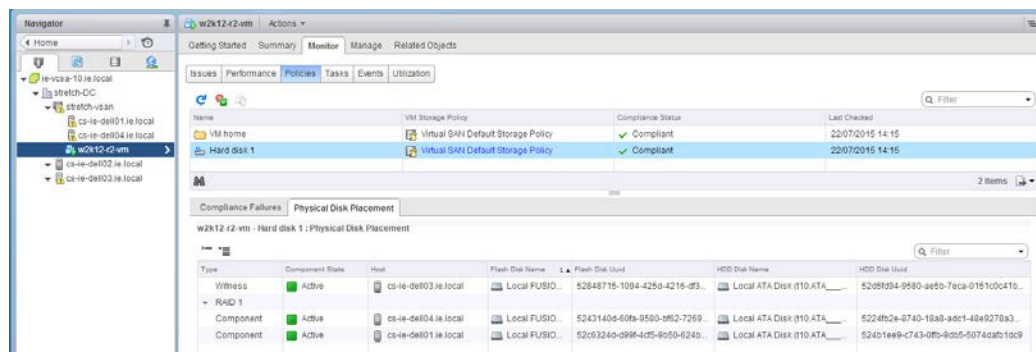
Maintenance mode on the witness host should be an infrequent event, as it does not run any virtual machines. Maintenance mode on the witness host only supports the *No data migration* option. Users check that all virtual machines are in compliance and that there is no ongoing failure before doing maintenance on the witness

Failure Scenarios

In this section, we will discuss the behavior of the Virtual SAN Stretched Cluster when various failures occur. In this example, there is a 1+1+1 Stretched VSAN deployment. This means that there is a single data host at site 1, a single data host at site 2 and a witness host at a third site.



A single VM has also been deployed. When the Physical Disk Placement is examined, we can see that the replicas are placed on the preferred and secondary data site respectively, and the witness component is placed on the witness site:



The next step is to introduce some failures and examine how Virtual SAN handles such events. Before beginning these tests, please ensure that the Virtual SAN Health Check Plugin is working correctly, and that all VSAN Health Checks have passed.

Note: In a 1+1+1 configuration, a single host failure would be akin to a complete site failure.

The health check plugin should be referred to regularly during failure scenario testing. Note that alarms are now raised in version 6.1 for any health check that fails. Alarms may also be reference at the cluster level throughout this testing.

Finally, when the term site is used in the failure scenarios, it implies a fault domain.

How read locality is established after failover to other site?

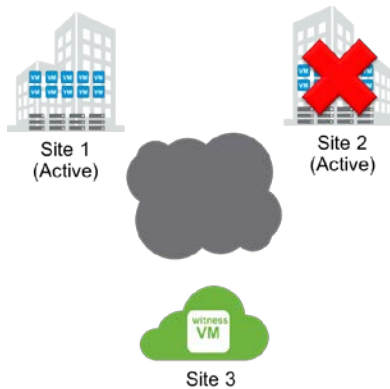
A common question is how read locality is maintained when there is a failover. This guide has already described read locality, and how in a typical Virtual SAN deployment, a virtual machine reads equally from all of its replicas in a round-robin format. In other words, if a virtual machine has two replicas as a result of being configured to tolerate one failure, 50% of the reads come from each replica. This algorithm has been enhanced for stretch clusters so that 100% of the reads comes from the local storage 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.

In the event of a failure or maintenance event, the virtual machine is restarted on the remote site. 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 levels of performance.

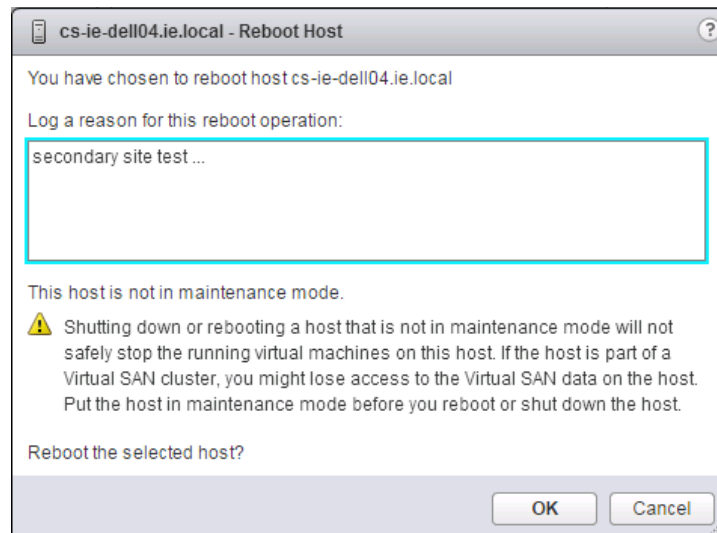
When the virtual machine starts on the other site, either as part of a vMotion operation or a power on from vSphere HA restarting it, Virtual SAN instantiates the in-memory state for all the objects of said virtual machine on the host where it moved. That includes the “owner” (coordinator) logic for each object. The owner checks if the cluster is setup in a “stretch cluster” mode, and if so, which fault domain it is running in. It then uses the 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 that is on the same site (but not necessarily the same host) as the virtual machine.

Single data host failure – Secondary site

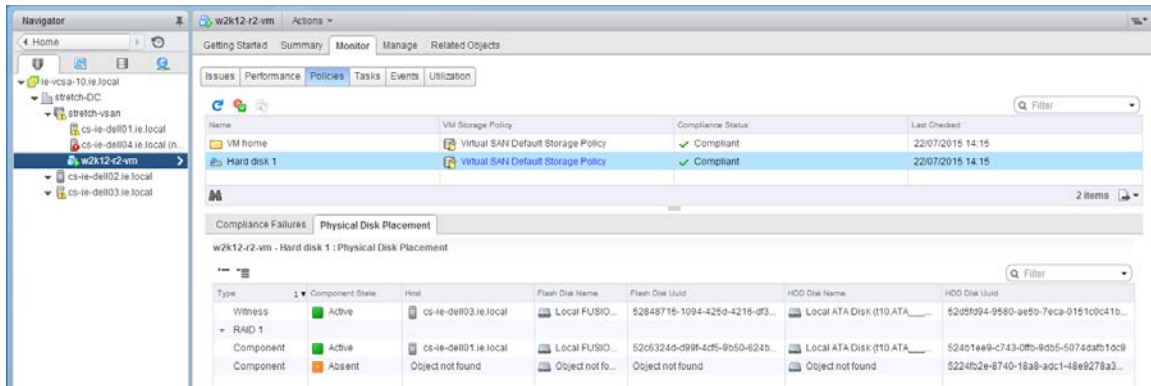
The first test is to introduce a failure on a host on one of the data sites, either the “preferred” or the “secondary” site. The sample virtual machine deployed for test purposes currently resides on the preferred site.



In the first part of this test, the secondary host will be rebooted, simulating a temporary outage.



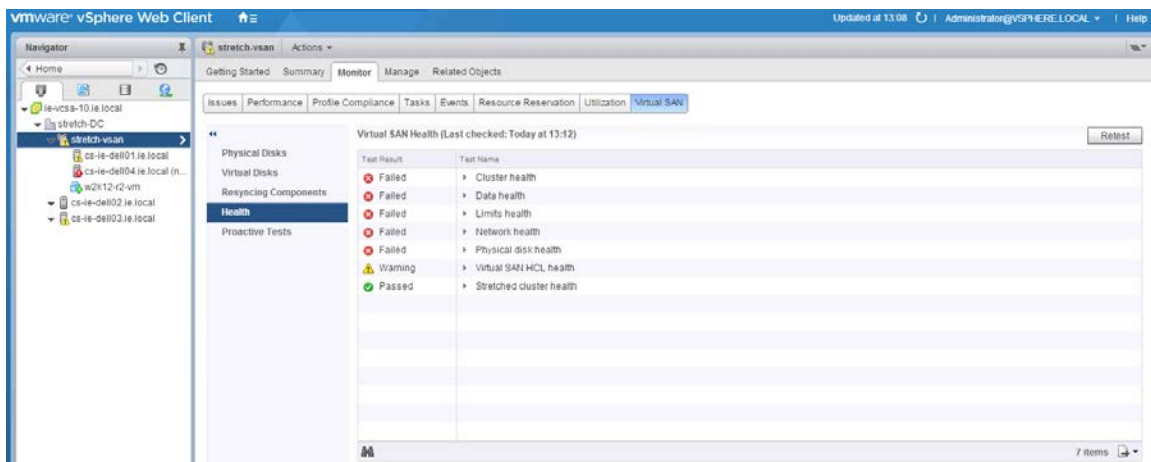
There will be some power and HA events related to the secondary host visible in the vSphere web client UI. Change to the Physical Disk Place view of the virtual machine. After a few moments, the components that were on the secondary host will go “Absent”, as shown below:



However, the virtual machine continues to be accessible. This is because there is a full copy of the data available on the host on the preferred site, and there are more than 50% of the components available. Open a console to the virtual machine and verify that it is still very much active and functioning.

Since the ESXi host which holds the compute of the virtual machine is unaffected by this failure, there is no reason for vSphere HA to take action.

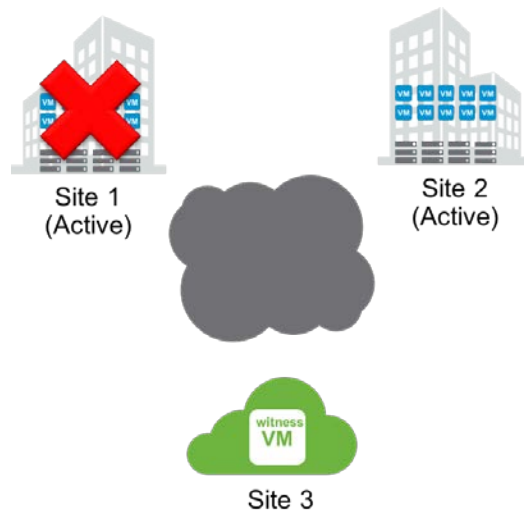
At this point, the VSAN Health Check plugin can be examined. There will be quite a number of failures due to the fact that the secondary host is no longer available, as one might expect.



Further testing should not be initiated until the secondary host has completed a reboot and has successfully rejoined the cluster. All "Failed" health check tests should show OK before another test is started. Also confirm that there are no "Absent" components on the VMs objects, and that all components are once again Active.

Single data host failure - Preferred site

This next test will not only check VSAN, but it will also verify vSphere HA functionality, and that the VM to Host affinity rules are working correctly. If each site has multiple hosts, then a host failure on the primary site will allow vSphere HA to start the virtual machine on another host on the same site. In this test, the configuration is 1+1+1 so the virtual machine will have to be restarted on the secondary site. This will also verify that the VM to Host affinity “should” rule is working.



A reboot can now be initiated on the preferred host. There will be a number of vSphere HA related events. As before, the components that were on the preferred host will show up as “Absent”:

The screenshot shows the vSphere Web Client interface for a VM named 'w2k12-r2-vm'. The 'Physical Disk Placement' tab is active, showing a table of components and their status. The table has columns for Type, Component State, Host, Flash Disk Name, Flash Disk UUID, HDD Disk Name, and HDD Disk GUID.

Type	Component State	Host	Flash Disk Name	Flash Disk UUID	HDD Disk Name	HDD Disk GUID
Witness	Active	cs-1e-dell03.l...	Local FUSI...	52848715-1094-4250-4216-df3...	Local ATA Disk (110.ATA_...	52d5f094-9580-ae50-7eca-0151...
RAID 1	Active	cs-1e-dell04.l...	Local FUSI...	5243148d-60fa-9580-bf62-7269...	Local ATA Disk (110.ATA_...	5224fb2e-8740-18a8-adc1-48e9...
Component	Absent	Object not found	Object not f...	Object not found	Object not found	52401ee9-c743-0fb-9db5-5074d...

Since the host on which the virtual machine’s compute resides is no longer available, vSphere HA will restart the virtual machine on another the host in the cluster. This will verify that the vSphere HA affinity rules are “should” rules and not “must” rules. If “must” rules are configured, this will not allow vSphere HA to restart the virtual machine on the other site, so it is important that this

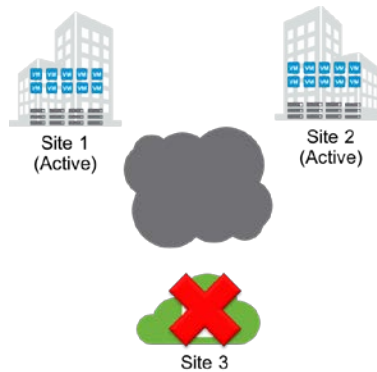
test behaves as expected. “Should” rules will allow vSphere HA to restart the virtual machine on hosts that are not in the VM/Host affinity rules when no other hosts are available.

Note that if there were more than one host on each site, then the virtual machine would be restarted on another host on the same site. However, since this is a test on a 1+1+1 configuration, there are no additional hosts available on the preferred site. Therefore the virtual machine is restarted on a host on the secondary site after a few moments. If you are testing this behavior on

As before, wait for all issues to be resolved before attempting another test. Remember: test one thing at a time. Allow time for the secondary site host to reboot and verify that all absent components are active, and that all health check tests pass before continuing.

Single witness host failure - Witness site

This is the final host failure test. In this test, the witness host (or virtual machine depending on the implementation) will be rebooted, simulating a failure on the witness host.



This should have no impact on the run state of the virtual machine, but the witness components residing on the witness host should show up as “Absent”.

First, verify which host is the witness host from the fault domains configuration. In this setup, it is host `cs-ie-dell03.ie.local`. It should be labeled “External witness host for Virtual SAN Stretched Cluster”.

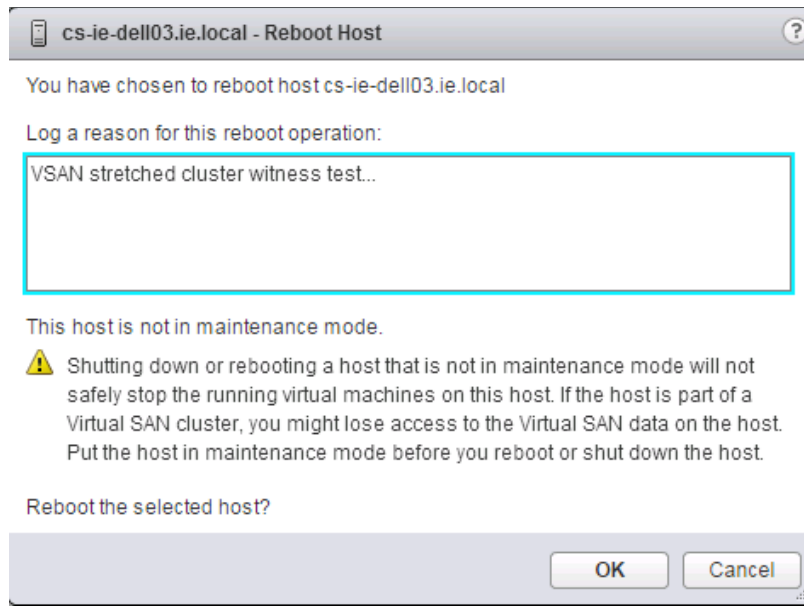
Virtual SAN Cluster Fault Domains

You can group Virtual SAN hosts that could potentially fail together by creating a fault domain and assigning one or more hosts to it. Failure of all hosts within a single fault domain is treated as one failure. If fault domains are specified, Virtual SAN will never put more than one replica of the same object in the same fault domain.

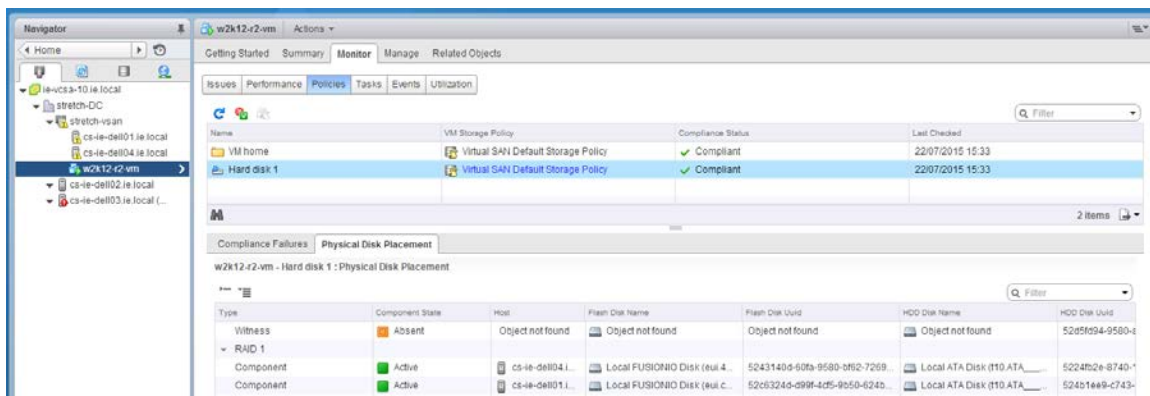
Filter

Fault Domain/Host
Hosts not in fault domain (0 hosts)
Secondary (1 host)
cs-ie-dell04.ie.local
Preferred (1 host) (Preferred fault domain for VSAN Stretched Cluster)
cs-ie-dell01.ie.local
External witness host for VSAN Stretched Cluster
cs-ie-dell03.ie.local

After verifying that there are no absent components on the virtual machine, and that all health checks have passed, reboot the witness host:



After a short period of time, the witness component of the virtual machine will appear as “Absent”:



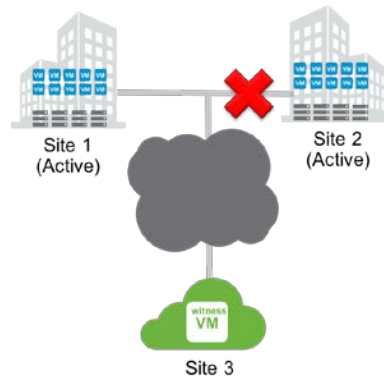
The virtual machine is unaffected and continues to be available and accessible.

Rule for virtual machine object accessibility: at least one full copy of the data must be available, and more than 50% of the components that go to make up the object are available.

Wait for the witness host to reboot. Verify that all virtual machine components are Active and that all of the Virtual SAN Health Checks pass before continuing with further testing.

Network failure - Data Site to Data Site

Before beginning this test, please revisit the vSphere HA configuration settings and ensure that the Host Isolation Response and Host Isolation address are configured correctly. Also, if there are non-VSAN datastores in your environment (NFS, VMFS), ensure that Datastore Heartbeats are disabled using the instructions earlier in this guide. As before, this test is on a 1+1+1 Virtual SAN Stretched Cluster. This test will simulate a network failure on a data site with a running virtual machine.



To test this functionality, there are various ways to cause it. One could simply unplug the VSAN network from the host or indeed the switch. Alternatively, the physical adapter(s) used for VSAN traffic are moved from active to “unused” for the VSAN VMkernel port on the host running the virtual machine. This can be done by editing the “Teaming and failover” properties of the VSAN traffic port group on a per host basis. In this case, the operation is done on a host on the “preferred” site. This results in two components of the virtual machine object getting marked as absent since the host can no longer communicate to the other data site where the other copy of the data resides, nor can it communicate to the witness.

The screenshot shows the vSphere Web Client interface. The left sidebar displays the navigation tree with the VM 'w2k12-r2-vm' selected. The main content area shows the 'Physical Disk Placement' table for the VM's 'Hard disk 1'.

Type	Component State	Host	Flash Disk Name	Flash Disk UUID	HDD Disk Name	HDD Disk UUID
Witness	Absent	cs-1e-d803.l...	Local FUSION#0 Disk (eul.d...	52848715-1094-425d-4216-df3...	Local ATA Disk (110.ATA_...	52d5f994-9588-4...
RAID 1						
Component	Absent	cs-1e-d804.l...	Local FUSION#0 Disk (eul.4...	5243140d-609b-958b-9852-7269...	Local ATA Disk (110.ATA_...	5224f2e-874b-*
Component	Active	cs-1e-d801.l...	Local FUSION#0 Disk (eul.c...	52c6324d-99f4-4c5f-9e50-024c...	Local ATA Disk (110.ATA_...	524b1ee8-c743-

From a vSphere HA perspective, since the host isolation response IP address is on the VSAN network, it should not be able to reach the isolation response IP address. The console to the virtual machine is also inaccessible at this time.

Note: *Simply disabling the VSAN network service will not lead to an isolation event since vSphere HA will still be able to use the network for communication.*

This isolation state of the host is a trigger for vSphere HA to implement the isolation response action, which has previously been configured to “*Power off VMs and restart*”. The virtual machine should then power up on the other site. If you navigate to the policies view after the virtual machine has been restarted on the other host, and click on the icon to check compliance, it should show that two out of the three components are now available, and since there is a full copy of the data, and more than 50% of the components available, the virtual machine is accessible. Launch the console to verify.

Note: *It would be worth including a check at this point to ensure that the virtual machine is accessible on the VM network on the new site. There is not much in having the virtual machine failover to the remaining site and not being able to reach it on the network.*

Remember that this is a simple 1+1+1 configuration of Virtual SAN Stretched Cluster. If there were additional hosts on each site, the virtual machine should be restarted on hosts on the same site, adhering to the VM/Host affinity rules defined earlier. Because the rules are “should” rules and not “must” rules, the virtual machine can be restarted on the other site when there are no hosts available on the site to which the virtual machine has affinity.

Once the correct behavior has been observed, repair the network.

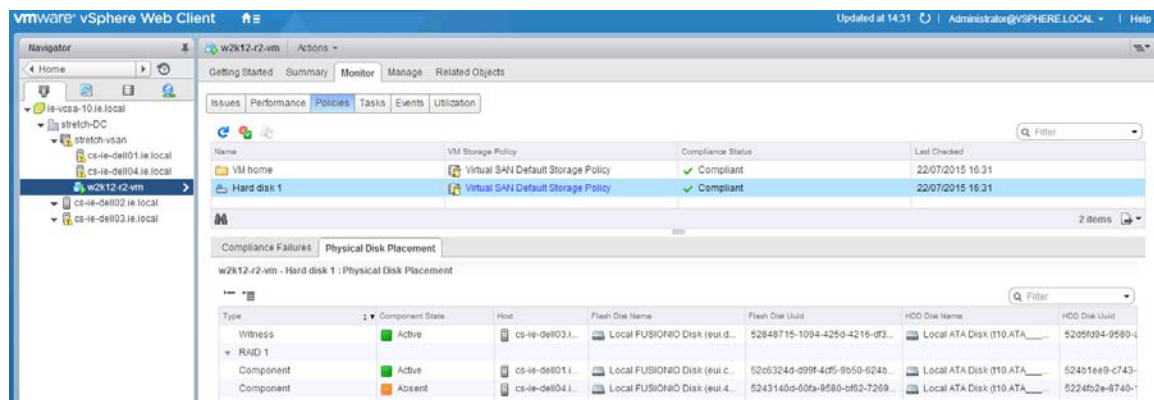
Note that the VM/Host affinity rules will trigger a move of the virtual machine(s) back to hosts on the preferred site. Run a VSAN Health Check test before continuing to test anything else. Remember with *NumberOfFailuresToTolerate* = 1, test one thing at a time. Verify that all absent components are active and that all health check tests pass before continuing.

Data network test with multiple ESXi hosts per site

If there is more than one host at each site, you could try setting the uplinks for the VSAN network to “unused” on each host on one site. What you should observe is that the virtual machine(s) is restarted on another host on the same site to adhere to the configured VM/Host affinity rules. Only when there is no remaining host on the site should the virtual machine be restarted on the other site.

Data network test on host that contains virtual machine data only

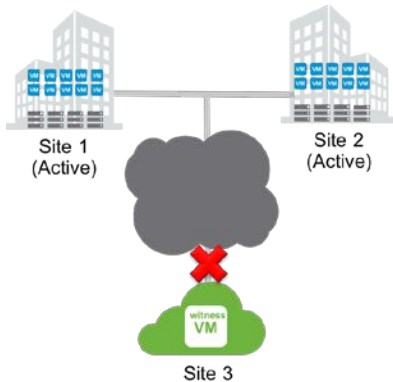
If the network is disabled on the ESXi host that does not run the virtual machine but contains a copy of the data, then the virtual machines on the primary site will only see one absent component. In this case the virtual machine remains accessible and is not impact by this failure. However, if there are any virtual machines on the secondary host running on the VSAN datastore, these will suffer the same issues seen in the previous test.



After the test, repair the network, and note that the VM/Host affinity rules will trigger a move of the virtual machine(s) back to hosts on the preferred site. Run a VSAN Health Check test before continuing to test anything else. Remember with *NumberOfFailuresToTolerate* = 1, test one thing at a time. Verify that all absent components are active and that all health check tests pass before continuing.

Network failure - Data Site to Witness Site

In this test, the VSAN network is disabled on the witness site.



As per the previous test, for physical witness hosts, the VSAN network can be physical removed from either the host or the network switch. Alternatively, the uplinks that are used for the VSAN network can be set to an “unused” state in the “Teaming and failover” properties of the VSAN network port group.

If the witness host is an ESXi VM, then the network connection used by VSAN can simply be disconnected from the virtual machine.

The expectation is that this will not impact the running virtual machine since there is one full copy of the data must be available, and more than 50% of the components that go to make up the object are available.

Once the behavior has been verified, repair the network, and run a VSAN Health Check test before continuing with further tests. Test one thing at a time. Verify that all absent components are active and that all health check tests pass before continuing.

Disk failure – Data Site host

In this test, a disk is failed on one of the hosts on the data site. This disk will contain one of the components belonging to an object that is part of the virtual machine. The expectation is that this will not impact the running virtual machine since there is one full copy of the data still available, and more than 50% of the components that go to make up the object are available. The missing data component will show up as absent in the vSphere web client UI.

Disk failure - Witness host

In this test, a disk is failed on the host on the witness site. The expectation is that this will not impact the running virtual machine since both copies of the data are still available, and more than 50% of the components that go to make up the object are available. The witness component will show up as absent in the vSphere web client UI.

VM provisioning when a sites is down

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

Compatibility:



Datastore does not match current VM policy.

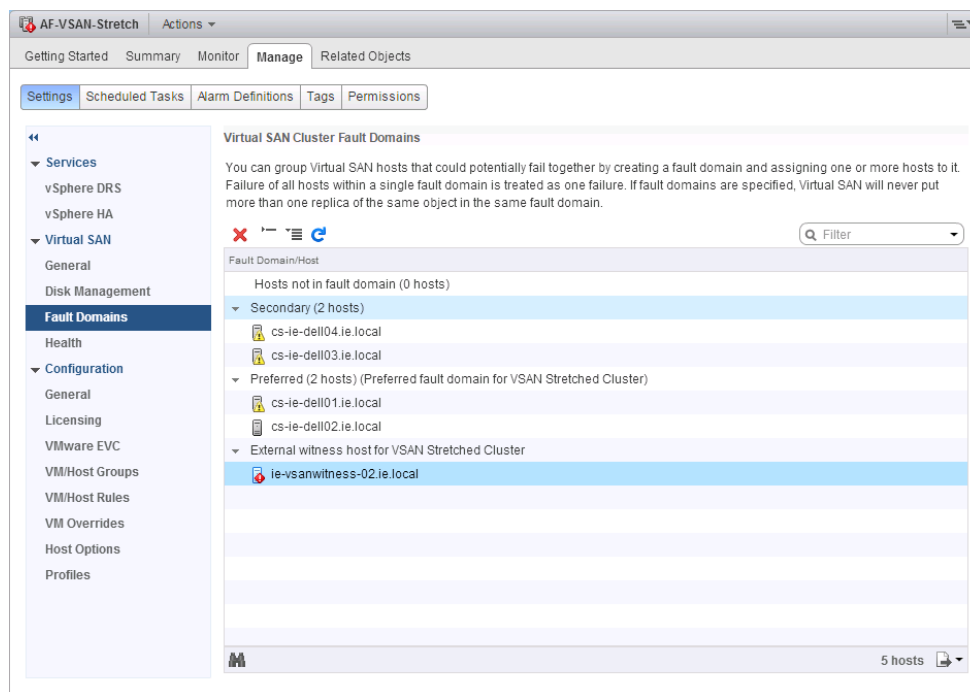
This storage policy requires at least 3 fault domains with hosts contributing storage but only 2 were found

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. This means that the virtual machine is provisioned with a *NumberOfFailuresToTolerate* = 0, meaning that there is no redundancy. Administrators will need to rectify the issues on the failing site and bring it back online. When this is done, Virtual SAN will automatically update the virtual machine configuration to *NumberOfFailuresToTolerate* = 1, creating a second copy of the data and any required witness components.

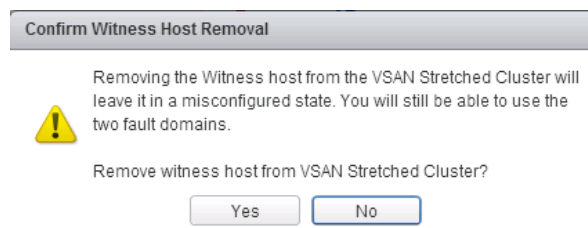
Replacing a failed witness host

Should a witness host fail in the Virtual SAN stretch cluster, a new witness host can easily be introduced to the configuration. If the witness host fails, there will be various health check failures, and all witness components will show up as absent, but all virtual machine continue to be available since there is a full copy of the virtual machine object data available as well as greater than 50% of the components (consider NumberOfFailuresToTolerate=1, there will be 2 replica copies available, implying 66% component availability).

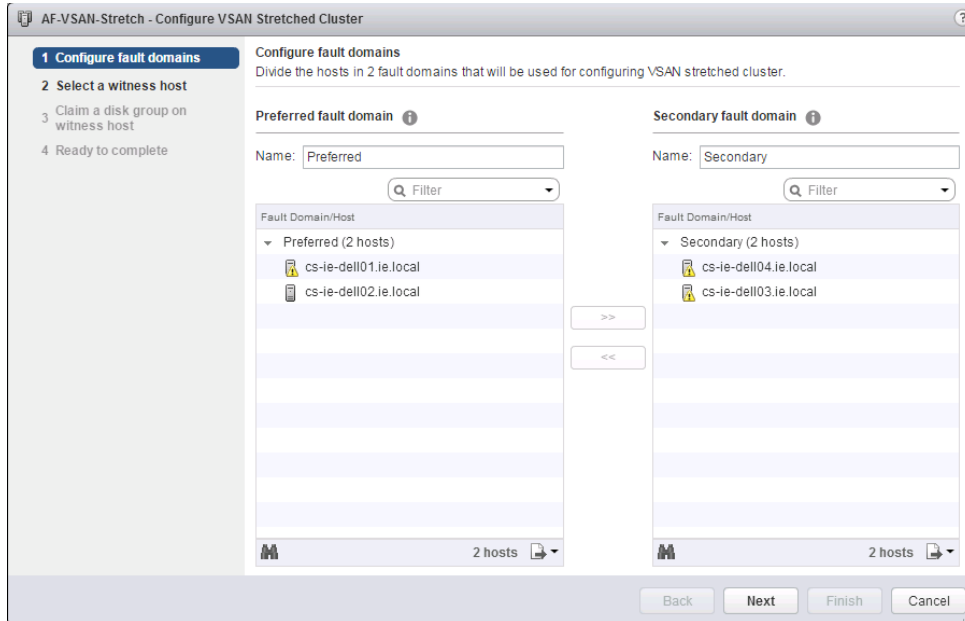
At this point, the failed witness needs to be removed from the configuration. Navigate to Cluster > Manage > Virtual SAN > Fault Domains. For this particular test, a 2+2+1 configuration is used, implying two ESXi hosts in the “preferred” data site, two ESXi hosts in the “secondary” data site and a single witness host.



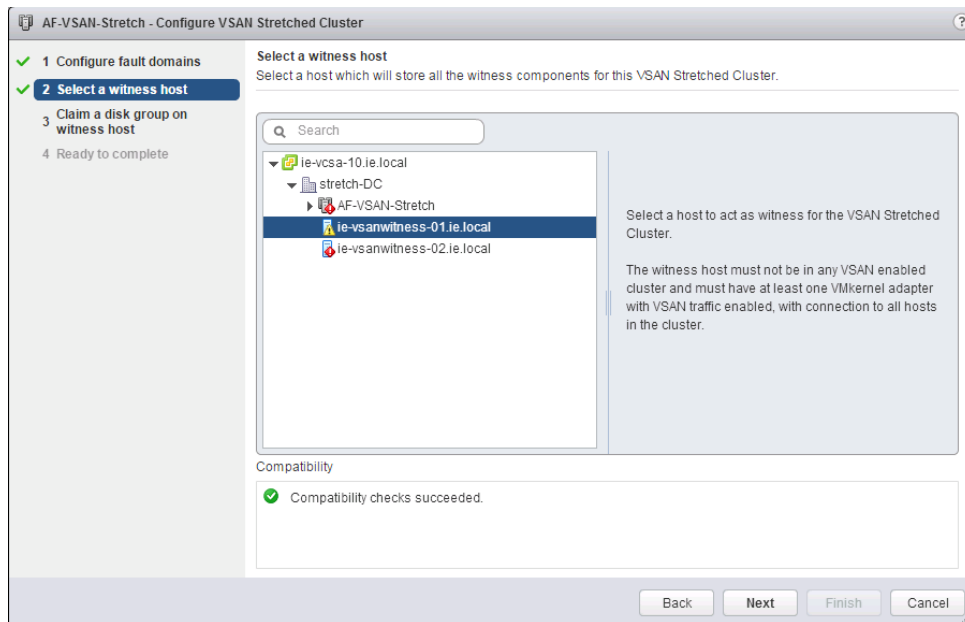
The failing witness host can be removed from the Virtual SAN Stretched Cluster via the UI (red X in fault domains view).



The next step is to rebuild the VSAN stretched and selecting the new witness host. In the same view, click on the “configure stretched cluster” icon. Align hosts to the preferred and secondary sites as before. This is quite simple to do since the hosts are still in the original fault domain, so simply select the secondary fault domain and move all the hosts over in a single click:



Select the new witness host:



Create the disk group and complete the Virtual SAN Stretched Cluster creation.

On completion, verify that the health check failures have resolved. Note that the Virtual SAN Object health test will continue to fail as the witness component of VM still remains “Absent”. When Clomd timer expires after a default of 60 minutes, witness components will be rebuilt on new witness host. Rerun the health check tests and they should all pass at this point, and all witness components should show as active.

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 Virtual SAN behavior when a site failure occurs and subsequently recovers. In the event of a site failure, Virtual SAN will now wait for some additional time for “all” hosts to become ready on the failed site before it starts to sync components. The main reason is that if only some subset of the hosts come up on the recovering site, then Virtual SAN will start the rebuild process. This may result in the transfer of a lot 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 Virtual SAN 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.

Appendix A: Additional Resources

A list of links to additional Virtual SAN resources is included below.

- [Virtual SAN 6.0 Proof Of Concept Guide](#)
- [Virtual SAN 6.1 Health Check Plugin Guide](#)
- [Virtual SAN Stretched Cluster Bandwidth Sizing Guidance](#)
- [Tech note: New VSAN 6.0 snapshot format vsanSparse](#)
- [Virtual SAN 6.0 Design and Sizing Guide](#)
- [Virtual SAN 6.0 Troubleshooting Reference Manual](#)
- [RVC Command Reference Guide for Virtual SAN](#)
- [Virtual SAN Administrators Guide](#)
- [Virtual SAN 6.0 Performance and Scalability Guide](#)

Location of the Witness Appliance OVA

The Witness appliance OVA is located on the Drivers & Tools tab of VSAN download page. There you will find a section called VMware Virtual SAN tools & plug-ins. This is where the "Stretch Cluster Witness VM OVA" is located. The URL is:

https://my.vmware.com/web/vmware/info/slug/datacenter_cloud_infrastructure/vmware_virtual_san/6_0#drivers_tools

Appendix B: CLI Commands for Virtual SAN Stretched Cluster

ESXCLI

New ESXCLI commands for Virtual SAN Stretched Cluster.

esxcli vsan cluster preferredfaultdomain

Display the preferred fault domain for a host:

```
[root@cs-ie-dell04:-] esxcli vsan cluster preferredfaultdomain
Usage: esxcli vsan cluster preferredfaultdomain {cmd} [cmd options]

Available Commands:
  get           Get the preferred fault domain for a stretched cluster.
  set           Set the preferred fault domain for a stretched cluster.

[root@cs-ie-dell04:-] esxcli vsan cluster preferredfaultdomain get
Preferred Fault Domain Id: a054ccb4-ff68-4c73-cbc2-d272d45e32df
Preferred Fault Domain Name: Preferred
[root@cs-ie-dell04:-]
```

esxcli vsan cluster unicastagent

An ESXi host in a Virtual SAN Stretched Cluster communicated to the witness host via a unicast agent over the VSAN network. This command can add, remove or display information about the unicast agent, such as network port.

```
[root@cs-ie-dell02:-] esxcli vsan cluster unicastagent
Usage: esxcli vsan cluster unicastagent {cmd} [cmd options]

Available Commands:
  add           Add a unicast agent to the Virtual SAN cluster configuration.
  list          List all unicast agents in the Virtual SAN cluster configuration.
  remove        Remove a unicast agent from the Virtual SAN cluster configuration.

[root@cs-ie-dell02:-] esxcli vsan cluster unicastagent list
IP Address  Port
-----  -----
172.3.0.16 12321
[root@cs-ie-dell02:-]
```

RVC – Ruby vSphere Console

The following are the new stretched cluster RVC commands:

vsan.stretchedcluster.config_witness

Configure a witness host. The name of the cluster, the witness host and the preferred fault domain must all be provided as arguments.

```
/localhost/Site-A/computers> vsan.stretchedcluster.config_witness -h
usage: config_witness cluster witness_host preferred_fault_domain
Configure witness host to form a Virtual SAN Stretched Cluster
cluster: A cluster with virtual SAN enabled
witness_host: Witness host for the stretched cluster
preferred_fault_domain: preferred fault domain for witness host
--help, -h: Show this message
/localhost/Site-A/computers>
```

vsan.stretchedcluster.remove_witness

Remove a witness host. The name of the cluster must be provided as an argument to the command.

```
/localhost/Site-A/computers> vsan.stretchedcluster.remove_witness -h
usage: remove_witness cluster
Remove witness host from a Virtual SAN Stretched Cluster
cluster: A cluster with virtual SAN stretched cluster enabled
--help, -h: Show this message
```

vsan.stretchedcluster.witness_info

Display information about a witness host. Takes a cluster as an argument.

```
/localhost/Site-A/computers> ls
0 Site-A (cluster): cpu 100 GHz, memory 241 GB
1 cs-ie-dell04.ie.local (standalone): cpu 33 GHz, memory 81 GB

/localhost/Site-A/computers> vsan.stretchedcluster.witness_info 0
Found witness host for Virtual SAN stretched cluster.
+-----+
| Stretched Cluster | Site-A |
+-----+
| Witness Host Name | cs-ie-dell04.ie.local |
| Witness Host UUID | 55684ccd-4ea7-002d-c3a9-ecf4bbd59370 |
| Preferred Fault Domain | Preferred |
| Unicast Agent Address | 172.3.0.16 |
+-----+
```



VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Copyright © 2012 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at <http://www.vmware.com/go/patents>. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdiction. All other marks and names mentioned herein may be trademarks of their respective companies.