vSAN 2-Node Cluster Guide

Recommendations for vSAN, as a part of VMware Cloud Foundation 9.0

June 17, 2025

vmware by Broadcom

Table of Contents

Introduction 4

Scope of Topics 4

vSAN 2-Node Cluster Concepts 5

vSAN 2-Node Clusters and Fault Domains 5

vSAN Witness Host 5

Data Path Optimizations for vSAN 2-Node Clusters 6

Requirements of a 2-Node Cluster 7

Witness Host Requirements 7

Networking Requirements 9

Configuration Minimums and Maximums 11

Virtual Machines Per Host 11

Hosts 11

Symmetry vs Asymmetry 11

Witness Host 12

Design Considerations 12

Storage Policies with 2-Node Clusters 12

Network Design 14

Cluster Settings - vSphere HA 16

Cluster Settings - DRS 17

Advanced Options (Cluster) 17

Initial Deployment 19

Management and Maintenance 19

Lifecycle Management 19

Healthy State of the Cluster 19

Failure Scenarios 20

Adaptive Quorum Control 20

Limitations 20

Summary 20

Additional Resources 20

About the Author 21

Introduction

vSAN 2-Node clusters a highly flexible, economical, and resilient vSAN configuration. The design of a 2-Node cluster minimizes the hardware investment and complexity of a deployment, yet provides a full vSAN HCI cluster experience, where VMs and their data can remain available in the event of a host outage.

2-Node clusters can be ideal for a variety of use cases at the edge, such as remote offices, or small disaster recovery locations. They can even be an option for small environments that need a small test and development clusters, or isolated environments to meet security requirements. But just because they are small doesn't mean that they cannot be powerful. With modern hardware, and vSAN ESA, customers can build incredibly robust 2-Node configurations that can achieve extraordinary levels of performance. This guide was developed to provide additional insight and information for the design, installation and configuration of vSAN stretched clusters. It will also cover operational procedures and explain how it handles failure scenarios unique to this topology.

A VMware 2-Node cluster is a cluster configuration that consists of two hosts, usually in the same geographic location such as a rack or a room at a remote location, and are typically connected directly to each other without the need of a network switch. In addition to the two hosts storing the data, will be the use of a vSAN Witness Host that resides at another location, usually located at a central data center, and is connected to both hosts via a low bandwidth, high latency link.

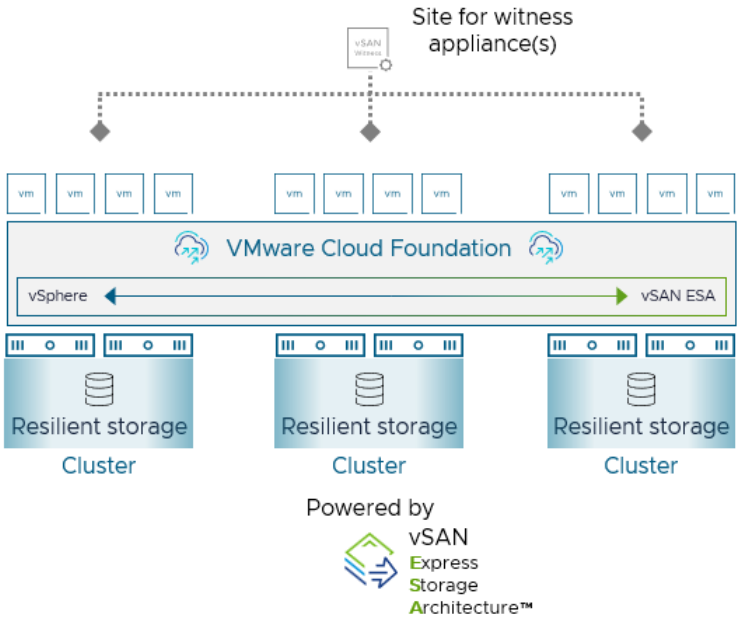


Figure. vSAN 2-Node clusters for edge and remote office

vSAN 2-Node clusters provide more than just data resilience across hosts. Thanks to its integration with vSphere HA, and vSphere DRS, it provides a comprehensive solution of high availability of mission critical workloads and the data it serves. vSAN provides this host-level resilience all as a cluster configuration type that is relatively easy to install and manage.

Scope of Topics

The information provided in this document will assume the use of vSAN 9.0. VMware Cloud Foundation (VCF) 9.0 may have specific workflows through the SDDC Manager to complete certain configuration tasks, or may only support certain features under specific conditions. VCF may also have additional requirements and support limitations that fall outside of the scope of this document. **For formal VCF installations, please refer to the "Administration Guide for VMware Cloud Foundation" as the authoritative resource for VCF installations.**

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

Many of the concepts shared in this document mirror that of a vSAN stretched cluster, as described in the "vSAN Stretched Cluster Guide." Differences will be noted where appropriate, and this documentation may refer to that guide for brevity.

vSAN 2-Node Cluster Concepts

vSAN 2-Node Clusters and Fault Domains

vSAN uses a construct known as a "fault domain" to help it distribute data in a resilient way. By default, vSAN treats each host as a fault domain, which helps keep data available in the event of a discrete host failure. vSAN has an optional feature known as "vSAN Fault Domains." When one defines a group of hosts as a "fault domain" it identifies this an additional logical boundary of failure. It can use this information to write the data in such a way that if that defined fault domain is offline, the data will remain available. The Fault Domains feature will commonly be used to ensure rack-level resilience in a data center and is compatible with both vSAN OSA, and ESA.

vSAN 2-Node clusters are a cluster deployment option within vSAN that use this same concept of fault domains. Instead of a fault domain representing a single rack or room in a data center, it represents one physical host. In a 2-Node cluster, there are two physical hosts responsible for storing the data in a resilient manner. With a witness host help vSAN determine the availability of the data. The defined fault domains in a 2-Node cluster ensures that the virtual machines and the data they consume will maintain availability in the event of a host outage. These sites are known in the product as:

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

2-Node clusters are most often directly connected to each other, while the witness host resides in another location, and communication occurs through the network topology.

vSAN Witness Host

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

Just like stretched clusters, 2-Node clusters need additional help to account for their topology. A witness host is used in these configurations, where it stores a very small amount of metadata associated with the object on the witness host. The metadata helps vSAN understand if there is enough resilient data active to keep the data available or not. The witness host is critical in in host isolation conditions, where one data host could be completely isolated from another site due to network connectivity between hosts. vSAN can determine quorum on if and where the data should remain available or unavailable. This is what prevents "split-brain" conditions where a VM and its data are updated independently.

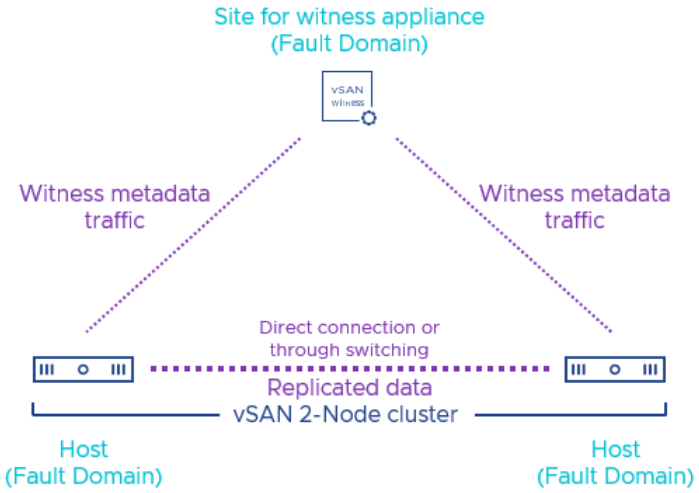


Figure. Relationship of the witness host and witness site to the data hosts in a vSAN 2-Node cluster.

The witness host must be managed by the same vCenter Server managing the vSAN 2-Node cluster. It will reside in the hosts and clusters inventory within vCenter Server but will not be a member of a specific cluster. The managing vCenter Server must be able to communicate with the two data hosts in the vSAN cluster, as well as the witness host. See the vSAN Witness Host Requirements section of this document for more information.

Data Path Optimizations for vSAN 2-Node Clusters

vSAN has several data path optimizations built into the product that work well for 2-Node topologies, as well as other settings that will activate when a 2-Node cluster is deployed.

Proxy Communication

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

Data Compression and Encryption

vSAN ESA will compress and encrypt the data prior to the data being transmitted across the network, which will essentially increase the effective bandwidth of a network and minimize the effort to transmit data between hosts. This is the same method used for vSAN stretched clusters, as discussed at: "Using the vSAN ESA in a Stretched Cluster Topology."

DRS after Recovery from a Failed Site

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

Requirements of a 2-Node Cluster

To provide a level of service and supportability for the data that is stored on a 2-Node cluster, additional requirements extend beyond the minimum requirements for a standard vSAN cluster. They include:

Witness Host Requirements

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

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

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

A virtual witness host appliance running on VMware Workstation or VMware Fusion is **NOT** a supported configuration. The virtual witness host appliance must run on a licensed vSphere host.

Witness Host Location

The witness host appliance for 2-Node clusters will generally reside at a location different than the location of the two physical data hosts. It is most common to see a virtual witness host appliance deployed at the primary data center, and the two data hosts of a 2-Node cluster deployed at the edge. Occasionally, some configurations will use a virtual or physical witness host at the same physical site or location as the two data hosts. If it is a virtual witness host appliance**, it cannot reside on one of the physical data hosts in the 2-Node cluster.** If one chooses to use a physical ESXi host to act as the witness, it will be subject to license consumption rules.

The witness host appliance can run on any vSphere environment with any supported storage, such as a VMFS datastore, NFS datastore, or a vSAN cluster. Ensure that the placement of the witness does not create any type of circular dependency. This will help avoid an unintended situation in which one failure initiates other failures.

Witness Host in the vSphere Client UI

When deployed, the virtual witness host appliance will show up in the “Hosts and Clusters” view of the vSphere client with a special color.

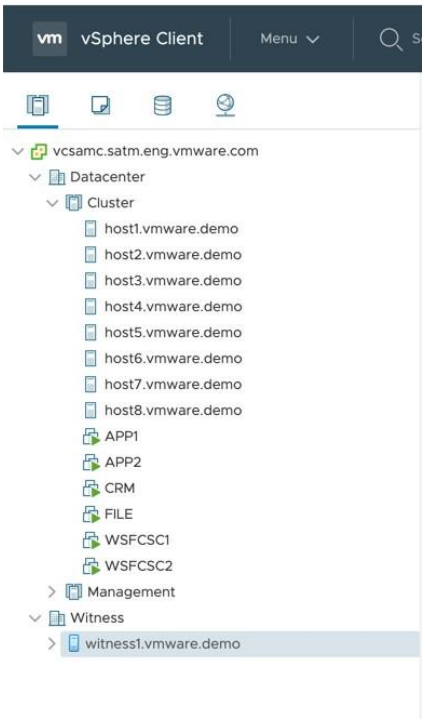


Figure. The virtual witness host appliance in the Hosts and Clusters view.

Unlike a stretched cluster, vSAN 2-Node clusters can share a witness host. While the witness host appliance can be shared with up to 64 2-Node clusters.

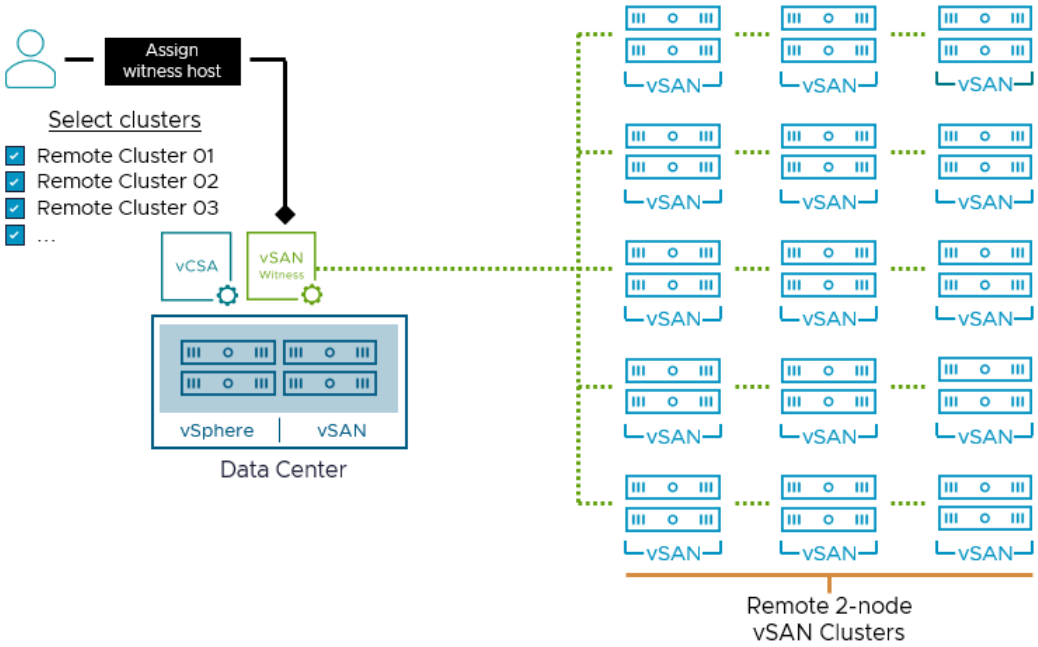


Figure. Shared witness host appliance for many vSAN 2-Node clusters.

While this is an intriguing capability that can reduce resources used for large-scale 2-Node cluster deployments, you will want to balance the desire to consolidate witness host appliances with the implications of increasing the dependency domain. For more information, see the post: "New Design and Operation Considerations for vSAN 2-Node Topologies."

Witness Host Lifecycle Management

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

Networking Requirements

Since vSAN is a distributed storage system, networking plays a critical role in any vSAN environment. While the networking considerations for a 2-Node cluster are not as substantial as with a vSAN stretched cluster, there are some networking-related aspects to be mindful of in a design and deployment of one or more vSAN 2-Node clusters.

Data Site to Data Site Network Bandwidth and Latency

vSAN 2-Node clusters generally adhere to the same bandwidth requirements imposed on vSAN stretched clusters, but are often much easier to achieve. For example, the minimum bandwidth between data hosts is 10Gbps with. However, in most cases, 2-Node clusters will directly connect the two data hosts without any network switching. This means that two hosts with a 25GbE or 100GbE NIC can easily achieve a high bandwidth, low latency connection between hosts. With vSAN ESA, these faster connections are highly recommended.

Data Site to Witness Site Network Bandwidth and Latency

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

The minimum latency required between witness site and the data sites should not exceed **500ms RTT**, or 250ms one way.

Firewalls, IDS and WAN Optimization

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

Witness Traffic Separation (Optional)

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

MTU Sizes

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

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

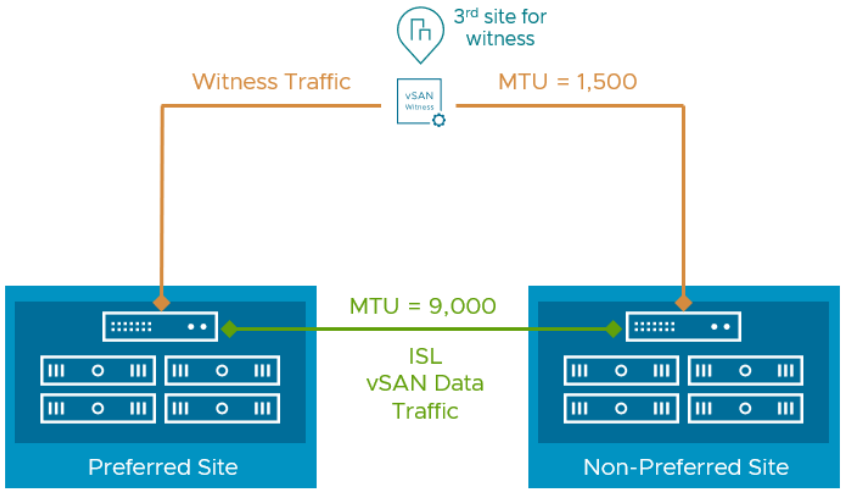
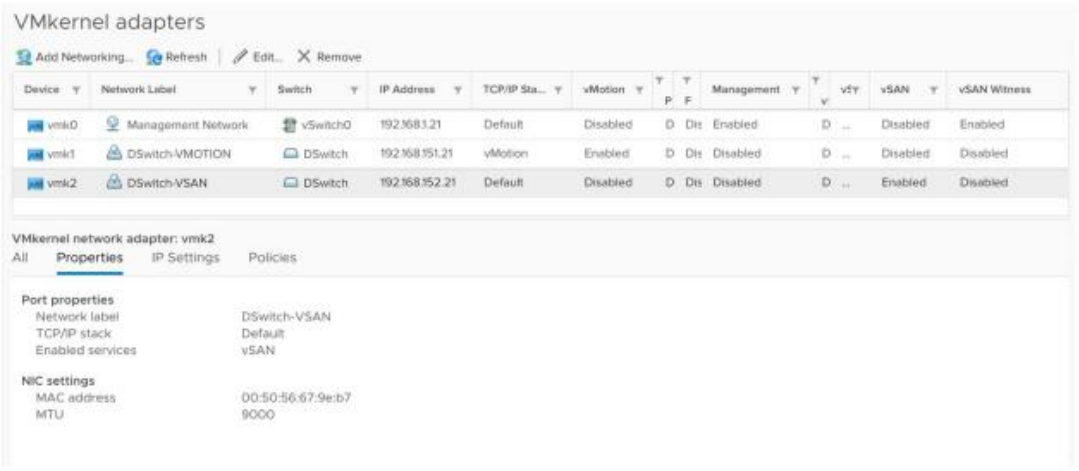


Figure. Illustrating how vSAN supports the use of mixed MTU sizes with 2-Node clusters and stretched clusters..

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



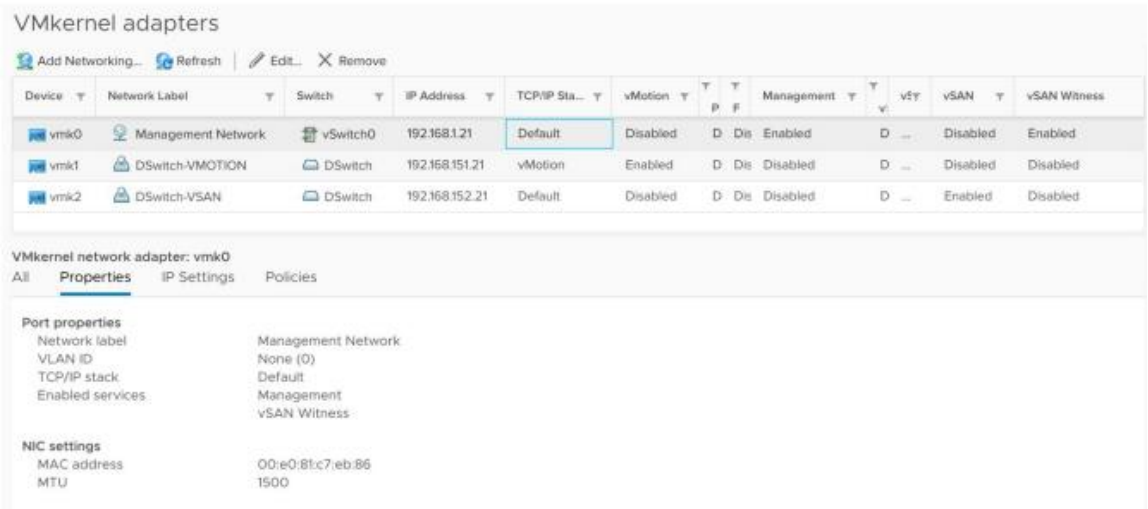


Figure. Configuration of witness traffic separation on the data hosts in a vSAN 2-Node cluster.

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

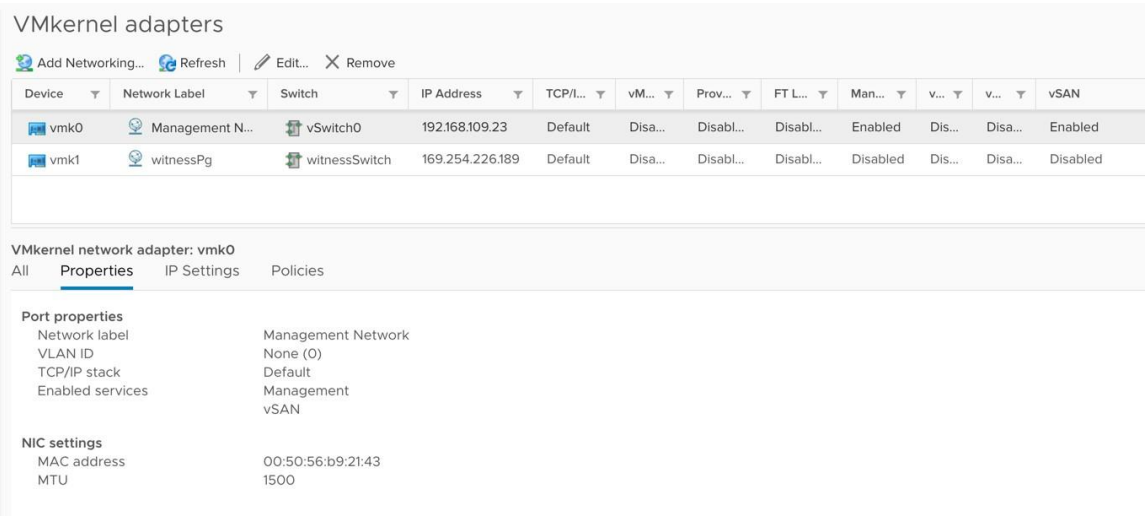


Figure. Configuration of the witness host appliance in a vSAN 2-Node cluster.

For stretched cluster and 2-Node clusters that use a different MTU size for the vSAN data payload and the vSAN witness traffic, this may trigger a health check warning. See KB 317670 "VMware vSAN Network Health Check for MTU check fails in a Stretched Cluster" for more information.

Configuration Minimums and Maximums

vSAN 2-Node clusters clusters do have considerations and limitations that are unique to the topology.

Virtual Machines Per Host

The number of virtual machines supported in vSAN is unaffected by a vSAN 2-Node cluster deployment. The maximum number of VMs supported is the same as standard vSAN cluster deployments - up to 200 VMs per host when running vSAN OSA, and up to 500 VMs per host when running vSAN ESA. Since 2-Node cluster are designed to provide resources in the event of a host failure, the following limits should be incorporated into your design exercise.

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

Hosts

Unlike single-site vSAN clusters, or vSAN stretched clusters, a 2-Node cluster has a strict design in terms of a host count. It will consist of the following:

- No more, or no less than 2 data hosts in the cluster.
- A third host serving as a witness. This is usually in the form of a virtual witness host appliance and can be shared with other 2-Node clusters.

Symmetry vs Asymmetry

vSAN supports asymmetrical configurations. With 2-Node clusters, reasonable levels of symmetry are recommended to help account for a host failure condition, where the one remaining host will take on the entire responsibility of running VM

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

Witness Host

2-Node clusters can use either a dedicated witness host for each cluster or share a witness host. While sharing a virtual witness host appliance across many 2-Node clusters can reduce the overall resource overhead needed for this responsibility, it does increase the dependency domain, where many 2-Node clusters are relying on a single witness host appliance to determine. **A better approach may be to only share a single witness host appliance across a smaller number of 2-Node clusters, instead of a large number of 2-Node clusters**. For more information, see the post: "Design and Operation Considerations for vSAN 2-Node Topologies."

Design Considerations

Storage Policies with 2-Node Clusters

Storage polices can account for the topology of a 2-Node cluster. These policies allow you to protect the data in a mirrored fashion across the two data hosts that make up the 2-Node cluster. If there are enough storage devices in each host, storage policies can provide a secondary level of resilience that would also protect against discrete disk failures. While this feature is available for 2-Node clusters using both OSA and ESA, it is much more practical using ESA. vSAN ESA can achieve this secondary level of resilience with as few as 3 storage devices per host, where vSAN OSA would need at least 6 storage devices per host.

When using a storage policy with 2-Node clusters, the "Site disaster tolerance" entry should use the "Host mirroring (2-Node)" policy. The "Failures to tolerate" entry will device the secondary level of resilience used. For 2-Node clusters, the secondary level of resilience will always use vSAN ESA's 2+1 RAID-5 erasure code if "1 failure – RAID-5" is used. vSAN ESA's 4+2 RAID-6 will be used when selecting "2 failures – RAID-6." For vSAN ESA, the latter option will require a minimum of 6 storage devices in each host.

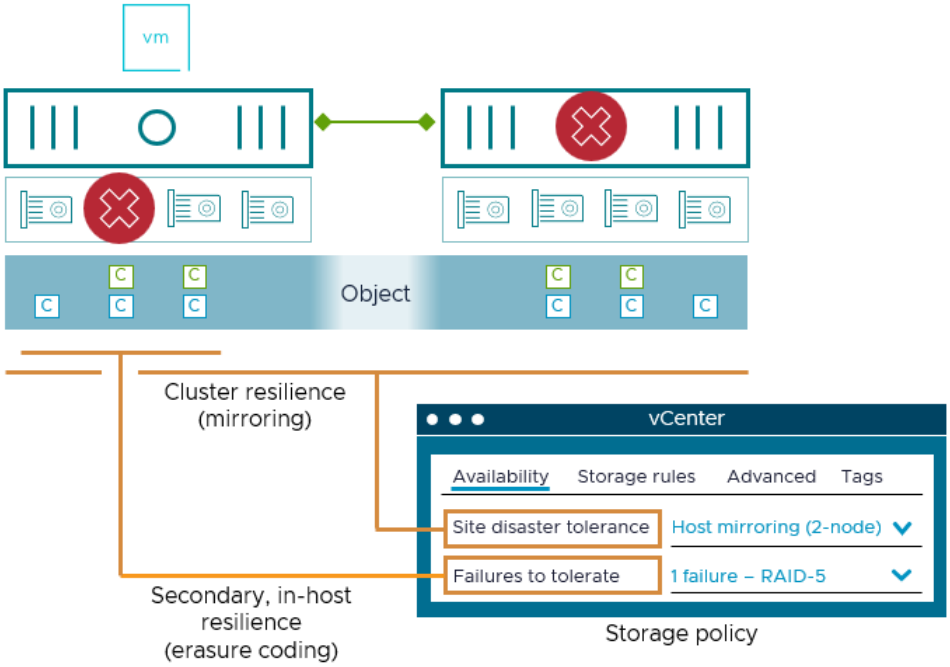


Figure. Mapping of a storage policy to vSAN 2-Node clusters

vSAN also allows you to apply storage policies for 2-Node clusters that do not provide resilience across hosts. This means that the VM data does not need to be synchronously written to both hosts, but rather, only the host where the VM resides. When paired with DRS host and VM rules, this can be useful for applications that either do not need to be replicated across

both sites, and use their own application-level replication. Even in cases where no site-level resilience is needed, one can still provide resilience within a site by applying the secondary level of failures to tolerate.

Storage Policy Rules

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

- **None – Standard Cluster.** Used for standard vSAN clusters that are not stretched.
- **Host mirroring – 2 node cluster.** Used exclusively for 2-Node clusters.
- **Dual site mirroring (stretched cluster).** Maintains availability of data in the event of an entire site.
- **None – Keep data on Preferred.** Stores data only on the preferred site.
- **None – Keep data on Non-preferred.** Stores data only on the non-preferred site.

vSAN 2-Node clusters rely on DRS settings such as VM and host group assignments and "should" rules to help balance VMs across the hosts in the cluster. For VMs do not need host-level resilience for the reasons stated above, DRS "must" rules help pin the VM instance same site that the VM data resides. A vSAN 2-Node cluster requires the configuration of these DRS rules for proper operation.

How 2-Node Cluster Storage Policy Rules Impact Capacity Requirements

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

Storage Policy Rule	vSAN Architecture	Capacity Consumption Multiplier in Accordance to Applied Policy	Capacity Consumption Result
Host mirroring without secondary level of resilience	Both	2x	200GB
Host mirroring with secondary FTT=1 using RAID-1	Both	4x	400GB
Host mirroring with secondary FTT=1 using RAID-5	OSA	2.66x	266GB
Host site mirroring with secondary FTT=1 using RAID-5	ESA	3x	300GB
Host site mirroring with secondary FTT=2 using RAID-6	Both	3x	300GB

Preferred Site (no host mirroring) with secondary FTT using RAID-1	Both	2x	200GB
Preferred Site (no host mirroring) with secondary FTT using RAID-5	OSA	1.33x	133GB
Preferred Site (no host mirroring) with secondary FTT using RAID-5	ESA	1.5x	150GB
Preferred Site (no site mirroring) with secondary FTT using RAID-6	Both	1.5x	150GB

Network Design

The three fault domains that comprise a 2-Node cluster must maintain network connectivity to each other during a normal operating state. This configuration will allow the loss of any one fault domain and still maintain availability of the data.

Connectivity and Network Types

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

Layer 2 versus Layer 3 Networking

2-Node clusters are most often directly connected to each other, without the use of any switches in between the data hosts. When this is the case, Layer 2 networking provides a simple and effective way to transmit data synchronously across the two hosts. If the two data hosts are connected via network switching, the networking between the two data hosts can also use Layer 3.

We do recommend that Layer 3 (routed) networking between the two data hosts, and the witness site. Layer 3 will help avoid spanning tree protocol (STP) from redirecting traffic across an undesirable link, such as the more bandwidth constrained link to the witness site.

When networking is properly configured, the following characteristics will exist

- The data sites will only be able to communicate with each other using the ISL
- The witness host should only be able to communicate with hosts in each data site directly, and not through another data site.

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

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

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

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

Layer 2 VLANs

While we typically recommend each discrete vSAN cluster use its own VLAN, in cases where you have one or more virtual witness host appliances providing witness services to several vSAN 2-Node clusters, there is no need to separate the witness traffic on discrete VLANs for the witness host appliance. One ore more witness host appliances can uses the same shared VLAN for witness communication.

Network Port Communication

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

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

Default Gateway on ESXi hosts in a vSAN 2-Node Cluster

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

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

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

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

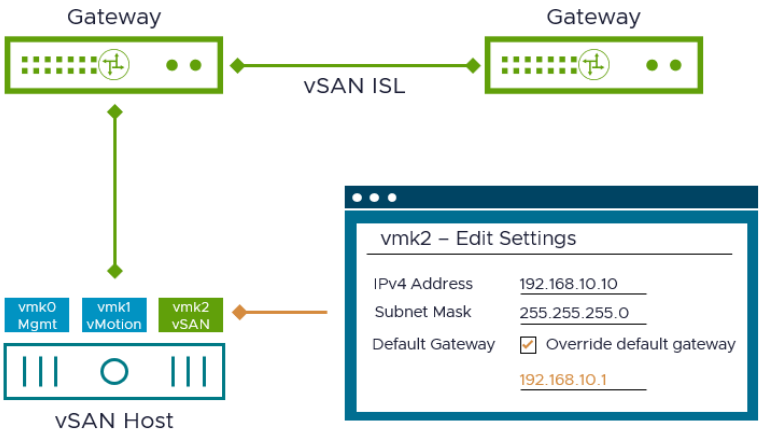


Figure. Changing the default gateway for VMkernel traffic tagged for vSAN

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

Routing of other Networks

For vSAN 2-Node clusters, there is typically no need to configure service networks such as vMotion or VM port group networks with static routes. The use of default gateway overrides, or static routing generally applies to the communication to and from the witness host. The witness host does not use these other networks. Note that the witness host will need to communicate with vCenter server, so in some topologies, a static route may need to be added on the witness host.

Custom TCP/IP Stacks

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

Network Bandwidth

Since most 2-Node clusters are directly connected to each other, network bandwidth and latency are rarely an issue. In fact, many 2-Node clusters directly connected using high-performance 25/100Gb networking can produce extraordinary performance. The bandwidth and latency requirements to the witness site should still be understood. For more information on calculating bandwidth requirements, see the "vSAN Stretched Cluster Bandwidth Sizing" guide.

Cluster Settings - vSphere HA

VMware vSAN is fully integrated and supported with vSphere HA, and should be enabled on all vSAN clusters. Some additional setting within the vSphere HA configuration are recommended for 2-Node clusters, and are detailed below:

Host monitoring

Host monitoring should be enabled on vSAN 2-Node cluster configurations. This feature uses network heartbeats to determine the status of hosts participating in the cluster, and if corrective action is required, such as restarting virtual machines on other nodes in the cluster.

Virtual Machine Response for Host Isolation

This setting determines what happens to the virtual machines on an isolated host, such as a host that can no longer communicate to other nodes in the cluster nor reach the isolation response IP address. For 2-Node clusters, setting the "Response for Host Isolation" to "Power off and restart VMs" is recommended. This is because a clean shutdown will not be possible as on an isolated host. The VM would be unable to write any data to disk in this condition.

Admission Control

Admission control ensures HA has sufficient resources to restart virtual machines after a failure. As a complete site failure is one scenario that needs to be considered in a resilient architecture, VMware recommends enabling vSphere HA Admission Control. Availability of workloads is the primary driver for most stretched cluster environments. Sufficient capacity must, therefore, be available for a total site failure. Since 2-Node clusters only have one data host in each fault domain, it is recommended that Admission Control be set to 50% for both memory and CPU.

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

Host Hardware Monitoring – VM Component Protection

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

Heartbeat Datastores

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

Advanced Options (HA)

The use of host isolation addresses is particularly important to ensure the proper behavior of vSAN stretched clusters during a failure condition, such as a site partition. When using vSAN 2-Node clusters where the two hosts are in the same location (directly connected, or connected to a ToR switch, etc.), there is no need to have a separate das.isolation address for each of the hosts.

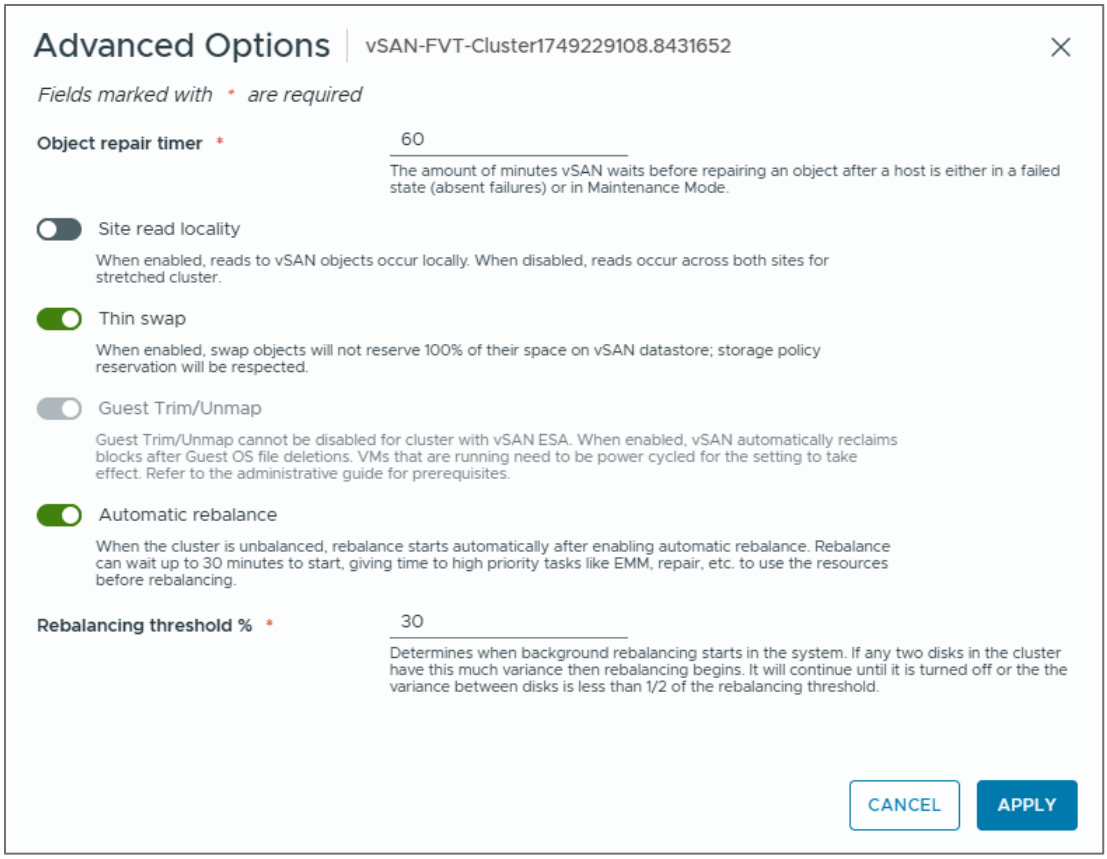
Cluster Settings - DRS

vSphere DRS is used to intelligently distribute VM instances across a cluster. It is fully integrated with vSAN its use is encouraged. vSAN 2-Node clusters will support DRS using multiple automation levels.

- **Partially Automated.** After a failure and subsequent recovery occurs, this level does not move workloads back until an administrator chooses. It tends to be used more with stretched clusters than 2-Node clusters, since stretched clusters use an ISL that may be constrained in bandwidth capabilities.
- **Fully Automated.** After a failure and subsequent recovery occurs, this level moves workloads back automatically. It is most often used with 2-Node clusters where the two physical hosts reside in the same location – directly connected, or through ToR switching.

Advanced Options (Cluster)

The cluster advanced options will look very similar to the advanced options with other vSAN clusters, and can be adjusted if desired. The "Site read locality" setting is more applicable to vSAN OSA configurations, where it was preferrable to disable this option in most 2-Node cluster arrangements. You may wish to disable it in vSAN ESA as well.


Advanced Options
vSAN-FVT-Cluster1749229108.8431652
Fields marked with * are required
Object repair timer *
60
The amount of minutes vSAN waits before repairing an object after a host is either in a failed state (absent failures) or in Maintenance Mode.
Site read locality
When enabled, reads to vSAN objects occur locally. When disabled, reads occur across both sites for stretched cluster.
Thin swap
When enabled, swap objects will not reserve 100% of their space on vSAN datastore; storage policy reservation will be respected.
Guest Trim/Unmap
Guest Trim/Unmap cannot be disabled for cluster with vSAN ESA. When enabled, vSAN automatically reclaims blocks after Guest OS file deletions. VMs that are running need to be power cycled for the setting to take effect. Refer to the administrative guide for prerequisites.
Automatic rebalance
When the cluster is unbalanced, rebalance starts automatically after enabling automatic rebalance. Rebalance can wait up to 30 minutes to start, giving time to high priority tasks like EMM, repair, etc. to use the resources before rebalancing.
Rebalancing threshold % *
30
Determines when background rebalancing starts in the system. If any two disks in the cluster have this much variance then rebalancing begins. It will continue until it is turned off or the the variance between disks is less than 1/2 of the rebalancing threshold.
CANCEL
APPLY

Initial Deployment

An initial deployment of a vSAN 2-Node cluster assumes the following:

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

The initially deployment will generally follow the same procedures as the installation of a vSAN stretched cluster. For more information, see the "Initial Deployment" section of the "vSAN Stretched Cluster Guide."

Management and Maintenance

This section covers some basic operational activities for vSAN 2-Node clusters. General operational guidance beyond 2-Node clusters can be found in the VMware Operations Guide.

Lifecycle Management

The process of patching and upgrading vSAN 2-Node clusters is very similar to traditional, single site vSAN clusters. The vSphere Lifecycle Manager (vLCM) is now responsible for all lifecycle management duties with the hypervisor, including the core hypervisor, software drivers, and some hardware firmware.

Updating a Cluster

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

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

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

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

Updating the Witness Host Appliance

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

Healthy State of the Cluster

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

Failure Scenarios

The behavior of a vSAN 2-Node cluster is very similar to that of a vSAN stretched cluster. The primary difference is that with a stretched cluster, a failure of a fault domain would represent all of the hosts in that given site. Whereas with a 2-Node cluster, a failure of a fault domain will be just that single host. Aside from this difference, the failure handling behavior will be the same, as described in the "Failure Scenarios" section of the "vSAN Stretched Cluster Guide."

Adaptive Quorum Control

Adaptive Quorum Control (AQC) is a technique that maintains data availability of objects during a site failure (or maintenance) followed by subsequent unavailability of the witness host. In a fully operational 2-Node cluster, quorum (determining availability of an object) is the result of account for object components in both sites and the witness host appliance. This is achieved through a simple voting mechanism. With this feature, when a data site has a planned or unplanned outage, vSAN will adjust the votes to favor the active site that still has quorum. This will allow sufficient votes to maintain quorum, which will keep the data available during times of a planned or unplanned outage of the witness host appliance. Depending on the size of the cluster, it may take a few seconds to a few minutes to adjust all of the component votes. As it completes each object, then that object is able to tolerate the failure of a witness host and still maintain availability. This capability will not protect against a simultaneous double failure of a data site and a witness.

Limitations

vSAN 2-Node clusters may have limited support of some features, deployment options, or topologies. This would include, but is not limited to the following:

- **vSAN storage clusters.** A 2-Node vSAN cluster cannot be deployed as a storage cluster
- **vSAN File Services.** 2-Node clusters can be deployed with vSAN File Services (vSAN 7 U2 and later).

Summary

vSAN 2-Node clusters provide a highly flexible and cost efficient way to provide storage and compute resources at the edge. Following the design and operational recommendations in this document will ensure that your experience with vSAN 2-Node clusters is successful for you and your organization.

Additional Resources

The following are a collection of useful links that relate to vSAN 2-Node clusters.

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

vSAN Stretched Cluster Guide. This guide provides design and operational guidance for vSAN stretched clusters, which mirror many of the same concepts found in this guide.

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

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

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

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

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

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

Official vSAN documentation. The location for all "how to" documentation on vSAN.

About the Author

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