VMware vCloud® Architecture Toolkit™
for Service Providers

Architecting VMware vSAN™ 6.2 for VMware Cloud Providers™

Version 2.9

January 2018

Martin Hosken

VMware, Inc.
3401 Hillview Ave
Palo Alto, CA 94304
www.vmware.com

Contents

List of Figures

List of Tables

Overview

VMware vCloud Providers™ are looking for ways to not only improve their storage infrastructure and services, but also increase capacity, simplify operations, and provide continuous up-time. As all technologists know, embracing changing technologies is integral to an organization's success. Therefore, if you are running into performance and capacity problems with your current cloud storage offerings, it is time to look at VMware Virtual SAN™.

The software-defined storage platform offered by vSAN is critical to delivering hybrid cloud service offerings in which infrastructure is provided by a building block hyper-converged solution. There are many different workload profiles in the data center, but from the perspective of storage, workloads have traditionally been separated into two profile types: file-based and block-based. While these technologies have had a long track record of success in the enterprise space, their track record has been somewhat limited in the cloud space.

VMware Cloud Providers require that their storage infrastructure not only be reliable and scalable, but also cost effective. The consumers of this storage expect this cost efficiency to be passed on to them. More often than not, these solutions are designed leveraging commodity servers in conjunction with internal storage capabilities.

There are several approaches for architecting a VMware software-designed storage solution and the approach depends on the use case and the technology deployed. However, the end goal is always the same. This *VMware vCloud Architecture Toolkit for Service Providers* (vCAT-SP) document describes a VMware Cloud Provider Program solution for supporting a software-defined storage solution using vSAN.

1.1 Enabling the Software-Defined Service Provider

VMware introduced the vision for the software-defined data center (SDDC) in 2012. It provides the infrastructure that enables the software-driven data center. The SDDC is the VMware cloud architecture in which all pillars of the data center, such as compute, storage, networks, and associated services are virtualized and automated. This document focuses on one aspect of the VMware SDDC, the storage pillar, and specifically discusses how vSAN fits into this vision for service providers.

1.2 VMware Software-Defined Storage Model

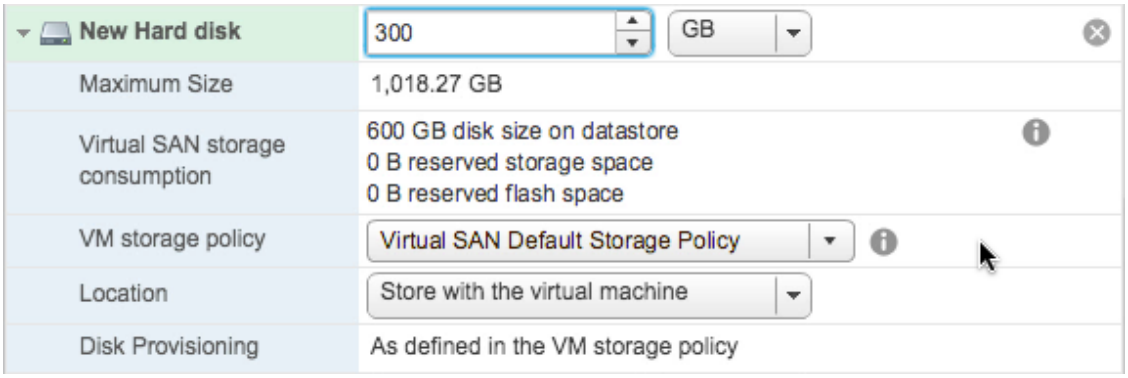
The VMware software-defined storage strategy focuses on a set of VMware initiatives regarding local storage, shared storage, and storage and data services. Software-defined storage is designed to provide storage services and service level agreement (SLA) automation through a software layer on the hosts that integrates with and abstracts the underlying hardware. With software-defined storage, virtual machine storage requirements can be dynamically instantiated. There is no need to repurpose LUNs or volumes. Virtual machine workloads might change over time, and the underlying storage can be adapted to the workload at any time.

A key factor for software-defined storage is Storage Policy-Based Management (SPBM), which was first featured in the VMware vSphere® 5.5 release, and can be considered the next generation of VMware vSphere storage profile feature. Storage Policy-Based Management offers a critical component for VMware in implementing software-defined storage. Using SPBM and VMware vSphere APIs, the underlying storage technology provides vSphere administrators with an abstracted pool of storage space for virtual machine provisioning. The technology's various capabilities relate to performance, availability, and storage services such as replication. A vSphere administrator can then create a virtual machine storage policy using a subset of the capabilities required by the application running in the virtual machine.

At the time of deployment, the vSphere administrator selects the virtual machine storage policy appropriate for the needs of that virtual machine. SPBM pushes the requirements down to the storage layer. Datastores that provide the capabilities included in the virtual machine storage policy are made available for selection. So, based on storage policy requirements, the virtual machine is always instantiated on the appropriate underlying storage. If the virtual machine's workload changes over time, a new policy with updated requirements that reflect the new workload is applied.

Storage Policy-Based Management plays a key role for vSAN because its policies are applied granularly to a virtual machine's virtual disks stored in the vSAN datastore. SPBM policies define which features of vSAN are applied to individual virtual disks.

Figure 1. VM Storage Policy Applied to a Virtual Machine's Disk



For further information about the VMware software-defined storage model, see *The VMware Perspective on Software-Defined Storage* white paper at https://www.vmware.com/files/pdf/solutions/VMware-Perspective-on-software-defined-storage-white-paper.pdf.

1.3 Target Audience

This document is targeted towards service provider architects, engineers, application owners, and technology leaders involved in the key decision making process and anyone else interested in guidance on designing a software-designed storage solution leveraging VMware technologies.

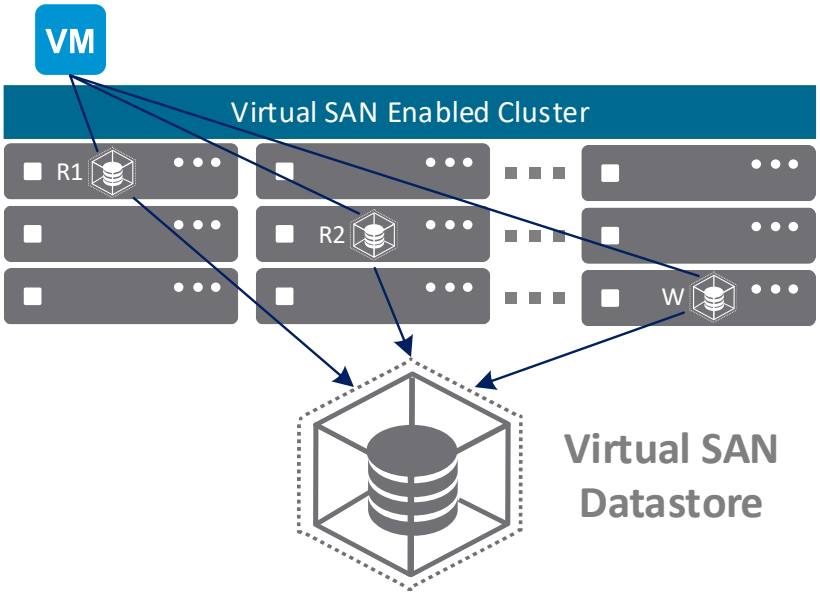
vSAN Introduction

VMware Hyper-Converged Software (HCS) transforms industry-standard x86 servers and directly attached storage into radically simple Hyper-Converged Infrastructure (HCI) to help eliminate high IT costs, management complexity, and performance limitations.

vSAN is a software-defined storage solution that is fully integrated with vSphere. vSAN aggregates locally attached disks in a vSphere cluster to create a storage solution that rapidly can be provisioned from VMware vCenter Server® during virtual machine provisioning operations. It is an example of a hypervisor-converged platform—that is, a solution in which storage and compute for virtual machines are combined into a single device, with storage being provided within the hypervisor itself as opposed to through a storage virtual machine running alongside other virtual machines.

vSAN is an object-based storage system designed to provide virtual machine-centric storage services and capabilities through an SPBM platform. Object-based storage is considered a leading technology for hybrid cloud deployments because many of its most prominent features, such as massive scalability, geographic independence, and multi-tenancy, have proven ideal for cloud storage. SPBM and virtual machine storage policies are solutions designed to simplify virtual machine storage placement decisions for vSphere administrators.

Figure 2. vSAN Cluster Example



vSAN has been tightly integrated with vCenter Server to deliver the leading HCI solution that offers simplicity, reliability, and performance for nearly any service provider use case. vSAN is fully integrated with core vSphere enterprise features, such as VMware vSphere High Availability (vSphere HA), VMware vSphere Distributed Resource Scheduler™ (vSphere DRS), and VMware vSphere vMotion®. The goal of vSAN is to provide both high availability and scale-out storage functionality. It also can be considered in the context of quality of service (QoS) because virtual machine storage policies can be created to define the levels of performance and availability required on a per–virtual machine basis.

vSAN is easy to implement with automated configuration and includes proactive tests to help verify functionality and performance. vSAN is optimized for modern all-flash storage with efficient nearline deduplication, compression, and erasure coding capabilities that lower TCO while delivering incredible performance.

New and improved features such as the performance and health services make it easier than ever to verify vSAN configurations and closely monitor key metrics such as IOPs, throughput, and latency at the cluster, host, virtual machine, and virtual disk levels. Quality of service can be managed by using IOPs limits on a per-virtual machine and per-virtual disk basis. vSAN 6.2, the latest iteration, is ready for any application with tested and validated deployments of several business critical applications, including SAP and Oracle RAC. vSAN 6.2 also provides VMware Cloud Providers with:

- A converged platform (compute + storage using commodity hardware)
- A simple to manage storage platform, as with CPU and memory
- The ability to scale on demand, in lock-step with application needs, but also in a flexible way
- Per-VM automated SLA management

In addition, one of the key features of vSAN is that the product runs as a kernel module, rather than running as a virtual appliance in a cluster, as many competitor solutions do. This allows for greater performance and better interaction with the vSphere hypervisor.

2.1 vSAN Key Requirements

As outlined above, vSAN runs as a part of VMware ESXi™ hypervisor, but has additional and more-restrictive requirements:

- A minimum of three vSAN nodes contributing storage to the cluster. (VMware strongly recommends four nodes.)
- A minimum of one SSD and one HDD (or capacity-tier SSD) per node, contributing to storage for *exclusive* use by vSAN.
- ESXi 5.5/U1 or later and equivalent vCenter Server version.
  - Use 5.5/U2/P05 if 5.5 is in use.
  - VMware strongly recommends 6.0 or later.
- The same storage controller must not serve both vSAN and VMware vSphere VMFS under the following circumstances:
  - If an attached VMFS volume is being used for persistent logging.
  - If VMs on a VMFS datastore are resident on a disk/RAID group attached to a controller that also serves vSAN disks.
- Servers, network interfaces, and so on must be on the vSphere HCL.
- Storage controllers, SSDs, and HDDs must be on the vSAN specific HCL.
  - Including driver and firmware revisions.
- VMware vSphere Web Client is exclusively used for vSAN operations and management.
- Dedicated 1-GbE link for vSAN *or* a 10-GbE link (preferred).
  - 10-GbE required for all-flash configuration.
- Upstream switches must handle multicast traffic.

2.2 vSAN Key Terminology

vSAN uses a number of terms that are good to define before proceeding:

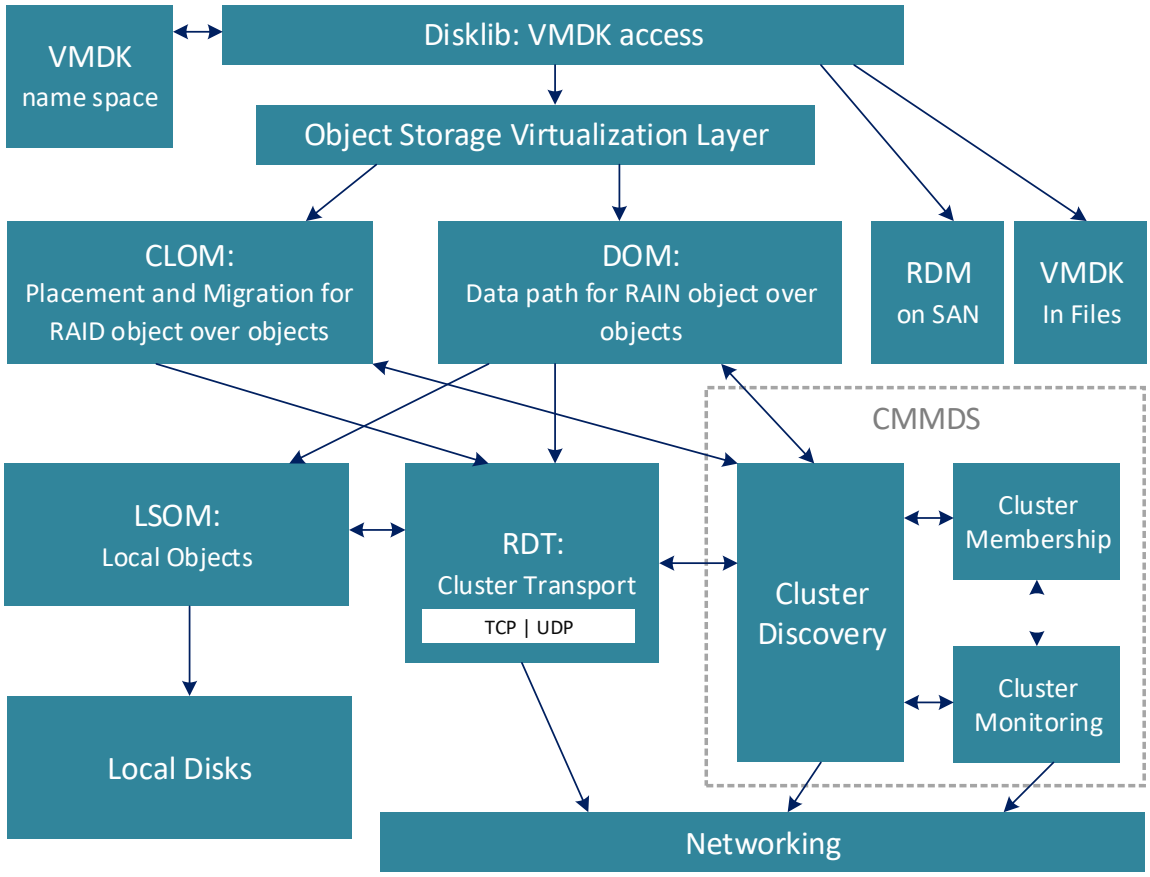
- Each vSAN host contributing storage does so in the form of a “disk group”.
- A disk group is a combination of exactly one cache-tier SSD and one or more capacity-tier drives.
- The cache tier is used for write buffering and/or read caching.
- The capacity tier is used for long-term persistence and data at rest.
- vSAN object – Every element of the VM (home directory, VMDKs, snapshots, swap files) is an object in vSAN, not a file.
- vSAN Component – Objects are distributed constructs made of components.
  - Components are the data and witnesses that allow the object to be assembled and deliver data to the end user (the VM).
- Witness – A witness is a metadata-only component. It exists to maintain quorum and to secure integrity during split-brain scenarios.

2.3 vSAN Internal Architecture

One of the vSAN unique advantages is that the product runs as a kernel module within the vSphere hypervisor, as opposed to running as a virtual appliance on the host. This architecture is therefore able to provide far greater performance and better interaction with the hypervisor. The following diagram illustrates how the modules interact with one another.

The remainder of this section addresses each of these key architectural components and its role in the vSAN integrated architecture.

Figure 3. vSAN Hypervisor Integrated Architecture



2.3.1 The Cluster-Level Object Manager (CLOM)

The CLOM exists as a userspace daemon on each ESXi host in the vSAN cluster and is responsible for:

- Orchestrating objects.
- Placement of objects during create operations.
- Ensuring that component placement satisfies the defined storage policy.
- Scheduling rebuilds should a component need to be replaced.
  - With a placement operation, it verifies that policies are satisfied.
  - If an object is reconfigured with a new storage policy, the CLOM handles the placement of any new components required.

2.3.2 The Distributed Object Manager (DOM)

The DOM exists in kernel space and there are no daemons that can be monitored or restarted without booting the host. The DOM is responsible for handling object availability and initial I/O requests. In addition:

- All objects exist at the DOM layer.
- All I/O is directed to the DOM client.
- One DOM client per vSAN enabled ESXi host.
- I/O is forwarded to the DOM owner.
- One DOM owner per object.
- There is no concept of locality—the DOM owner might be a host that does not own any component of the object.
- The DOM owner can change.
- The DOM is responsible for ensuring that all sides of a mirror are consistent.
- The DOM handles the synchronous I/O.

2.3.3 Local Log-Structured Object Manager (LSOM)

Like the DOM, the LSOM exists in kernel space, and there are no daemons that can be monitored or restarted without booting the ESXi host. All components exist at the LSOM layer and the LSOM is responsible for handling I/O and ensuring consistency of the components resident on local disks. In addition to this, the LSOM:

- Handles write-buffering, read caching and the destaging of data to the capacity tier.
- Receives I/O from the DOM.
- Returns acknowledgements to the DOM when write operations are complete and returns payloads for read operations.
- Is unaware of distribution, quorum, I/O synchronization, and so on. The DOM addresses all of that. The LSOM is responsible only for handling I/O.

2.3.4 Object Store Filesystem (OSFS)

The OSFS is a userspace daemon on each host in the vSAN cluster, which:

- Provides a filesystem-like construct for compatibility.
- Is responsible for the creation of the vsanDatastore.
    - There are no directories as such in vSAN. “Directories” in vSAN are actually objects formatted with VMFS.
- Handles the initial formatting of the “VM namespace” objects, mapping friendly names, and so on.
- Is responsible for mounting namespace objects and making them available in the vsanDatastore container.

2.3.5 The Cluster Management, Monitoring, and Directory Service (CMMDS)

The CMMDS handles all vSAN inventory of objects, hosts, disks, network interfaces, policies, names, and so on. It is the central directory which:

- All hosts join to perform elections through it.

- Other architecture elements (like DOM/LSOM) publish into CMMDS to update the directory about objects, components, and so on.

In addition:

- If CMMDS cannot communicate or update information, it will cause production problems in the vSAN cluster.
- After something is published into CMMDS, information about it can be found from any node that is part of the cluster using cmmds-tool.
- Network-partitions and communication problems typically manifest through CMMDS.

2.3.6 Reliable Datagram Transport (RDT)

The RDT is used for cluster network communication. It is optimized to send very large files. It will be used again later to send object data between hosts.

What RDT does is deliver datagrams, potentially very large if necessary, between logical endpoints (conceptually fault-tolerant clients and servers), typically over multiple paths. Based on link health status changes published by the CMMDS, the RDT sets up and tears down transport connections very quickly, to minimize delay in datagram transport due to link failures.

2.3.7 Storage Policy-Based Management (SPBM)

As outlined previously, SPBM is a core component of the VMware software-defined storage model and allows user control of storage policies on a per-object basis in vSAN. The Storage Policy-Based Management framework:

- Exists at the vCenter Server level through the VMware vSphere Storage Profile.
- Is enabled through VMware vSphere API for Storage Awareness™. (There is a vSphere API for Storage Awareness provider on each ESXi host.)
- Sends the defined storage policies (if selected) to CLOMD during provisioning.
- In the event that new policy is applied to a VM/disk, or if the policy is modified, those changes are reflected in SPBM and forwarded to CLOMD.
- Reflects in the vSphere Web Client object status (distribution, reconfiguration/resync, and so on).

vSAN Technology and Features Overview

vSAN transforms industry-standard x86 servers and directly attached storage into radically simple Hyper-Converged Infrastructure (HCI) to help eliminate high IT costs, management complexity, and performance limitations. The tightly integrated software stack includes vSphere with vSAN to deliver a radically simple, enterprise-class native storage; and vCenter Server, the unified and extensible management solution.

Being natively integrated with vSphere allows vSAN to be configured with just a few mouse clicks. Because disks internal to the vSphere hosts are used to create a vSAN datastore, there is no dependency on external shared storage. In addition, virtual machines can be assigned specific storage policies based on the needs of the applications. These workloads can then benefit from this dependable shared storage with predictable performance and availability characteristics.

vSAN version 6.2 also further reduces TCO by providing up to 10x greater storage efficiency in an optimized all-flash storage solution, which delivers efficient near-line deduplication, compression, and erasure coding capabilities that enable high performance all-flash systems for as low as one dollar per GB of usable capacity, which might be up to 50 percent less than the cost of lower-performing hybrid solutions or shared storage solutions from the leading industry competitors.

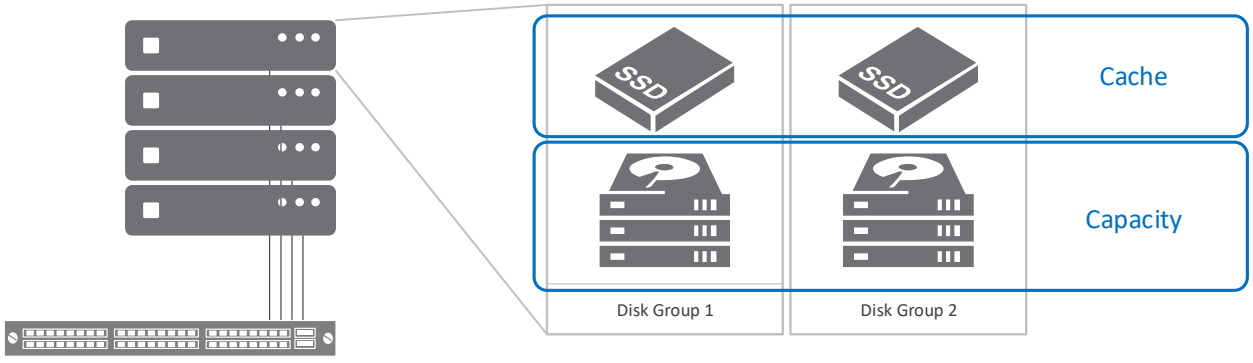
3.1 Architecture Overview

vSAN clusters must be made up of physical hosts that contain either a combination of mechanical disks and flash devices (hybrid configuration) or all flash devices (all-flash configuration) that contribute cache and capacity to the vSAN distributed datastore.

In a hybrid configuration, one flash device and one or more mechanical drives are configured as a disk group, with a disk group being able to maintain a maximum of seven mechanical drives. One or more disk groups are utilized in a vSphere host. In a hybrid configuration, the flash device serves as read-and-write cache for the vSAN datastore, while the mechanical drives make up the capacity aspect of the datastore. By default, vSAN uses 70 percent of the flash capacity as read cache and 30 percent as write cache. It is typically not recommended to modify this ratio.

In an all-flash configuration, the flash device in the cache tier is used for write caching only (no read cache) because read performance from the capacity flash devices is more than sufficient for even the most demanding enterprise application. In an all-flash configuration, two different grades of flash devices are commonly used, a lower capacity, higher endurance device for the cache layer, and more cost effective, higher capacity, lower endurance devices for the capacity layer. Writes are performed at the cache layer and then de-staged to the capacity layer, only as needed. This helps extend the usable life of the lower endurance flash devices in the capacity layer.

Figure 4. vSAN All-Flash Architecture



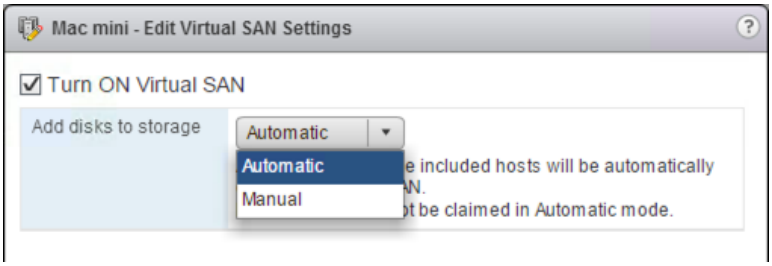
3.2 vSAN Key Features

The features described in this section are referring to vSAN 6.2, which is available through vSphere ESXi version 6.0 Update 2.

3.2.1 Management

vSAN requires vCenter Server, with both the Microsoft Windows version and the VMware vCenter Server Appliance™ capable of managing vSAN. vSAN is configured and monitored exclusively from the vSphere Web Client.

Figure 5. Enabling vSAN



3.2.2 Space Efficiency

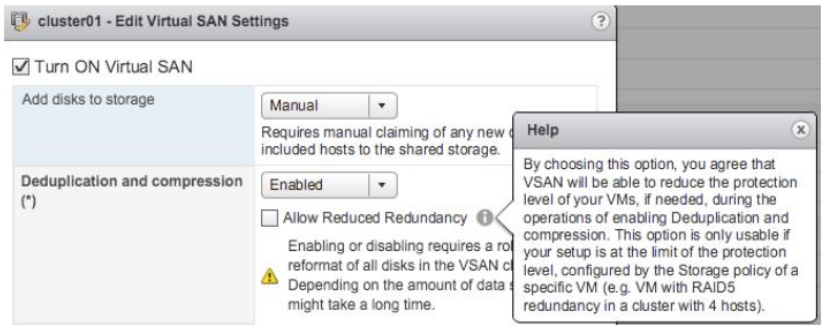
vSAN 6.2 introduced space efficiency technologies, optimized for all-flash configurations, to minimize storage capacity consumption while ensuring performance and availability. This feature includes deduplication, compression, and RAID5/6 erasure coding to reduce capacity consumption, for a lower total cost of ownership, while still ensuring the same levels of availability and performance.

3.2.3 Deduplication and Compression

Enabling deduplication and compression can reduce the amount of storage consumed by as much as seven times. However, the actual reduction value varies, because this depends primarily on the types of data present, number of duplicate blocks, how much these data types can be compressed, and distribution of these unique blocks. For example, video files do not compress well, while documents and spreadsheets typically yield more favorable results. The following figure shows deduplication and compression efficiency being enabled in an all-flash vSAN environment.

Deduplication and compression is a cluster-wide setting that is disabled by default and can be enabled using a simple drop-down menu, illustrated here. Note that a rolling reformat of every disk group on every host in the vSAN cluster is required, which can take a considerable amount of time. However, this process does not incur virtual machine downtime. In addition, deduplication and compression are enabled as a single parameter. It is not possible to enable deduplication or compression individually.

Figure 6. Features: Deduplication and Compression



Deduplication occurs when data is destaged *nearline* from the cache tier to the capacity tier of an all-flash disk group. The deduplication algorithm utilizes a 4K fixed-block size and is performed within each disk group independently. In other words, redundant copies of a block within the same disk group are reduced to one copy, but redundant blocks across multiple disk groups are not deduplicated. The deduplication at the disk group level by vSAN using a 4K block size helps provide a good balance between space efficiency and performance.

The compression algorithm is applied after deduplication has occurred, just before the data is written to the capacity tier flash devices. Considering the additional compute resource and allocation map overhead of compression, vSAN only stores compressed data if a unique 4K block can be reduced to 2K or less. Under all other circumstances, the block is written uncompressed.

There is a storage policy implication to be aware of when deduplication and compression are enabled. This is especially true when upgrading from a previous version of vSAN, where one or more storage policies contain an *object space reservation* rule with a value other than 0 or 100 percent. When deduplication and compression are enabled, object space reservation rules must be set to 0 or 100 percent. Values from 1 to 99 percent are not supported when deduplication and compression are enabled. An object that is assigned a storage policy containing an object space reservation rule of 100 percent is analyzed for deduplication, but no space savings are realized because capacity for the entire object is reserved. Before upgrading vSAN, all policies containing an explicit object space reservation rule must be configured to 0 or 100 percent.

Note The implicit default value for this rule is 0 percent, so there is no need to adjust a policy that does not have the object space reservation explicitly defined.

Naturally, the processes of deduplication and compression on any storage platform require additional CPU cycles and can potentially impact performance in terms of latency and maximum IOPS. vSAN is no exception. However, considering deduplication and compression are only supported in an all-flash vSAN configuration, these effects are negligible in the majority of use cases due to the high levels of performance available from modern enterprise flash devices.

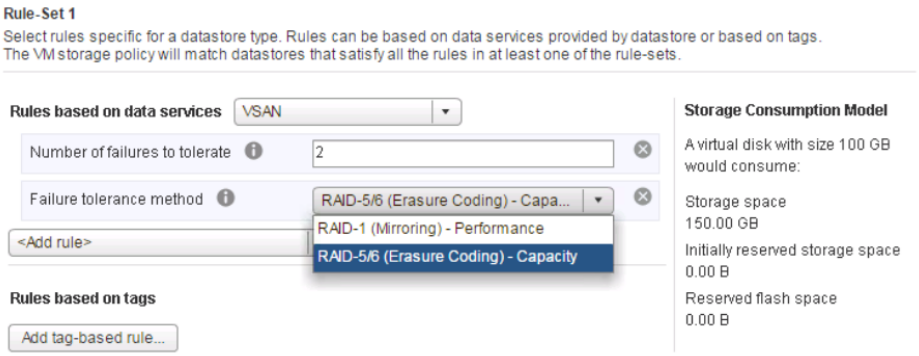
3.2.4 Host and vSphere Cluster Failure Tolerance

vSAN 6.2 introduce a new Storage Policy-Based Management (SPBM) rule, called *fault tolerance method*, which allows virtualization administrators to choose which method of fault tolerance to employ.

Prior to vSAN 6.2, RAID-1 (mirroring) was used as the failure tolerance method. However, vSAN 6.2 introduced RAID-5/6 (erasure coding) to all-flash configurations. While mirroring techniques excel in workloads where performance is the most important factor, they are expensive in terms of capacity required. RAID-5/6 (erasure coding) data layout can be configured to help provide the same levels of availability, while consuming less capacity than RAID-1 (mirroring). Use of erasure coding reduces capacity consumption by as much as 50 percent compared with mirroring at the same fault tolerance level. This method of fault tolerance does require additional write overhead in comparison to mirroring as a result of data placement and parity.

RAID-5/6 (erasure coding) is configured as a storage policy rule and can be applied to individual virtual disks or an entire virtual machine. Note that the failure tolerance method in the rule set must be set to RAID5/6 (erasure coding).

Figure 7. Storage Policy-Based Management Fault Tolerance Methods



RAID-1 (mirroring) in vSAN employs a 2*n*+1 host or fault domain algorithm, where *n* is the number of failures to tolerate. RAID-5/6 (erasure coding) in vSAN employs a 3+1 or 4+2 host or fault domain requirement, depending on 1 or 2 failures to tolerate respectively. RAID-5/6 (erasure coding) does not support 3 failures-to-tolerate. The following table details the host and capacity requirements.

Table 1. vSAN Data Protection Space Consumption

	Tolerated Failures	RAID-1		RAID-5/6		Erasure Coding Space Savings vs Mirroring
		Minimum Hosts Required	Total Capacity Requirement*	Minimum Hosts Required	Total Capacity Requirement*	
FTT=0	0	3	1x	n/a	n/a	n/a
FTT=1	1	3	2x	4	1.33x	33% less
FTT=2	2	5	3x	6	1.5x	50% less
FTT=3	3	7	4x	n/a	n/a	n/a

***** Without Deduplication/Compression taken into account.

Erasure coding can provide significant capacity savings over mirroring, but it is important to consider that erasure coding incurs additional overhead, which is common among any storage platform for this type of data striping. Because erasure coding is only supported in all-flash vSAN configurations, effects to latency and IOPS are negligible in most use cases due to the inherent performance of flash devices.

3.2.4.1 Data Integrity (Software Checksum)

Software checksum enables service providers to detect the corruptions that could be caused by hardware/software components, including memory, drives, and so on during the read or write operations. In case of drives, there are two basic kinds of corruption. The first are *latent sector errors*, which are typically the result of a physical disk drive malfunction. The second type are *silent corruption errors*, which can happen without warning (these are typically called silent data corruption). Undetected or completely silent errors can lead to lost or inaccurate data and significant downtime. There is no other effective means of detection for these errors without an end-to-end integrity checking mechanism.

During the read/write operations, vSAN checks for the validity of the data based on the checksum. If the data is not valid, vSAN takes the necessary steps to either correct the data or report it to the user to take action. These actions could be as follows:

- To retrieve a new copy of the data from other replica of the information, stored within the RAID1, RAID5/6 constructs. This is referred to as recoverable data.
- If there is no valid copy of the data found, an error is returned. These are referred to as non-recoverable errors.

In the case of data errors, issues are reported in the vSphere Web Client user interface and in log files. These include impacted blocks and their associated VMs, allowing the administrator to see the following:

- List of the VMs/blocks that are hit by non-recoverable errors within the vSphere Web Client user interface.
- Historical/trending errors on each drive from the vSphere Web Client user interface.

The data integrity feature uses a CRC32 algorithm, which also supports CPU offload to reduce overhead. In addition, there are two levels of scrubbing employed:

1. Component-level scrubbing – every block of each component is checked. If there is a checksum mismatch, the scrubber tries to repair the block by reading other components.
2. Object-level scrubbing – for every block of the object, data of each mirror (or the parity blocks in RAID-5/6) is read and checked. For inconsistent data, all data in the affected stripe is marked as bad.

The repair can either happen during normal I/O operations by the DOM owner of that object, or by the scrubber, although the repair mechanism for a mirrored or RAID-5/6 operation is different. If checksum verification fails, the scrubber or DOM owner reads the other copy of the data (or other data in the same stripe in case of RAID-5/6), and rebuilds the correct data by writing it out to the bad location.

This end-to-end checksum of the data aims to prevent data integrity issues, which could be caused by silent disk errors, with the checksum being calculated and stored on the write path, or with silent corruptions being detected when reading the data through checksum data.

When checksum verification fails, vSAN automatically read a different copy of the data (or other data in the same stripe in the case of RAID-5/6), then rebuild the corrected data, writing it out to the bad location, based on a 4KB block size.

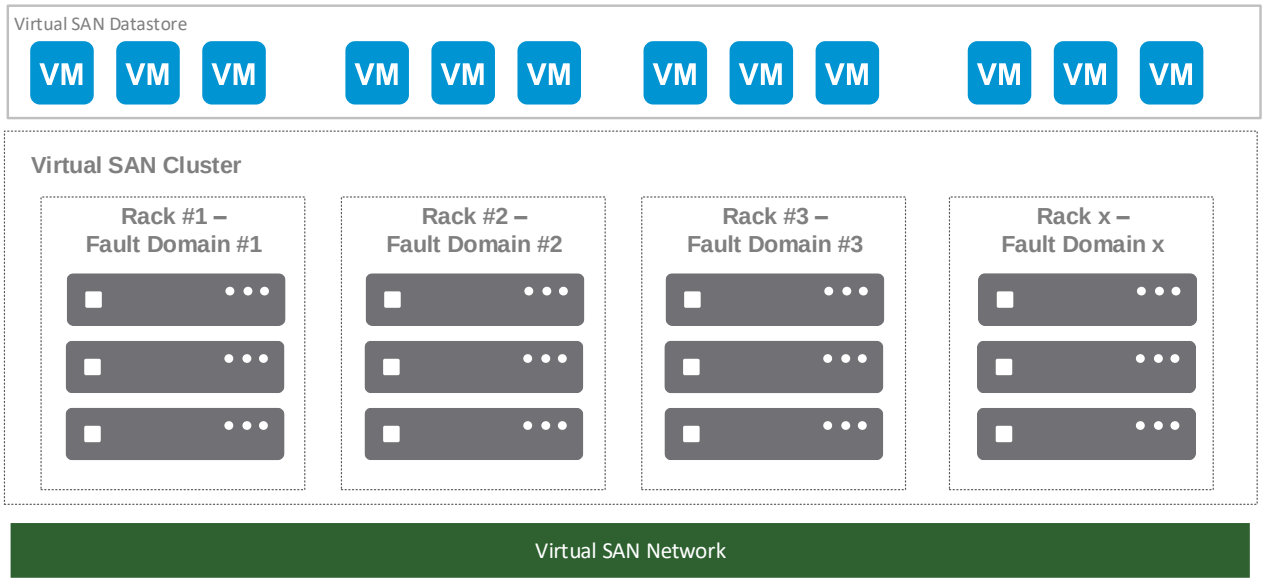
Data Integrity (Software Checksum) is a cluster wide setting, which by default is switched on. However, this value can be disabled on a per-object basis, using storage policies.

3.2.4.2 Rack Awareness (Fault Domains)

vSAN fault domains provide the ability to tolerate rack failures and power failures in addition to disk, network, and host failures.

The idea behind fault domains is to tolerate groups of hosts (chassis or racks) failing without requiring additional data copies. Their implementation allows vSAN to save replica copies of the virtual machine data in different domains, for example, different racks of compute. An example of using fault domains is shown in the following figure.

Figure 8. Rack Awareness (Fault Domains) Example



When working with fault domains, to tolerate *n* number of failures, rather than needing 2*n*+1 hosts, you require 2*n*+1 fault domains instead. This allows for replica data to be spread out across the domains rather than hosts and increases the ability to handle a greater than single host failure.

A minimum of three fault domains are required, however, VMware recommends having four or more, for redundancy purposes. In addition, to have a minimum level of redundancy, VMware also recommends having six hosts (two per fault domain).

Fault domains are used to further protect the environment from failure, assuming there are enough hosts to properly support the configuration. Further details on fault domains can be found in the *VMware Virtual SAN Design and Sizing Guide* (https://www.vmware.com/files/pdf/products/vsan/VSAN_Design_and_Sizing_Guide.pdf).

3.2.5 Snapshots and Clones

vSAN 6.0 introduced a highly efficient VM-centric snapshot and clone mechanism with support for up to 32 snapshots per clones per VM and 16K snapshots per clones per cluster. The new snapshot and clones offer performance improvements over the previous versions. For more information about snapshots in vSAN 6.0 and later, see the following two papers:

- *vsanSparse – Tech Note for Virtual SAN 6.0 Snapshots*

  https://www.vmware.com/files/pdf/products/vsan/Tech-Notes-Virtual-San6-Snapshots.pdf
- *VMware Virtual SAN Snapshots in VMware vSphere 6.0*

  https://www.vmware.com/files/pdf/techpaper/vsan-snapshots-vsphere6-perf.pdf

3.2.6 Swap Efficiency

Virtual swap files are created when virtual machines are powered on. In cases where physical host memory is exhausted, the virtual swap file is used in place of physical memory for a virtual machine. Virtual swap files are sized according to the allocated memory minus reserved memory.

A virtual machine with 4 GB of RAM allocated and a 2-GB memory reservation will create a 2-GB virtual swap file. On vSAN, that swap file is created with a mirrored policy, resulting in 4 GB of space consumed. 100 virtual machines with the same configuration consume 400 GB of capacity. In large service provider

deployments of thousands of virtual machines, this additional capacity could be substantial and require significant capacity.

In addition to the use of deduplication and compression, as well as erasure coding, it can be advantageous to use the vSAN *SwapThickProvisionedDisabled* advanced host setting for additional space savings.

Enabling this setting creates virtual swap files as a sparse object on the vSAN datastore. Sparse virtual swap files only consume capacity on vSAN as they are accessed. The result can be significantly less space consumed on the vSAN datastore, provided virtual machines do not experience memory over commitment, requiring use of the virtual swap file.

3.2.7 Quality of Service

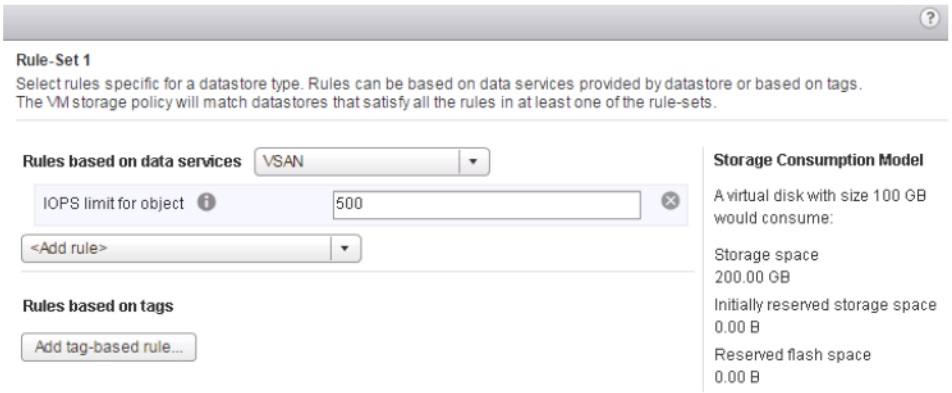
vSAN 6.2 adds a quality of service feature that limits the number IOPS an object can consume. In underutilized configurations, limits might or might not be necessary, because objects likely have sufficient resources to effectively meet the needs of their workload. While it is entirely desirable to have more than enough resources, it does not come without cost. Efficiently sized configurations are typically a good mix of cost and available resources. The metrics of appropriate resources for workloads can change over time, especially as utilization grows, or as workloads are added over the lifecycle of a platform.

There are several situations where it might be advantageous to limit the IOPS of one or more virtual machines. The term *noisy neighbor* is often used to describe when a workload monopolizes available I/O or other resources, which negatively affects other workloads on the same platform. In environments where there is a mix of both low and high utilization, it could be detrimental for a virtual machine with low utilization during normal operations to change its pattern and start to consume massive resources, in turn starving others for enough to operate properly. Particularly at a larger scale, the impact of this situation might affect multiple business units, tenants, or customers.

A good example of a noisy neighbor situation includes end-of-month reporting. Consider those virtual machines generating reports residing on the same four-node Hybrid vSAN cluster as other Tier 1 applications that have stringent service level agreement requirements. What would happen if one or more large reports were generated, consuming a larger percentage of IOPS than were available for the minimum service level requirements of the Tier 1 applications? It is quite possible the Tier 1 applications would not be able to satisfy the requirements of the service level agreement. Implementing IOPS limits on the reporting solution VMs could prevent the situation where the reporting solution starves the Tier 1 applications for IO.

With the Quality of Service addition to vSAN 6.2, IOPS limits are now available. Quality of service for vSAN 6.2 is a Storage Policy-Based Management (SPBM) rule. Because quality of service is applied to vSAN objects through a storage policy, it can be applied to individual components or the entire virtual machine without interrupting the operation of the virtual machine.

Figure 9. vSAN QoS Enforced by SPBM



Quality of service for vSAN is normalized to a 32-KB block size, and treats reads the same as writes. An example with an IOPS limit of 500 (regardless of block size up to 32 KB) results in 500 IOPS, while a block size of 64 KB results in 250 IOPS. It is important to consider the workload profile when configuring IOPS limits.

3.2.8 Data Locality (Locality of Reference)

In computer science, *data locality*, also known as *locality of reference,* is the behavior of computer programs according to which a workload accesses a set of data entities or storage locations within some period of time with a predictable access pattern.

There are two main types of data locality:

- Temporal locality – The probability that if some data (or a storage location) is accessed at one point in time, it will be accessed again soon afterwards.
- Spatial locality – The probability of accessing some data (or a storage location) soon after some nearby data (or a storage location) on the same medium has been accessed. Sequential locality is a special case of spatial locality, where data (or storage locations) are accessed linearly and according to their physical locations.

Data locality is particularly relevant when designing storage caches. For example, flash devices offer impressive performance improvements, at a cost, so efficient use of these resources becomes an important design factor.

Like any storage system, vSAN makes use of data locality. vSAN uses a combination of algorithms that take advantage of both temporal and spatial locality of reference to populate the flash-based read caches across a cluster and provide high performance from available flash resources.

Examples include:

- Every time application data is read by a virtual machine, vSAN saves a copy of the data in the Read Cache portion of the flash device associated with the disk group where the copy of the data resides. Temporal locality implies that there is high probability that said data will be accessed again before long. In addition, vSAN predictively caches disk blocks in the vicinity of the accessed data (in 1MB chunk at a time) to take advantage of spatial locality as well.
- vSAN uses an adaptive replacement algorithm to evict data from the Read Cache when it is deemed unlikely that the data will be accessed again soon, and uses the space for new data more likely to be accessed repeatedly.
- vSAN makes replicas of storage objects across multiple servers for protection purposes. Reads are distributed across the replicas of an object for better load balancing. However, a certain range of logical addresses of an object is always read from the same replica. This approach has two important benefits:
    - Increases the chances that the data accessed is already in the Read Cache.
    - A data block is never cached in more than one flash device.

To be clear, a fundamental design decision for vSAN is to not implement a persistent client-side local read cache. The decision was based on the following observations regarding local read caching on flash:

- Local read caching results in very poor balancing of flash utilization (both capacity and performance) across the cluster.
- Local read caching requires transferring hundreds of gigabytes of data and cache re-warming when virtual machines are migrated using vSphere vMotion between hosts to keep compute resources balanced.
- Local read caching offers negligible practical benefits in terms of performance metrics, such as latency.

3.2.9 Scalability

vSAN 6.0 or later scales up to 64 nodes per cluster for both hybrid and all-flash architectures and matches vSphere node/cluster supportability. You can scale up to 200 virtual machines per host and 6400 virtual machines per cluster, for both hybrid and all-flash architectures. The maximum virtual disk (VMDK) size is 62 TB, matching classic VMFS/NFS VMDK size.

3.2.9.1 Support for High-Density Storage Systems with Direct-Attached JBOD

vSAN allows the management of this type of externally connected disk enclosures, therefore potentially leveraging an existing investment in a blade-based architecture.

3.2.9.2 Capacity Planning

What-if scenario analysis and reporting on how much of the vSAN datastore (flash device and mechanical disk capacity) has been or will be utilized when a virtual machine storage policy is created or edited is fully supported within VMware vRealize® Operations Manager™ through the VMware vRealize Operations Management Pack™ for Storage Devices.

3.2.9.3 vSAN 6.2 Scalability Limits

Requirements for vSAN scalability are shown in the following table. Always validate these limits with the latest updates to the *VMware vSphere 6.0 Configuration Maximums* guide (https://www.vmware.com/pdf/vsphere6/r60/vsphere-60-configuration-maximums.pdf).

Table 2. vSAN Scalability Limitations

Option	Limit
vSAN ESXi host	
vSAN disk groups per host	5
Mechanical disks per disk group	7
SSD disks per disk group	1
Spinning disks in all disk groups per host	35
Components per vSAN host	9000
vSAN Cluster	
Number of vSAN nodes in a cluster	64
Number of datastores per cluster	1
vSAN virtual machines	
Virtual machines per host	200
Virtual machines per cluster	6400
Virtual machine virtual disk size	62 TB
vSAN VM storage policy	

Option	Limit
Disk stripes per object	12
Percentage of flash-read cache reservation	100
Failures-to-tolerate	3 if VM disk is <= 16 TB 1 if VM disk is > 16 TB
Percentage of object space reservation	100
Virtual networking	
vSAN networks/physical network fabric	2

Note These scalability limits for the release of vSAN version 6.2 are correct at time of this writing.

3.2.10 Enterprise Availability and Data Protection

3.2.10.1 vSphere High Availability Integration

vSphere HA and vSAN on the same cluster is fully supported and interoperable, as with traditional datastores, and vSphere HA provides the same level of protection for virtual machines on vSAN datastores. However, this level of protection imposes specific restrictions when vSphere HA and vSAN interact.

In addition, vSAN uses its own logical network. Therefore, when vSAN and vSphere HA are enabled for the same cluster, the HA inter-agent traffic flows over this storage network rather than the management network. vSphere HA uses the management network only when vSAN is disabled. vCenter Server automatically chooses the appropriate network when vSphere HA is configured on a host.

Note You must disable vSphere HA before you enable vSAN on the cluster. Then you can re-enable vSphere HA.

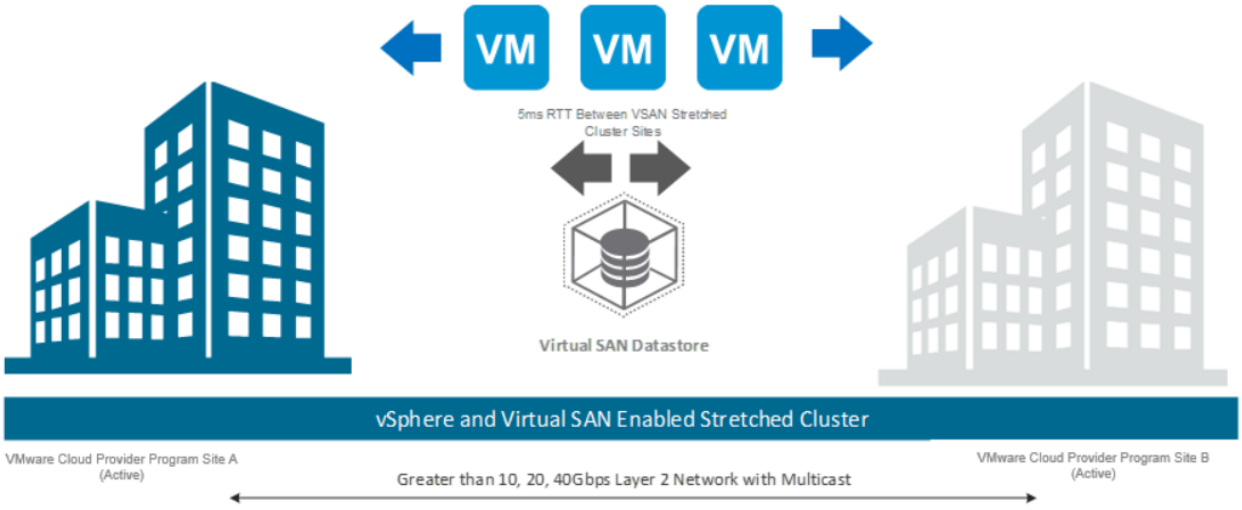
3.2.10.2 vSAN Stretched Cluster

vSAN 6.1 introduced the ability to create a stretched cluster between two or more geographically separated sites, synchronously replicating data between sites.

This stretched cluster feature introduced enterprise-level availability where an entire site failure can be tolerated, with no data loss and near zero downtime. In addition, the feature enables disaster and downtime avoidance by allowing the proactive migration of virtual machines from one site to another in order to avoid an impending outage, or for planned maintenance purposes.

vSphere vMotion and vSphere DRS are natively integrated into a stretched cluster architecture and can help load balance workloads between two active sites, while vSAN insures high performance by making sure that virtual machines use the local site copy for read and write operations.

Figure 10. vSAN Stretched Cluster



Because of the tight integration with the VMware advanced features like vSphere vMotion, vSphere DRS and vSphere HA, the stretched cluster solution works seamlessly with little additional operational overhead, and expands the knowledge of vSphere administrators, who are already familiar with how to operate and run everything around vSphere.

vSAN removes all the complexity around storage management traditionally seen in a multisite environment and also simplifies storage tasks to a few clicks in the vSphere Web Client.

3.2.10.3 Remote and Branch Offices

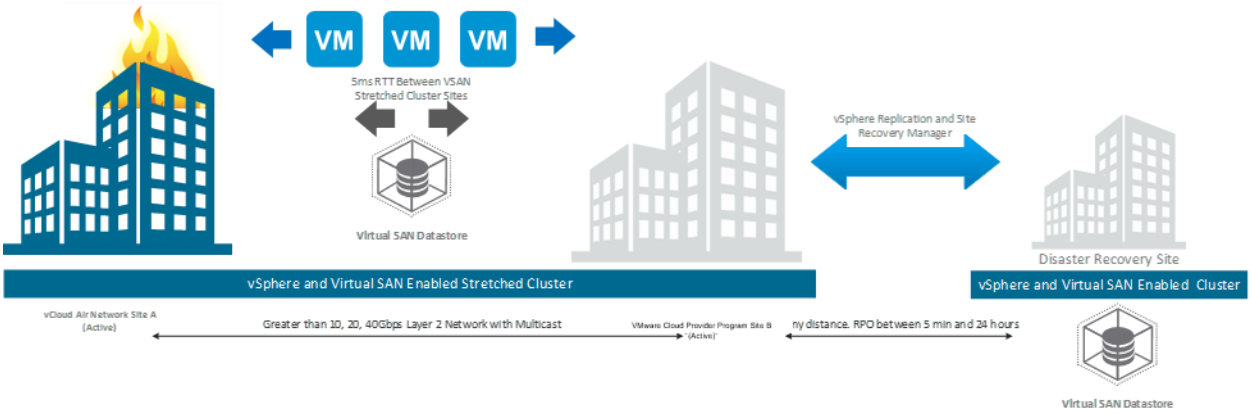
vSAN provides the ability and support to deploy vSAN clusters for ROBO scenarios. VMware Cloud Providers can now deploy a large number of two-node vSAN clusters that can be centrally managed from a centralized data center, through an individual vCenter Server instance.

For a more detailed use case description of two-node cluster architecture for service providers, see the *VMware vSAN Two-Node Architecture Service Provider Use Cases* vCAT-SP paper (http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vcat/vmware-virtual-san-two-node-architecture-service-provider-use-cases.pdf).

3.2.10.4 vSAN Replication

vSAN can also utilize VMware vSphere Replication™ for its replication data services. vSphere Replication with vSAN supports Recovery Point Objectives (RPOs) from 5 minutes to 24 hours. VMware Site Recovery Manager™ can be used as part of a solution to deliver a fully orchestrated disaster recovery solution for tenant workloads.

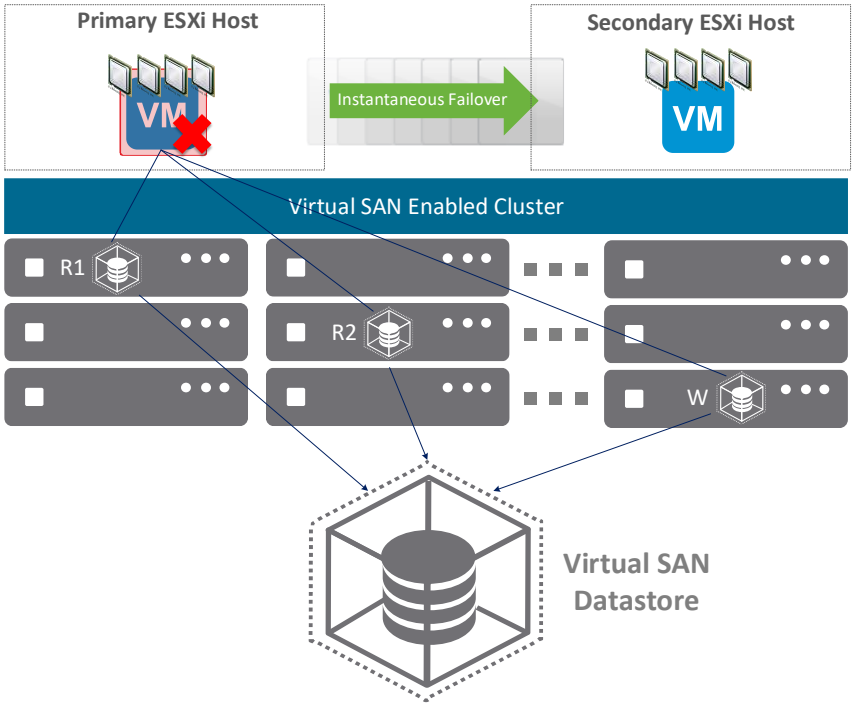
Figure 11. Three-Site Disaster Recovery Architecture



3.2.10.5 Support for Multi-Processor Fault Tolerance

vSAN is able to support the vSphere multi-CPU Fault Tolerance (SMP-FT) feature and provide continuous availability with zero downtime for mission critical applications, even in the event of hardware failure.

Figure 12. vSAN with Fault Tolerance



3.2.11 Encryption

For support information on controller-based *data-at-rest* encryption to protect sensitive data on the disk using controller-based encryption, see vSAN HCL for certified controllers (https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan).

3.2.11.1 HyTrust DataControl

With the vSAN 6.2 release, VMware recommends HyTrust to further protect vSAN environments.

HyTrust DataControl is a complete cloud and virtualization encryption solution including both encryption and key control, and is available for both Windows and Linux. Initial installation takes just a few seconds and data is securely encrypted at rest and in motion from the point of install until the virtual machine is eventually securely retired.

Because the encryption software is part of the virtual machine, encryption travels with the VM from one physical host to another, or from private to public cloud and back again. Operationally DataControl is transparent, allowing encryption and rekeying “on-the-fly” with no need to reboot, dismount, or incur downtime.

3.2.12 Maximum Performance and Low Latencies

As flash and non-volatile memory technologies continue to evolve rapidly, with the emergence of new flash form factors, and new flash interfaces, these technologies have the potential to further reduce storage latencies by an order of magnitude. In fact, read and write latencies will soon become much faster than doing a network hop. As a result, the only way to really leverage the enormous performance gains of these new technologies is to eliminate that network hop and bring the data much closer to the compute—onto the same server.

VMware continues the performance improvement of vSAN by expanding the support of new types of hardware devices and interfaces. For instance:

- ULLtraDIMM – ULLtraDIMM SSDs connect flash storage to the memory channel via DIMM slots, achieving very low (<5us) write latency. UltraDIMM provides even greater density and performance. For example, it allows for a 12-TB all-flash vSAN host in a thin blade form factor, as well as three times improvement in latency compared to external arrays.
- NVMe – Non-volatile Memory Express (NVMe) is a new communications interface developed especially for SSDs. NVMe provides greater parallelism for both hardware and software, and as a result, enables performance improvements. Leveraging NVMe in a vSAN all flash deployment resulted in 3.2M IOPS measured on a 32-node cluster ~100K IOPS/host.

3.2.13 Application Support

vSAN is a great fit for all types of workloads, large or small. VMware Cloud Providers are adopting vSAN at an increasing rate since its release. vSAN has proven to be a robust and effective platform for business critical applications, virtual desktop infrastructure (VDI), test and development, and a multitude of other use cases.

As vSAN is increasingly utilized for business critical applications, it has been important to provide additional certification, support, and validation specific to several key workloads. Examples include support for Microsoft Windows Server Failover Clustering, when using a File Share Witness, as well as, support for Oracle Real Application Clusters. vSAN 6.2 also introduces support for core SAP applications.

Business critical applications, virtual desktops, and other application types have increased scalability with vSAN, without the sizing challenges often associated with legacy storage platforms. As a result, vSAN built on vSphere can easily be sized for small workloads, medium or large workloads, and anything in between.

3.2.14 vSAN Hybrid Architecture

The vSAN hybrid architecture was the original disk group configuration for vSAN, and uses mechanical disks for the capacity layer and flash memory devices for the cache layer.

A vSAN hybrid configuration uses the local attached storage to create a cost effective, highly available and high-performance shared storage for virtual machines. The flash devices are used for read cache and write buffer, while mechanical HDD drives are used for the persistent capacity storage requirements of virtual disk files (VMDKs), snapshots and swap. This option is ideal for workloads that have typical storage performance requirements.

3.2.15 vSAN All-Flash Architecture

vSAN 6.0 introduced the ability to create an all-flash architecture in which flash cache devices and flash capacity devices are intelligently used as a write cache and also provide high endurance data persistence.

This all-flash architecture allows tiering of PCI-e devices—a write-intensive, high endurance performance tier for the writes, and a read-intensive, cost-effective capacity tier for data persistence, thereby reducing the overall cost of an all-flash architecture.

This approach with a vSAN 6.0 all-flash architecture provides consistent, predictable performance, with up to 90K IOPS per host and sub-millisecond response times, making it ideal for Tier 1 or business-critical workloads.

3.2.16 Comparing Hybrid and All-Flash Configurations

As described previously, vSAN 6.0 and later provides support for two different configuration options—a hybrid configuration that leverages both flash-based devices and mechanical disks, and an all-flash configuration. The hybrid configuration is the traditional approach used in the previous release of vSAN. Hybrid configurations use a flash-based device to provide the caching tier, and mechanical disks to provide capacity and persistent data storage. An all-flash configuration employs entirely flash-based devices for both the caching and capacity tiers. The cost of using all-flash configurations can be significant, because 10-GbE networking is an absolute requirement, in addition to the additional costs of flash storage.

The all-flash vSAN configuration brings improved, highly predictable and uniform performance, regardless of workload, and additional features when compared with hybrid configurations. However, it do not automatically assume that an all-flash configuration delivers improved performance over a hybrid design. While it is true to say that performance is more consistent when employing an all-flash environment, the actual performance, in terms of read and write latency, is very much dependent on the workload's dataset, the hardware being employed, and overall design.

Both hybrid and all-flash clusters carry a 10 percent of consumed capacity recommendation for the flash cache layer, even though cache is used differently in each configuration:

- Hybrid clusters – The caching algorithm attempts to maximize both read and write performance. By default, 70 percent of the cache is allocated for storing frequently read disk blocks, which minimizes accesses to slower mechanical disks. The remaining 30 percent is used for write caching.
- All-flash clusters – Employ two types of flash: very fast and durable write cache, and larger and cost-effective capacity flash. In an all-flash configuration, 100 percent of cache is allocated for writes, because read performance from capacity flash is more than sufficient to handle demanding enterprise workloads. In addition, far more writes are held by the cache, and written to the capacity layer only when needed, extending the life of the capacity flash tier.

Consider the following general guidelines about drives (flash or mechanical):

- Compatibility – The model of the PCIe or SSD devices must be listed in the vSAN section of the *VMware Compatibility Guide*.

- Performance – PCIe devices generally have faster performance than SSD devices.
- Capacity – The maximum capacity available for PCIe devices is generally greater than the maximum capacity that is currently listed for SSD devices for vSAN in the *VMware Compatibility Guide*.
- Write endurance – The write endurance of the PCIe or SSD devices must meet the requirements for capacity or for cache in all-flash configurations, and for cache in hybrid configurations.
- Cost – PCIe devices generally have higher cost than SSD devices.

In addition to these general guidelines, the following are considerations for using all-flash configurations:

- vSAN 6.0 at a minimum must be used.
- 10-GbE network is required.
- Maximum number of all-flash nodes is 64.
- Flash Read Cache reservation SPBM capability is *not* used.
- All drives must be marked as flash.
- Drive endurance becomes an important design consideration.

vSAN Monitoring

Monitoring the vSAN environment is critical to the service provider's successful implementation of the hyper-converged infrastructure.

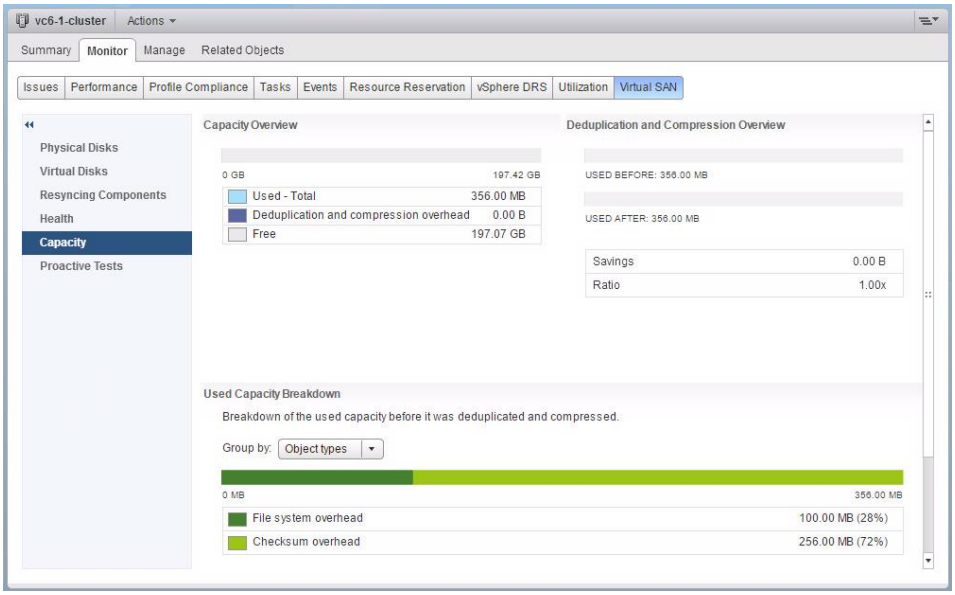
Because vSAN is a policy-driven storage solution, provisioning and management are significantly simplified. vSAN automatically and dynamically matches requirements with underlying storage resources. With vSAN, many manual storage tasks are automated to deliver a more efficient and cost-effective operational model. However, from an operational aspect, providers have additional concerns and requirements that go beyond the simplification and streamlining of management. vSAN enables additional capabilities that allow providers to monitor and manage their vSAN infrastructure. This section describes considerations for monitoring best practices and design.

vSAN supports a number of monitoring capabilities. For basic monitoring of the datastores, vSAN can be monitored from the vSphere Web Client, where monitoring can occur on different objects, including clusters and datastores. When hosts participate in a vSAN cluster, they can also be monitored in the same way as any other host that vCenter Server manages.

From the cluster level, you can monitor the hosts' physical disks and virtual disks that are participating with vSAN. When reviewing at the cluster level, capacity, operational status, health status, policy information, and compliance status can be seen for each host, VM, and disk on the vSAN datastore.

At the vSAN datastore level, you can view the standard performance information, state of the disks, status of the volume, and partition group information. In addition to these statistics, in vSAN 6.2, there is also a capacity monitoring page available as well as a number of new graphs and data points that provide performance information at the cluster, host, virtual machine, and virtual disk levels. In addition, the Time Range can also be modified to show information from the last few hours or a custom date and time range.

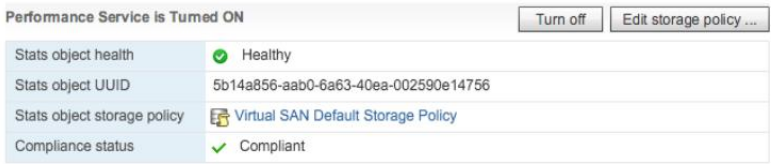
Figure 13. Capacity Overview Interface



4.1 Performance Service

The performance service is enabled at the cluster level, and it is important to remember that the performance history database is not a component of vCenter Server. The performance history database is stored as a vSAN object, independent of vCenter Server. A storage policy is assigned to the object to control space consumption, availability, and performance of that object. If the object becomes unavailable, performance history for the cluster is unavailable until access to the object is restored. The performance history database can consume up to 255 GB of capacity on the vSAN datastore.

Figure 14: Configuration of Performance Service

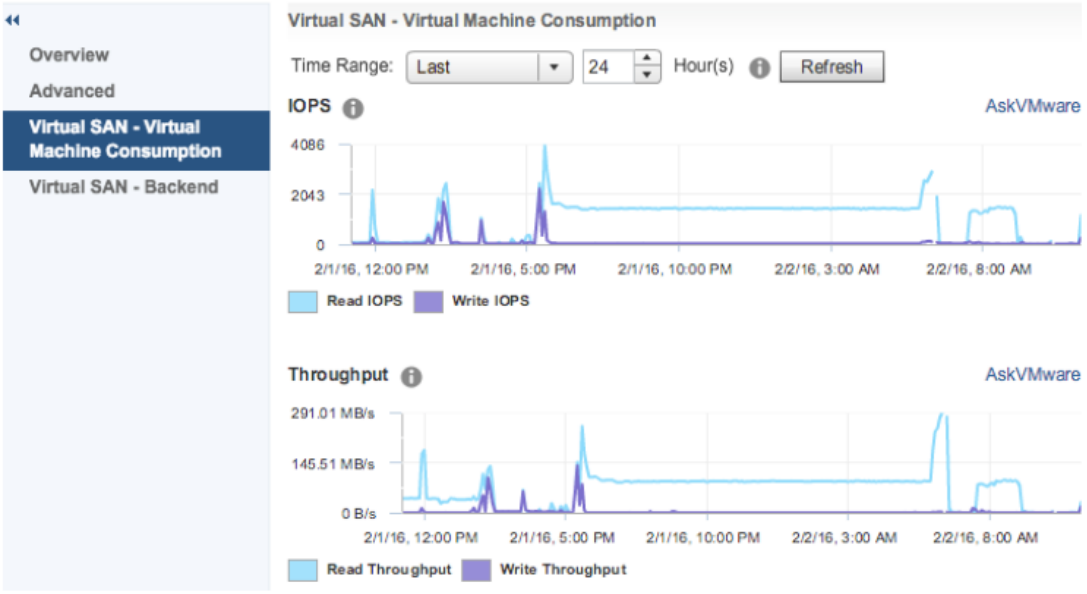


4.1.1 Cluster Metrics

At the cluster level, the performance monitoring service shows performance metrics for virtual machines running on vSAN as well as the vSAN back end. These metrics provide quick visibility to the entire vSAN cluster, showing how vSAN and the objects that reside on it are performing.

The following figure shows metrics such as IOPS, throughput, latency, and outstanding I/O not only for virtual disks, but for all vSAN objects in the cluster.

Figure 15. Performance Metrics



Back end vSAN metrics help show what is required to deliver the expected performance, and what is visible at the virtual machine and object level. It is noteworthy that the I/O overhead from different storage policies can show significantly different back end metrics than what is observed at the virtual machine or object level.

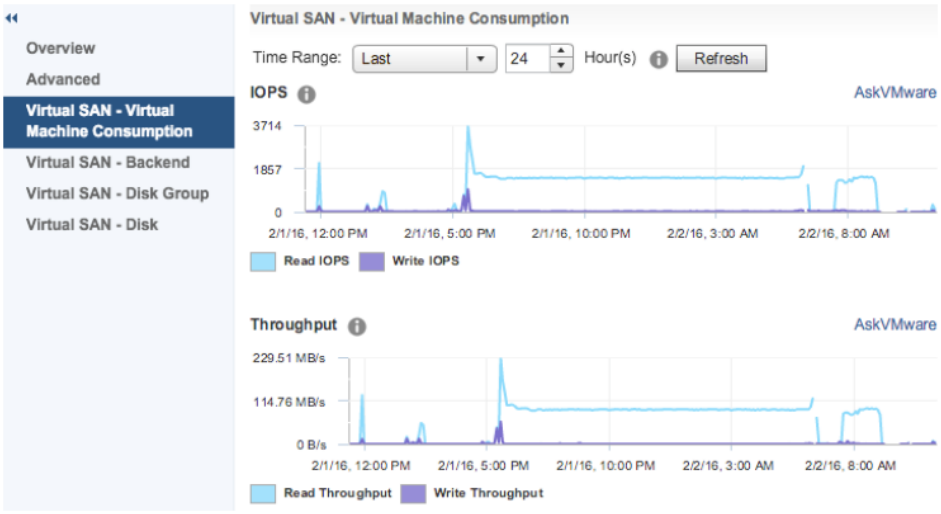
Take the example of a virtual machine with a Number of Failures to Tolerate of 1, with a failure tolerance method of RAID-1 (Mirroring). For every write I/O to a virtual disk, two are seen on the back end. This is because two mirrors both receive a write, despite the fact that the virtual disk only had a single write.

Changing the fault tolerance method to RAID-5/6 (erasure coding) now requires four writes, because data and parity in a RAID-5 configuration is comprised of three data writes and one parity write.

4.1.2 Host Metrics

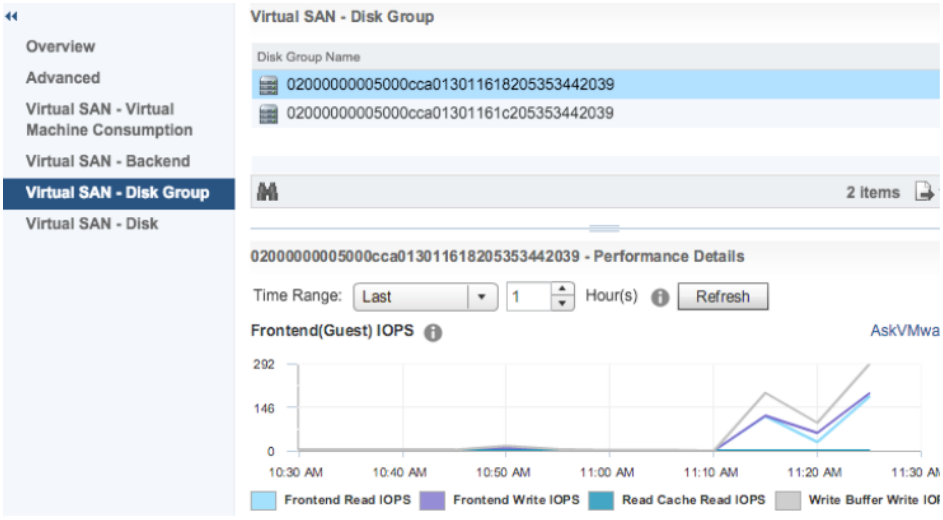
While visibility of cluster performance is important, so too is performance at the host level. Virtual machine consumption as well as back end metrics are visible specific to the host selected, as well as additional metrics.

Figure 16. Performance Metrics per Virtual Machine



In addition, because vSAN requires at least one disk group, but can be configured with up to five disk groups on each host in the cluster, it is important to be able to distinguish, from an operational perspective, how each disk group is performing, independent of any other disk groups on the same host.

Figure 17. Performance Metrics per Disk Group



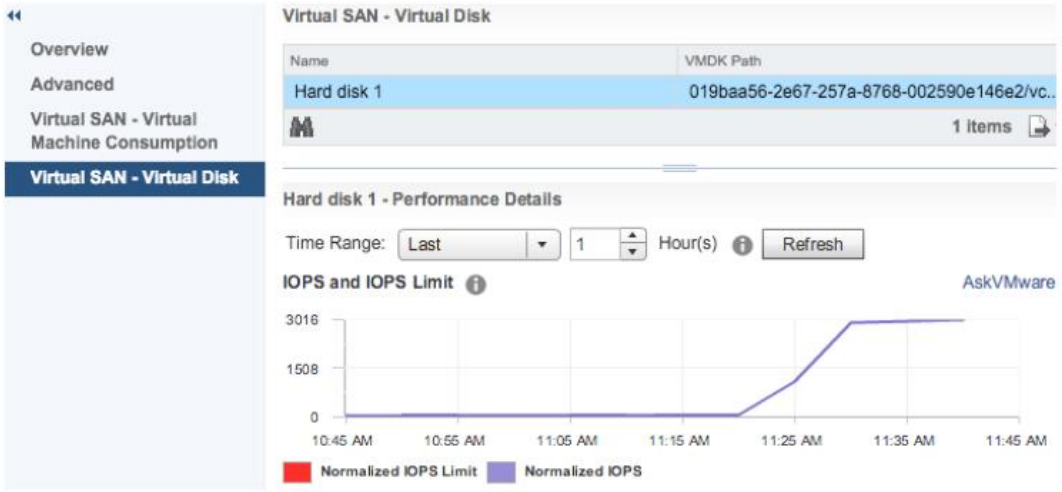
As previously outlined, each disk group in a vSAN host requires one cache device and one to seven capacity devices. Just as it is important to be able to distinguish different disk groups' performance, it is also important to have visibility to the performance metrics of a single device.

4.1.3 Virtual Machine Metrics

Performance metrics are also available for all of the upper-level components of vSAN. With vSAN 6.2, we can also view performance data at the cluster, host, disk group, and devices. In addition, as illustrated in Figure 16, it also provides the same visibility for specific virtual machines and their respective disks.

Being able to view metrics at this level of virtual machine consumption shows what the virtual machine is experiencing from an application perspective. Then, going deeper into the individual metrics which are available on a per disk basis provides even further granularity, as it is very common to have virtual machines configured with more than one virtual disk. Of course, with vSAN leveraging Storage Policy-Based Management (SPBM), it is entirely possible to have completely different storage policies applied to different virtual disks. However, different properties like an SPBM policy, workload, capacity, and more can also contribute to the performance characteristics of a virtual disk.

Figure 18. Performance Metric per Virtual Disk



Capacity details are also available in the vSAN user interface making it easy for administrators to understand how much capacity various object types are consuming, as illustrated in the following figure.

Figure 19. Capacity Metrics

vSAN includes the vSAN API, which is an extension of the vSphere API. The vSAN API centers on a small set of managed objects, which enable administrators to query runtime state, as well as configure vSAN. The API is exposed as a web service, running on both vCenter Server systems and ESXi systems. Managed objects are available for cluster-level and host-level operations.

4.2 Health Service

vSAN 6.2 also features an improved health service. The health service actively tests and monitors a number of items such as:

- Hardware compatibility
- Network connectivity
- Cluster health
- Capacity consumption
- Limits health
- Physical disk health

The health service is enabled by default and configured to check the health of the vSAN environment every 60 minutes. The plug-in provides details on:

- Whether the vSAN deployment is fully supported, functional, and operational
- Immediate indications to a root cause in the event of a failure, leading to a speedier resolution time
- The ability to test different aspects of the configuration

The health service is quite thorough in the number of tests it performs. As an example, proper network configuration is key to a healthy vSAN cluster and there are 11 tests in the **Network** section of the vSAN Health user interface.

If an issue is detected in the environment, a warning is visible in the vSAN user interface. Clicking on the warning provides more details about the issue. For example, a controller driver that is not on the hardware compatibility list for vSAN will trigger a warning. In addition to providing details about the warning, the vSAN Health Service user interface also has an **Ask VMware** button, which brings up the relevant VMware Knowledge Base article.

vSphere and vSAN support a wide variety of hardware configurations. The list of hardware components and corresponding drivers that are supported with vSAN can be found in the *VMware Compatibility Guide*. It is very important to use only hardware, firmware, and drivers found in this guide to help maintain stability and performance of a vSAN environment. The list of certified hardware, firmware, and drivers is contained in a hardware compatibility list (HCL) database. The vSAN user interface makes it very easy to update this information, for use by the Health Service tests. If the environment has internet connectivity, updates can be obtained directly from VMware. Otherwise, HCL updates can be downloaded as a file to enable offline updates.

Figure 20. Health Service – Physical Server Specifications Support

Finally, if an issue does arise that requires the assistance of VMware Support, it is very easy to upload support bundles to help expedite the troubleshooting process. Clicking the **Upload Support Bundles to Service Request** button enables an administrator to enter an existing support request (SR) number and upload the necessary logs with just a few mouse clicks.

4.3 VSAN Observer

For more granular information on how vSAN is operating, you can utilize VSAN Observer, a tool that provides deep visibility into vSAN performance metrics and counters. This tool is included with vSphere 6.0. As a part of this release, the Ruby vSphere Console (RVC) provides an interactive command interface that can be used to manage, monitor, and troubleshoot. RVC includes functions for the following:

- vSAN configuration
- vSAN health monitoring
- vSAN disks statistics
- vSAN performance statistics
- VSAN Observer

You can use VSAN Observer to monitor information on vSAN. Observer is a graphical user interface utility that displays vSAN-related statistics from the vSAN client perspective. It can be used to understand vSAN performance characteristics and for analysis.

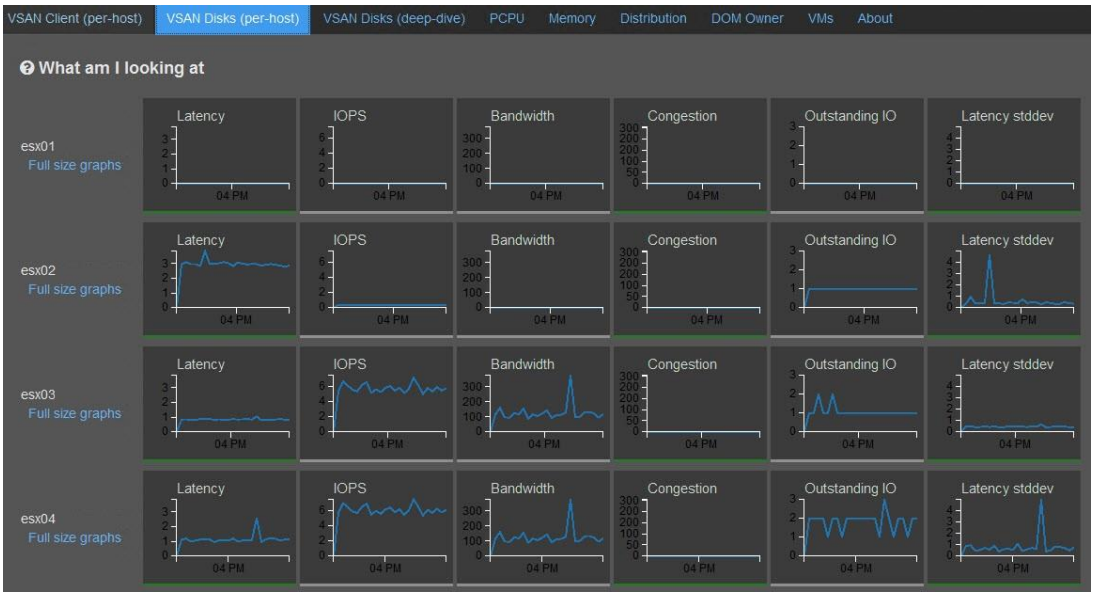
Note VSAN Observer is primarily a troubleshooting tool. VMware does not recommend running it continuously.

The VSAN Observer user interface displays performance information for the following:

- Statistics of the physical disk layer
- Extensive physical disks group details
- CPU usage statistics
- Consumption of vSAN memory pools
- Physical and in-memory object distribution across vSAN clusters

An example of the interface is shown in the following figure.

Figure 21. VSAN Observer



For more detailed information on troubleshooting using the VSAN Observer tool, see the following VMware documentation:

- *VMware Virtual SAN Diagnostics and Troubleshooting Reference Manual*

  http://www.vmware.com/files/pdf/products/vsan/VSAN-Troubleshooting-Reference-Manual.pdf

- *VMware Ruby vSphere Console Command Reference for Virtual SAN*

  https://www.vmware.com/files/pdf/products/vsan/VMware-Ruby-vSphere-Console-Command-Reference-For-Virtual-SAN.pdf

- *Monitoring VMware Virtual SAN with Virtual SAN Observer*

  http://blogs.vmware.com/vsphere/files/2014/08/Monitoring-with-VSAN-Observer-v1.2.pdf

4.4 vRealize Operations Manager Monitoring

The vRealize Operations Manager Management Pack for Storage Devices (MSPD) supports the collection of data from vSAN. The management pack can also connect to any storage device that has a vSphere API for Storage Awareness provider as well as SAN/NAS switches from Brocade or Cisco using SMI-S. Performance data is also collected from host HBAs, NICS, VMs, and SAN/NAS switches.

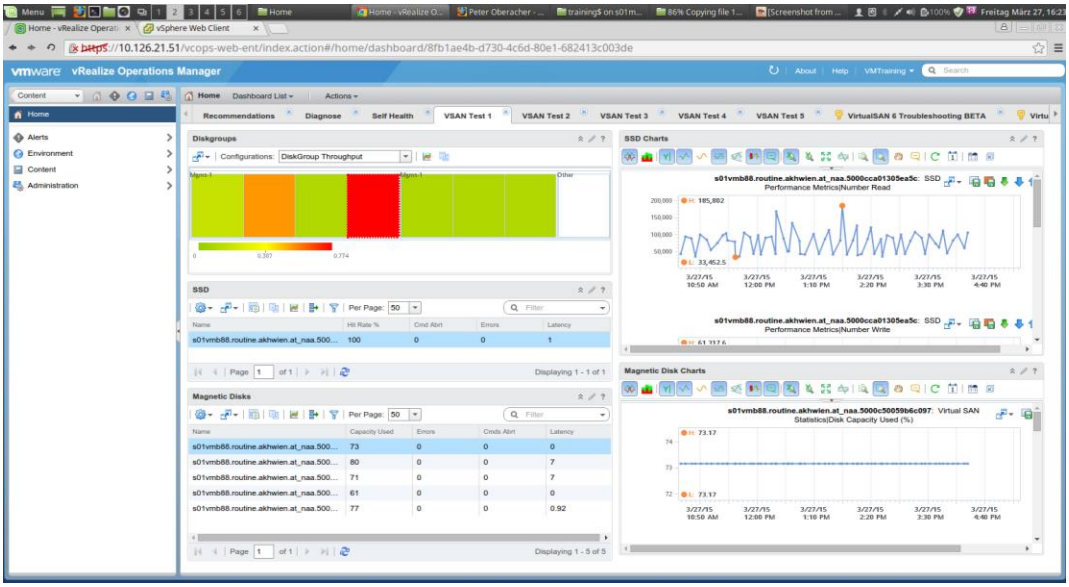
In typical vRealize Operations fashion, out-of-the-box monitoring is provided across clusters, HDD/SSD devices, network, CPU, and memory components. The vRealize Operations Management Pack for Storage Devices is located in the Cloud Marketplace on the *VMware Solution Exchange* web site at https://solutionexchange.vmware.com/store

Discover more at https://blogs.vmware.com/management/2015/08/mpsd-management-pack-for-storage-devices-for-vrealize-operations-available-now.html.

The MPSD allows for comprehensive monitoring of a vSAN environment with vRealize Operations Manager creating dashboards and alerts as required. The default set of dashboards include:

- Troubleshooting – Shows state of the cluster and any anomalies
- Heatmap – Can be used to show heatmap based on different metrics, such as throughput (cluster, controller and disk level)
- Entity Usage – Throughput, latency, bandwidth metrics across controllers, flash devices, and mechanical disks. Useful for locating bottlenecks
- Device insights – Includes SMART (Self-Monitoring, Analysis and Reporting Technology) metrics, such as wear levelling or reallocated sector count usage
- Cluster Insights – Metrics on cluster as a whole

Figure 22. vRealize Operations with Management Pack for Storage Devices (MPSD)



vSAN Design Overview

Designing storage resources for a cloud model differs from the traditional vSphere approach used in defining storage for non-cloud onsite data center storage. Platform features such as VMware vSphere Storage DRS™ and storage policies assist in balancing workloads across storage resources, allowing providers to offer differentiated storage. This enables the provisioning of flexible pools of storage resources while maintaining consistent performance for consumers. Users can choose the appropriate storage tier for a particular type of workload.

- Historical data is usually the best guide for designing and determining the size and scale of a vSAN infrastructure. For example, if the average virtual machine requirements, in terms of storage, memory, IOPS, and compute can be correlated, this information can be used to efficiently design your vSAN infrastructure.
- Application data is also required in the design. Information, such as whether the application is cache-friendly, requires file services, and has availability requirements, is critical. After this information has been calculated, it must be correlated and aligned to future growth needs and the financial requirements of the provider with respect to operational and capital expenditures as well as profitability goals. The objective is to design the most efficient architecture that meets current and future demand.

5.1 vSAN Hardware Compatibility List

vSAN supports a very large number of different hardware components, such as SSDs, mechanical disks, and storage controllers. This list is constantly being added to, so refer directly to the online Hardware Compatibility List at https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan.

5.2 vSAN Ready Systems

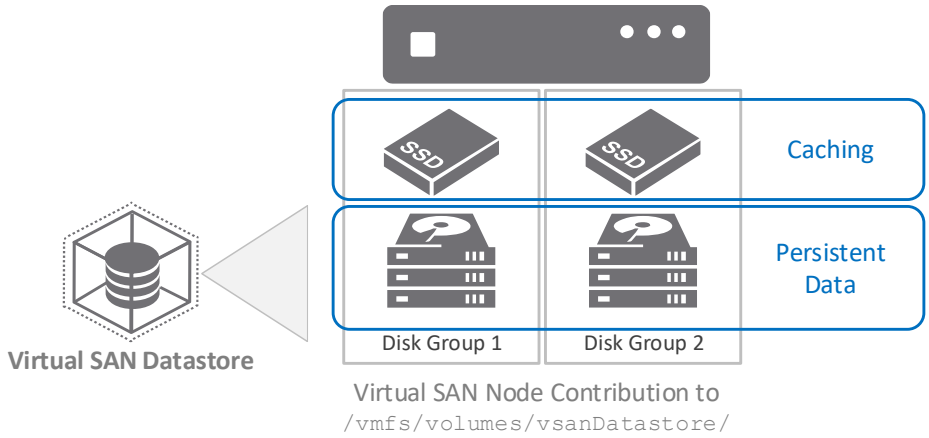
If the service provider plans to purchase new hardware, VMware has partnerships with Dell, Cisco and Supermicro, which have vSAN ready systems. These systems are out-of-the-box, preconfigured to fit the most common needs of customers and service providers for vSAN, and are compatible and tested with product. The list of vSAN ready systems can also be accessed directly online at http://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan.

5.3 Single Node Design

vSAN is capable of addressing up to 40 storage media devices per node. These devices can be organized in up to five disk groups, each containing at least one SSD device or flash card and at least one hard disk (hybrid configuration) or SSD device (all-flash configuration), up to a total of 7 devices.

The server configuration containing two SSDs and six 10k SAS drives uses the vSAN design of a single node as shown in the following figure. The storage media devices will be split equally into two disk groups, each containing an SSD and three SAS hard disk drives (HDDs).

Figure 23. Single Node Design with Two Disk Groups



5.4 vSAN Cluster Design

Generally, for a good design, VMware highly recommends that each node contained within a vSphere cluster is enabled for vSAN, and has the exact same configuration. When designing a vSAN cluster, use similarly configured and sized hosts. This allows for an evenly balanced vSAN cluster configuration. Otherwise, the performance and capacity expectations will be compromised.

Most VMware Cloud Providers have a preferred server vendor and, more than likely, they have a preconfigured and validated server (Ready Node) architecture for vSAN. If not, providers can also build their own vSAN nodes using the information provided by the compatibility guide. From an operational standpoint, the Ready Nodes have been the preferred choice of infrastructure for vSAN.

vCloud Director management clusters have different requirements than a resource cluster. Traditionally, providers would allocate dedicated storage for management-only clusters, which would ultimately increase the infrastructure costs. Deployment of a management cluster would normally entail the purchase of an additional storage array to provide the expected performance and availability for what is typically a high I/O throughput and highly available virtual environment. These costs, in turn, would inherently be passed on to the consumer. With vSAN, these costs are significantly reduced, making it an ideal solution for a dedicated management environment. Using vSAN for the management cluster provides an entire integrated stack that is automated by vCenter Server. The storage is locally attached and backed by SSD-based cache. Take the following best practices into consideration when designing a management cluster backed by vSAN:

- Use a minimum of four nodes with storage
- Use a balanced cluster with identical host configurations
- Do not use a stateless boot image. Use an SD card / USB / SATADOM (Serial ATA Disk on Module)

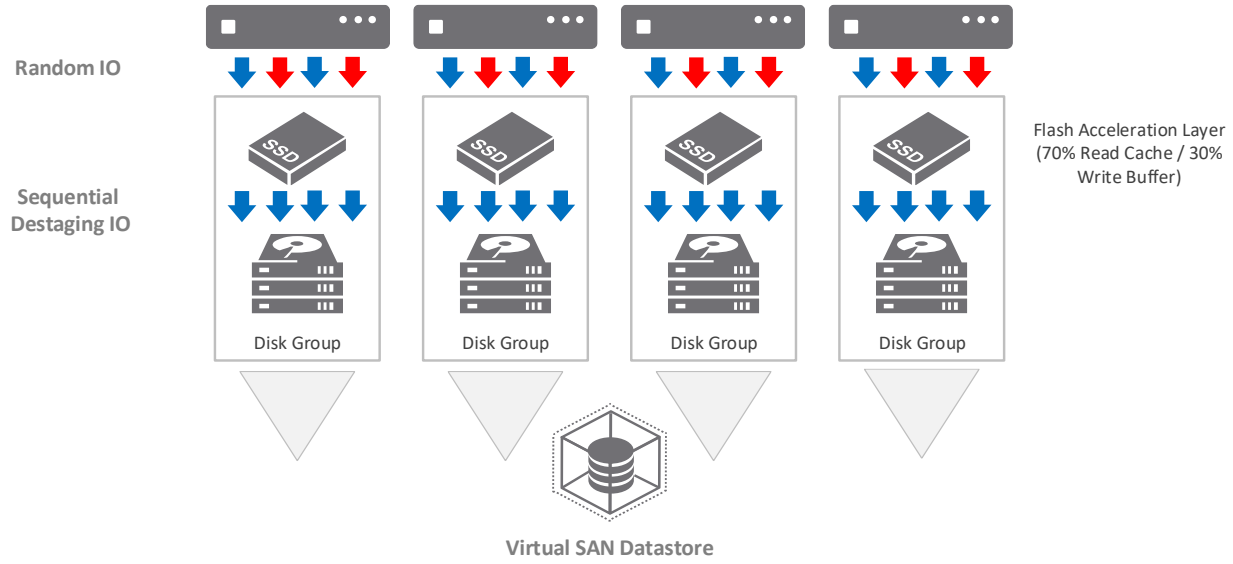
In addition, as a good practice, enable vSphere High Availability (HA) on the vSAN cluster to protect it from host and virtual machine failure.

A single vSAN cluster can contain up to 64 nodes. As a starting point, it is a commonly defined configuration for service providers to implement 20 to 24 node clusters, with each node configured with two disk groups to provide flexibility of growth. This approach also provides flexibility over Fault Domain design, for improved data availability and protection.

5.5 vSAN Design Principles

A key design principle of vSAN is to optimize to aggregate-consistent performance across a dynamic environment. One way that vSAN does this is through minimizing the I/O blender effect in virtualized environments.

Figure 24. I/O Blender Effect



The I/O blender effect is caused by multiple virtual machines simultaneously sending I/O to a storage subsystem, causing sequential I/O to become highly randomized, which can increase latency in the storage system. vSAN mitigates this through its tiered write design, using a flash acceleration layer which acts as a write buffer and read cache (hybrid) combined with capacity. The majority of reads and all writes are served by flash in a properly sized and designed vSAN solution, allowing for much improved performance in environments with highly random I/O.

Note When data must be destaged to spinning disk from the flash acceleration layer, the destaging operation consists primarily of sequential I/O, efficiently taking advantage of the full I/O capability of the underlying spinning disk.

A second design principle used in optimizing vSAN for aggregate consistent performance is not depending on data locality to guarantee performance. This concept is reviewed in-depth in the white paper *Understanding Data Locality in VMware Virtual SAN* (https://www.vmware.com/files/pdf/products/vsan/VMware-Virtual-SAN-Data-Locality.pdf).

5.6 vSAN Requirements

5.6.1 vSphere Requirements

vSAN requires vCenter Server. Both the Microsoft Windows version of vCenter Server and the VMware vCenter Server Appliance™ can manage vSAN enabled clusters. vSAN is configurable and monitored exclusively from only VMware vSphere Web Client.

vSAN requires a minimum of three vSphere hosts contributing local storage capacity to form a supported cluster (with the exception of ROBO clusters). The minimum, three-host, configuration enables the cluster to meet the lowest availability requirement of tolerating at least one host, disk, or network failure. However, the recommended minimum number of hosts for a production cluster is four nodes.

To enable vSAN 6.2 and to benefit from the features described in this document, the vSphere hosts require vSphere version 6.0 Update 2.

5.6.1.1 Host Memory Requirements

vSAN memory requirements are defined based on the number of disk groups and disks that the ESXi hypervisor manages. To support the maximum number of disk groups, 32 GB of RAM is required.

Note A best practice is to assign a consistent amount of RAM to each host in the cluster.

If there are separate host configurations with disparate amounts of RAM in the hosts, it is recommended to create multiple vSAN clusters. Other considerations when calculating memory requirements for hosts participating in a vSAN cluster include:

- vSAN scales back on its memory usage when hosts have less than 32 GB of memory.
- vSAN consumes additional memory when the number of nodes in the cluster is greater than 32.
- All flash vSAN configurations consume additional memory resources when compared to hybrid configurations.

For more information about sizing and designing a vSAN cluster, see the following documents:

- *VMware Virtual SAN 6.0 Design and Sizing Guide*

  http://www.vmware.com/files/pdf/products/vsan/VSAN_Design_and_Sizing_Guide.pdf
- *VMware Virtual SAN Diagnostics and Troubleshooting Reference Manual*

  http://www.vmware.com/files/pdf/products/vsan/VSAN-Troubleshooting-Reference-Manual.pdf

5.6.1.2 Host CPU Overhead

vSAN does not have specific CPU requirements because it introduces a less than 10 percent CPU overhead. This might impact available resources in high consolidation ratio workloads and CPU-intensive applications. Therefore, when designing and sizing the environment, take the projected resource requirements for the environment into account so that performance does not suffer due to a lack of resources or the overcommitment of available resources.

5.6.2 vSAN Cluster and Datastore Design

vSAN simplifies the storage configuration because there is only a single datastore for virtual machines. In addition, vSAN uses the concept of objects and components for storage of virtual machine data, where an object consists of multiple components that are distributed across a vSAN cluster, based on the assigned policy for the object.

vSAN 6.2 provides the following different object types:

- VM home namespace – Location for VM configuration and log files.

- VM swap object – Created for the VM swap file (only created when the VM is powered on).
- VMDK – Stores the data that is on a disk.
- Snapshot delta VMDKs – Created for use when a VM has a snapshot created on it.
- Memory object – Created when the snapshot memory option is selected when creating or suspending a virtual machine.

Each object can be a maximum of 255 GB in size. If they are larger, they are split into multiple components. Currently, vSAN 6.2 has a maximum of 9,000 components per host, which can be a limiting factor as to how far it can be scaled. More components are required for larger and more redundant virtual machines.

For instance, one virtual machine with a 500 GB disk (and no snapshots) would always consume the following components:

- Two for VM home namespace (failures to tolerate is always one)
- Two for VM swap objects (assuming that there is less than 255 GB of RAM in the machine)
- Two for the VMDKs (assuming no mirroring and no failures to tolerate)

Therefore, sizing the environment appropriately means that the limit on the number of components can be avoided during the design stage.

5.6.2.1 vSAN Disk Format

For vSAN with vSphere 6.0 and later, the disk format has been upgraded. The following configurations are now available.

Table 3. On-Disk File Format Version

vSAN Version	Format Type	On-Disk Version	Overhead	Supported Hosts
5.5 / 6.0	VMFS-L	v1	750 MB per disk	ESXi 5.5 U1+
6.0	VirstoFS	v2	1% of physical disk capacity	ESXi 6.0+
6.2	VirstoFS	v3	1% of physical disk capacity	ESXi 6.0 U2 +

The newer on-disk versions are not supported on older ESXi hosts. For example, with vSAN 5.5 v2, on-disk versioning is not supported.

The recommendation is to use the VirstoFS (v3) format unless there are backward compatibility concerns for the cluster. This configuration provides the broadest set of features available with vSAN.

5.6.2.2 Disk Group Design

Disk groups can be thought of as storage “containers” on vSAN hosts. They contain a maximum of one flash cache device and up to seven capacity devices. Either mechanical disks (hybrid configuration) or flash devices (all-flash configuration) are used as capacity devices. Each disk group assigns a cache device to provide the cache for a given capacity device. The recommendation is to have at least a 10 percent cache-to-capacity ratio. This provides a degree of control over performance as the cache-to-capacity ratio is based on disk group configuration. This also needs to be taken into account when planning future growth. For instance, you want to make sure that the flash layer devices are large enough

to scale the capacity layer for growth. Otherwise, you will not be able to maintain the minimum flash-to-capacity ratio. Depending on the use case, it might be necessary to design with additional cache up front to allow for future growth of the capacity layer.

To rebuild components after a failure, the design must be sized so that there is a free host's worth of capacity to tolerate each failure. There must be at least one full host's worth of capacity free for maintenance. The number of failures to tolerate determines whether there is a requirement for additional host capacity. For example, to rebuild components after one failure (FTT=1), there must be one full host's worth of capacity available. To rebuild components after a second failure (FTT=2), there must be two full hosts' worth of capacity free.

When evaluating hardware for cluster nodes in a hybrid cloud environment, the hardware must be identical, with special attention given to the storage I/O controllers. Queue depth must be as large as possible. At a minimum, the queue depth must be able to accommodate the throughput of current and future devices. In general, SATA drives have the lowest queue depth of the supported mechanical disks, and for this reason, they are not recommended in a cloud environment. Equally important, verify that the storage I/O controller supports pass-through mode. RAID 0 is not recommended in a hybrid cloud environment due to the increased maintenance of setting up and replacing disks.

Key design decisions must be made about number of disk groups and the flash-to-mechanical disk or flash-to-flash ratio in vSAN. Consider that vSAN:

- Supports up to one flash device for cache and a maximum of seven mechanical disks for capacity per disk group in a hybrid configuration.
- Supports up to one flash device for cache and a maximum of seven flash devices for capacity per disk group in an all-flash configuration.
- Supports up to a maximum of five disk groups per host.
- The number of mechanical disks matters in hybrid configurations due to the eventual destaging of read cache. Multiple disk spindles can speed up this process. Having more, smaller mechanical disks often provides better performance than fewer, larger disks in hybrid configurations.
- Allow 30 percent slack space when designing capacity.
- vSAN begins automatic rebalancing when a disk reaches the 80 percent of full threshold.
- Target configurations must be approximately 10 percent of the 80 percent threshold.
- Multiple disk groups typically provide better performance and smaller fault domains, but might sometimes come at a cost and consume additional disk slots.

The more disks that are configured per disk group, the more cache is needed, and the more capacity that is available for virtual machines. However, this leads to additional costs due to the disk group limits. Multiple disk groups require one flash device per group for cache and at least one device for capacity.

Disk group sizing is also important to consider when designing the volume. Include the following data points when deciding on the number of disk groups per host:

- Available space on the vSAN datastore.
- Number of failures you want to tolerate in the cluster.

The optimal number of disk groups is a balance between hardware and space requirements for the vSAN datastore. More disk groups increase space and provide higher availability. However, adding disk groups can be cost-prohibitive.

Thus, the total amount of space in the configuration can be quite significant, based on the number of disks configured in the hosts. Configure the disk groups based on:

- The end size of the datastore.
- The tolerance for failure required for the design (both failures-to-tolerate and fault domains).

A good starting point is to utilize two disk groups per host, containing three HDDs. This means that each host contains two SSDs and six HDDs. However, you can use more or fewer disks to meet the sizing requirements for vSAN (and the estimated sizing required for the overall design).

5.6.2.3 Datastore Sizing

Sizing the vSAN volume can be approached in several ways, depending on the factors that are most important to the environment. All different policy settings can impact the available space in the cluster. Also, the resultant sizing can mean additional hardware is required or that disk groups must be configured in a specific way.

To reduce complexities when estimating the size of the vSAN configuration, VMware released the VMware vSAN TCO and Sizing tool (https://vsantco.vmware.com/). To use the tool, simply enter the parameters. The calculator will generate the optimal configuration.

The following figure shows the results of a typical sizing recommendation from the tool.

Figure 25. vSAN TCO and Sizing Calculator

Note For manual sizing details outside of this tool, see the *VMware Virtual SAN 6.0 Design and Sizing Guide* (https://www.vmware.com/files/pdf/products/vsan/VSAN_Design_and_Sizing_Guide.pdf).

VMware recommends using the vSAN Sizing tool to determine the correct number of hosts and sizing for the cluster. For proper recoverability, use eight or more hosts in the cluster to support the configured fault domains. The end datastore size and number of hosts is dependent on the end workloads being hosted on vSAN, and it must be sized accordingly. Therefore, before selecting a drive size, consider disk groups, sizing, and expected future growth.

5.6.2.4 Hosts per Cluster

Determine the number of hosts for a cluster based on the following:

- Amount of available space on the vSAN datastore
- Number of available failures you wish to tolerate in the cluster

The number of hosts is generally a balance between hardware and space:

- More ESXi hosts and/or disk groups means higher hardware costs
- Fewer ESXi hosts and/or disk groups means resource availability could suffer

For instance, if the vSAN cluster has only three ESXi hosts, only a single failure is supported. If a higher level of availability is required, additional hosts are also required. For this reason, VMware recommends that the minimum production configuration include four hosts per cluster. However, a maximum of 64 is supported.

5.6.2.5 The Impact on Sizing of the Number of Failures-to-Tolerate Policy

The number of failures-to-tolerate policy setting is one of the core availability mechanisms with vSAN. This policy controls the number of replicas (mirrors) of a virtual machine component, although the policy can be applied to all of a virtual machine's disks, or individual VMDKs. This policy is important when planning and sizing storage capacity because it directly relates to the consumption of capacity that the virtual machine has on the storage.

For every *n* failures-to-tolerate, *n*+1 copies of the data are needed, and *2n*+1 hosts contributing storage are required. For instance, if the Number of Failures-to-Tolerate capability is set to 1, the virtual machine or disk will have two replica mirrored copies of its components created across the cluster. If the number is set to 2, three mirror copies are created, and so on. If a failure occurs, the mirrors take over.

As you can see, this leads to greater storage utilization than is actually configured because of the replicas created, but they also consume additional components. Configure this setting based on the availability requirements of the virtual machine or disk.

These requirements are defined in a storage policy and applied to the appropriate workloads. The default is to have one failure to tolerate unless the policy is changed to be a different value. The maximum number of failures to tolerate is three.

The recommendation is to keep the failures to tolerate setting at the default value of 1, unless there is specific need to configure a separate policy that provides a higher level of fault tolerance. If this is the case, use this policy sparingly with higher priority virtual machines.

5.6.3 Storage Device Requirements

5.6.3.1 Disk Controllers

The I/O controllers are just as important to a vSAN configuration as the selection of disk drives with performance being highly dependent on the choice of I/O controller. Each vSphere host that contributes storage to the vSAN cluster requires a disk controller. This can be a SAS or SATA host bus adapter (HBA) or a RAID controller. However, the RAID controller must function in one of two modes:

- Pass-through mode (preferred)
- RAID 0 mode

Pass-through mode, commonly referred to as JBOD or HBA mode, is the preferred configuration for vSAN because it enables vSAN to manage the RAID configuration settings for storage policy attributes based on availability and performance requirements that are defined on a virtual machine. These are the only two modes that are supported, with many storage adapters supporting both modes.

Consider the following when selecting a storage adapter:

- What modes does it support? (RAID 0, pass-through, or both.)
  - In RAID 0 mode, SSD performance must be reviewed.
  - RAID 0 mode also comes with operational overhead that can impact performance.
- Storage controller interface speed.
- Number of devices supported for the controller.
- Number of controllers to be used. Multiple controllers can reduce the failure domain and increase speed, but they also increase the cost.
- Controller queue depth is important for performance. Ideally, select a queue depth of 256 or higher. This is a significant determining factor for the performance of vSAN.

Regardless of the choice made, vSAN requires complete control of the drives. Performance between pass-through and RAID 0 modes is generally very similar for most interfaces.

When utilizing RAID 0 mode, disable the storage controller cache so that it does not conflict with the SSD drive caches that are controlled by vSAN. The storage controller cache is configurable on some, but not all, storage controllers.

When the storage controller cache cannot be completely disabled in a RAID 0 configuration, VMware recommends that you configure the storage controller cache for 100 percent read cache, effectively disabling the write cache.

The main consideration when utilizing RAID 0 mode for storage controllers within vSAN is the impact on the operational model. RAID 0 mode controllers typically require interaction with the storage controller software to manage the addition and removal of drives, because each drive is configured as a separate array, individually presenting each disk to vSAN, rather than simply providing a JBOD solution. vSAN performance and reliability can be impacted if this is not configured correctly.

The recommended configuration for VMware Cloud Providers is to use I/O controllers suited for the design characteristics of the environment, including:

- Model of the SSDs
- Model of the HDDs
- Queue depth of the controller
- Number of disks (and corresponding disk groups) being configured

For a list of the latest vSAN certified hardware and supported controllers, see the *VMware Compatibility Guide* at https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan.

5.6.3.2 Mechanical Disk Devices

When employing vSAN 6.2 hybrid disk groups, each vSphere host must have at least one SAS, near-line SAS (NL-SAS), or SATA mechanical device (HDD) to participate in the vSAN cluster. Mechanical disk devices, also referred to as capacity devices, account for the storage capacity of the vSAN shared datastore.

The HDDs in a vSAN hybrid environment are used primarily for data storage capacity, although they also are a determining factor in the available stripe width for virtual machine storage policies. If a specific stripe width is required, you must confirm that a particular stripe width is available across all hosts in the cluster to meet the requirement. Where a virtual machine has a high failure-to-tolerate setting, additional HDDs are necessary because each component must be replicated to meet the requirement.

VMware supports the following three types of mechanical disks:

- Serial Attached SCSI (SAS)
- Near Line Serial Attached SCSI (NL-SAS)

- Serial Advanced Technology Attachment (SATA)

NL-SAS can be thought of as enterprise SATA drives with a SAS interface. The best results can be obtained with SAS and NL-SAS drives. Only use SATA mechanical disks in capacity-centric environments where performance is not prioritized.

Choose the speed of the HDDs to meet the environmental characteristic for which the cluster is designed. VMware defines HDD characteristics and speed in the following table.

Table 4. vSAN HDD Environmental Characteristics

Characteristic	Revolutions per Minute
Capacity	7,200
Performance	10,000
Additional Performance	15,000

VMware recommends that you use an SAS HDD configuration suited to the characteristics of the environment being designed. If high performance is not required, a lower-cost disk will enable a higher number of failures to be tolerated. If there are no specific requirements, selecting 10,000 RPM drives achieves a balance between cost and availability.

Note VMware warns against mixing and matching HDD speeds to achieve a blend of different characteristics in the environment because there is only a single volume in the vSAN datastore. A best practice is to select one type of HDD per cluster. If a different characteristic is required, create a separate vSAN cluster for a higher performing configuration.

5.6.3.3 Flash-Based Devices

In vSAN 6.2 architecture flash-based devices can be used for both caching tier as well persistent capacity tier. In hybrid architectures, each vSphere host must have at least one flash-based caching—SAS, SATA, or PCI-e—to participate in the vSAN cluster. Flash-based devices provide both a write buffer and a read cache.

In a hybrid architecture, the larger the flash-based device capacity is per host, the larger the number of I/Os that can be cached and the greater the performance results that can be achieved. This scenario does not apply to the all-flash architecture.

The SSD used in a hybrid configuration is important because all reads and writes go to the SSD first. This is critical because it is what accounts for the speed that can be achieved with the vSAN solution.

In a hybrid system, the use of the SSD is split between a non-volatile write cache (30 percent) and a read buffer (70 percent). The endurance and the number of I/O operations per second that the SSD is capable of sustaining are important factors in the performance of the solution.

In all-flash architectures, each vSphere host must have at least one flash-based capacity—SAS, SATA, or PCI-e—marked as a capacity device and one for performance to participate in the vSAN cluster. The Virtual VSAN all-flash architecture is based on a two-tier model for performance and capacity. For an all-flash system, the cache is set to 100 percent writes because read performance is not a factor.

For endurance of the SSD used, standard industry write metrics are the primary measurements used to gauge the reliability of the drive. The measurement VMware uses has been updated to industry standard Terabytes Written (TBW) over the vendor's warranty. Previously the specification used was Drive Writes per Day (DWPD).

VMware recommends selecting an endurance class based on the requirements of the environment. The different endurance classes are defined in the VMware Compatibility Guide for vSAN

(https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan) as shown in the following table.

Table 5. SSD Endurance Classes

Endurance Class	Terabytes Written in 5 Years (TBW)
Class A	>=365 TBW
Class B	>=1,825 TBW
Class C	>=3,650 TBW
Class D	>=7,300 TBW+

In addition to the generic classes, the following table lists the VMware recommendation for the endurance class based on the SSD tier as defined in the *VMware Virtual SAN 6.0 Design and Sizing Guide* (https://www.vmware.com/files/pdf/products/vsan/VSAN_Design_and_Sizing_Guide.pdf).

Table 6. SSD Endurance Classes by Tier Classes (Caching Drives)

Endurance Class	SSD Tier	TB Writes Per Day	Terabyte Writes in 5 Years
Class A	All-Flash – Capacity	0.2	365
Class B	Hybrid - Caching	1	1,825
Class C	All-Flash – Caching (Medium Workload)	2	3,650
Class D	All-Flash – Caching (High Workload)	4	7,300

For optimal performance of vSAN, select a higher performance class of SSD. VMware defines classes of performance in the VMware Compatibility Guide for vSAN (https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan) as shown in the following table.

Table 7. SSD Performance Classes (Capacity Drives)

Performance Class	Writes Per Second
Class A	2,500 – 5,000
Class B	5,000 – 10,000
Class C	10,000 – 20,000
Class D	20,000 – 30,000
Class E	30,000 – 100,000
Class F	100,000+

When designing a vSAN platform, it must be understood that a direct correlation exists between the SSD performance class and the level of vSAN performance. In general, the highest-performing hardware supports optimal performance of the solution. Cost, therefore, is the determining factor and might make a lower class of hardware more attractive, even though the performance or size might not be ideal.

The recommended strategy is to select an SSD size that is, at a minimum, 10 percent of the anticipated size of the consumed capacity storage, before the Number of Failures to Tolerate capability is considered. For instance, the SSD should be at least 100 GB if usage is estimated to be 1 TB of capacity storage, consumed in a 2-TB disk group.

Note In hybrid and all-flash architectures, flash-based caching devices do not contribute to the overall size of the distributed vSAN shared datastore. Because they are utilized for read and write caching, they count only toward the capacity of the vSAN caching tier or write buffer. In all-flash architecture, flash-based devices marked as capacity devices make up the size of the distributed vSAN datastore.

5.6.4 Network Requirements

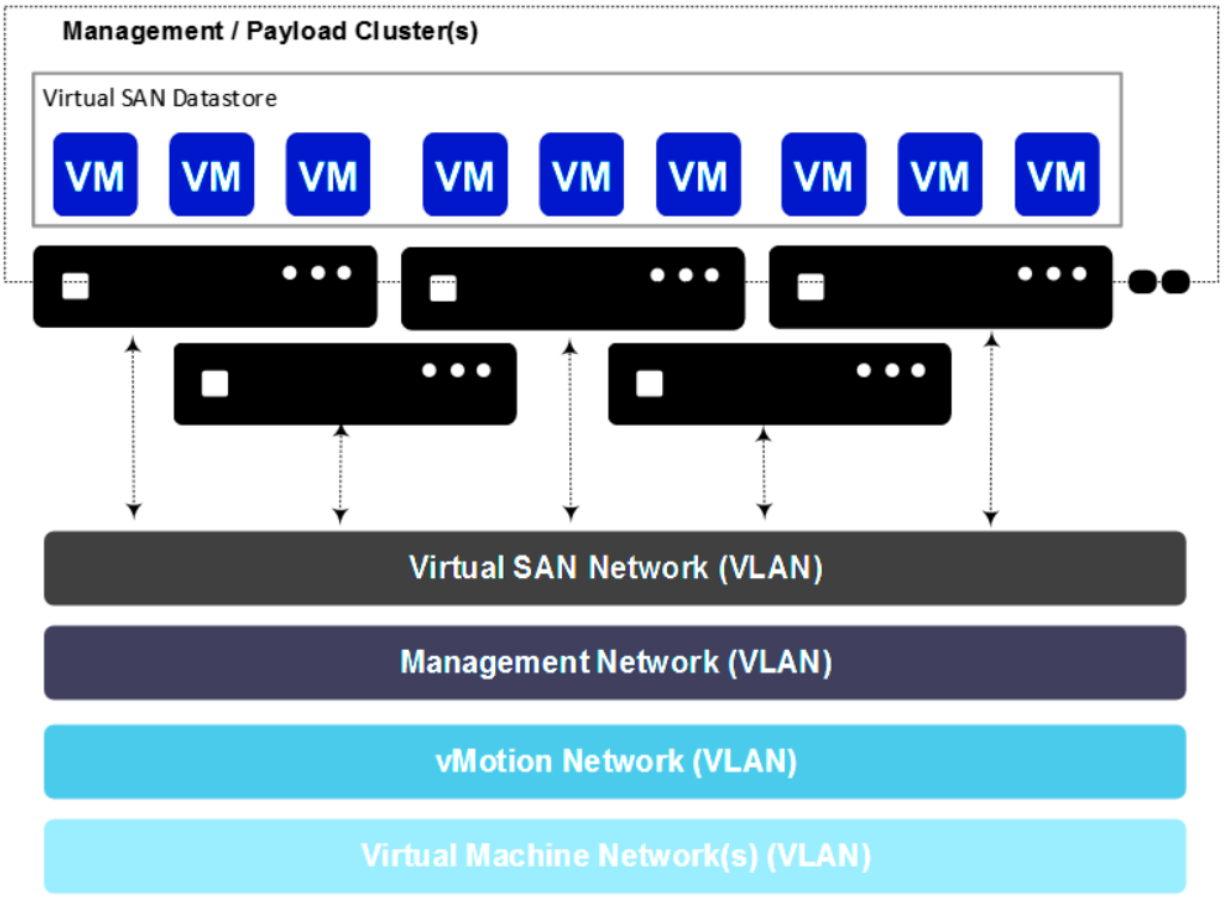
5.6.4.1 Network Design

vSAN uses the network to transport all information, including communication between the cluster nodes and virtual machine I/O operations. The required network transport is accomplished by a specially created VMkernel port group that must be configured on all hosts in the cluster, whether the hosts are contributing storage resources to the cluster or not.

When designing the network, the architect must consider how much replication and communication traffic is running between hosts, because with vSAN, the amount of traffic directly correlates to the number of virtual machines that are running in the cluster, how write-intensive the I/O is for the applications running, and whether an all-flash configuration is being used.

As with other communications, such as vSphere vMotion, isolate vSAN traffic on its own Layer 2 network segment. You can do this with dedicated switches or ports, or by using VLANs, as shown in the following figure.

Figure 26. Conceptual Network Diagram



5.6.4.2 Network Interface Cards (NIC)

In vSAN hybrid architectures, each vSphere host must have at least one 1-Gb Ethernet or 10-Gb Ethernet capable network adapter, although VMware recommends 10-GbE for production environments. The all-flash architectures are only supported with 10-Gb Ethernet capable network adapters.

For redundancy and high availability, a team of network adapters can be configured on a per-host basis. The teaming of network adapters for link aggregation to improve performance is not supported. However, the teaming of network adapters to provide increased availability is addressed in further detail later in this section. VMware considers this to be a best practice, but it is not necessary in building a fully functional vSAN cluster.

5.6.4.3 Virtual Switches

vSAN is supported on both the VMware vSphere Distributed Switch™ (VDS) and the vSphere standard switch (VSS). No other third-party virtual switch types are supported with vSAN. For most designs, the vSphere Distributed Switch offers advantages through the inclusion of the VMware vSphere Network I/O Control and QoS features.

The benefit of using vSphere Distributed Switch configurations is that they allow Network I/O Control to be used, which allows for prioritization of bandwidth when there is contention in an environment. In addition, the vSphere Distributed Switch using Network I/O Control is an attractive option for

environments that employ a limited number of 10-GbE network ports, because it allows the interfaces to be shared, while it prioritizes bandwidth in contention scenarios.

For these reasons, VMware recommends using a vSphere Distributed Switch for the vSAN port group, so that priority can be assigned using Network I/O Control to separate and reserve the bandwidth for vSAN traffic in the environment.

5.6.4.4 VMkernel Network Port Group

On each vSphere host, a VMkernel port for vSAN communication must be created and used for synchronization and replication activities. The VMkernel virtual adapter type has been added to vSphere for vSAN. This VMkernel port is labeled vSAN traffic. This port group is typically dedicated and isolated to vSAN communication. However, if a 10-GbE network interface is being employed, it can be shared.

For 1-GbE networks, VMware recommends using a dedicated port group, dedicated network card, and an isolated network for vSAN traffic. This also increases security and prevents other traffic from impacting the performance of vSAN.

As highlighted previously, the VMkernel interface is used for host intra-cluster communications as well as for read and write operations whenever a vSphere host in the cluster is the owner of a particular virtual machine, but the actual data blocks making up that virtual machine's objects are located on a remote host in the cluster. In this case, I/O must traverse the network configured between the hosts in the cluster. If this interface is created on a VDS, the Network I/O Control feature can be used to set shares or reservations for the vSAN traffic.

5.6.4.5 Network Speed Requirements

vSAN supports the following configurations:

- Hybrid vSAN configurations support 1-Gb or 10-Gb Ethernet for network uplinks. Depending on usage, the amount of activity on the vSAN might overwhelm a 1-Gb network and might be the limiting factor in I/O-intensive environments, such as the following:
    - Rebuild and synchronization operations
    - Highly-intensive disk operations, such as cloning a VM
    - High-density environments with a large number of VMs
- All-flash vSAN configurations are supported only when using 10-Gb Ethernet network uplinks. The improved performance with an all-flash configuration consumes significantly more network bandwidth due to the increased speed of the disks in the configuration.

A 10-GbE network is required to achieve the highest performance (IOPS). Without it, a significant decrease in vSAN performance can be expected. For this reason, VMware recommends a 10-Gb Ethernet connection for use with vSAN in all production configurations.

5.6.4.6 Jumbo Frames

vSAN supports using jumbo frames for vSAN network transmissions. The environment is supported fully whether or not jumbo frames are employed. The performance gains are often not significant enough to justify the underlying configuration necessary to enable jumbo frames properly on the network.

For this reason, VMware only recommends using jumbo frames for vSAN traffic if the physical environment is already configured to support them, and they are part of the existing design, or if the underlying configuration does not create a significant amount of added complexity to the environment.

5.6.4.7 VLANs

VMware recommends segregating vSAN traffic on its own VLAN. When multiple vSAN clusters are used, each cluster must use a dedicated VLAN or segment for their traffic. This prevents interference between

clusters and aid in troubleshooting cluster configuration. VMware strongly recommends that a production design always employ separate VLANs for vSAN traffic.

5.6.4.8 Multicast Requirements

vSAN requires that IP multicast be enabled on the physical switches and routers that handle vSAN traffic along the Layer 2 path, and optionally the Layer 3 path. It is utilized for intra-cluster communication, specifically for heartbeats and exchange of metadata between the ESXi hosts participating in the cluster. Multicast traffic can be limited to specific port groups by using IGMP (v3) snooping. This is considered a best practice. VMware recommends not implementing multicast flooding across all ports, rather enabling multicast only on the segments or ports that are being used for vSAN.

By default, multicast communication is not configured on most hardware switches. As a result, there is no IGMP snooper setup, and therefore traffic is not passed. To be sure that multicast traffic is passed, make sure to configure the IGMP snooper so traffic can be passed, or explicitly disable IGMP snooping on the port.

5.6.4.9 Networking Failover, Load Balancing and Teaming Considerations

Business continuity and disaster recovery (BCDR) is critical in any environment in case of a network failure. vSAN supports teaming configurations for network cards to enhance the availability and redundancy of the network.

Note vSAN does not currently leverage teaming of network adapters for the purpose of bandwidth aggregation and load balancing to enhance performance.

For a predictable level of performance with the use of multiple adapters, configure each adapter with the following teaming algorithms:

- Route based on originating virtual port – Active/Passive failover configuration of adapters in the team.
- Route based on IP hash – Active/Active failover configuration with:
  - Static EtherChannel for standard switches.
  - Link Aggregation Control Protocol (LACP) port channel for vSphere distributed switches.
- Route based on physical network adapter load – Active/Active failover configuration with LACP port channel for distributed switches. Unavailable for standard switches.

Typically, VMware recommends that configurations use an Active/Active redundancy with a route based on physical adapter load for the teaming in the environment. In this configuration, idle network cards do not wait for a failure to occur. This assumes that LACP has been, or can be, configured in the environment.

5.6.4.10 Network QoS and Network I/O Control

Quality of Service (QoS) can be implemented using Network I/O Control for a vSAN environment. This allows a dedicated amount of the network bandwidth to be allocated to the vSAN traffic. When using Network I/O Control, no other traffic will impact the vSAN network, and likewise, through the use of shares, other traffic types can also be protected. This is useful in a shared environment, but does require that a vSphere distributed virtual switch be used.

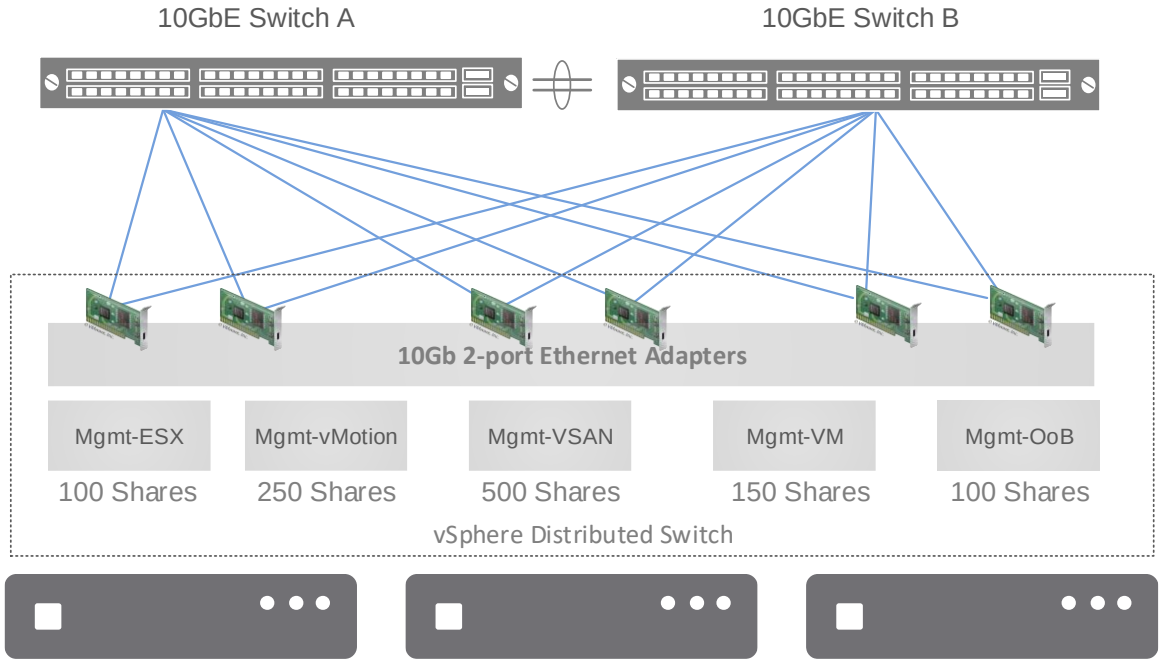
In Network I/O Control, configure reservation and shares for the vSAN outgoing traffic as follows:

- Set a reservation so that Network I/O Control guarantees that minimum bandwidth is available on the physical adapter for vSAN.
- Set shares so that when the physical adapter assigned for vSAN becomes saturated, certain bandwidth is available to vSAN to prevent vSAN from consuming the entire capacity of the physical adapter during rebuild and synchronization operations. For example, the physical adapter might

become saturated when another physical adapter in the team fails, and all traffic in the port group is transferred to the other adapters in the team.

In addition to this, priority tagging can be set. Priority tagging is a mechanism to indicate to the connected external network devices that vSAN traffic has higher QoS demands. You can assign vSAN traffic to a certain class and accordingly mark the traffic with a Class of Service (CoS) value from 0 to 7 by using the traffic filtering and marking policy of vSphere Distributed Switch. The lower the CoS value is, the higher the priority of the vSAN data.

Figure 27. Network I/O Control



For most designs, make use of Network I/O Control and QoS when the NIC is shared between traffic types, such as if a 10-GbE network card team is used for all traffic on a host system.

5.6.5 vSAN Policy Design

After vSAN is enabled and configured, storage policies that define the virtual machine storage characteristics must be created. These characteristics allow configuration of different levels of service to be provided to a virtual machine. If no specific policy is applied, a default policy tolerating a single failure and having a single disk stripe is applied. Creating multiple policies with vSAN allows for policies to be applied on a per virtual machine basis and changed as needed.

As discussed earlier, vSAN is an object storage technology. Each virtual machine deployed on vSAN comprises a set of objects. VMDKs, snapshots, VM swap space, and the VM home namespace are all objects. Each of these objects is comprised of a set of components that are determined by capabilities configured in the VM storage policy. For instance, if a virtual machine is deployed with a policy to tolerate one failure, objects will be made up of two replica components. If the policy contains a stripe width, the associated object will be striped across multiple devices in the capacity layer. Each stripe is a component of the object.

As you plan the design with respect to component maximums, be aware that vSAN might decide that an object needs to be striped across multiple disks, even with the default policy of one stripe in place. Normally, this is the result of an administrator requesting that a VMDK be created that is too large to fit on a single physical drive. As previously highlighted, the maximum component size is 255 GB. To elaborate, objects that are greater than 255 GB in size are automatically divided into multiple components. For

example, if an administrator deploys a 2-TB VMDK, it will result in eight or more components being created in the same RAID-0 stripe configuration.

vSAN guarantees the storage policies that have been configured. If a policy cannot be met, virtual machine provisioning fails, unless force provisioning has been set. Almost any combination of the possible attribute capabilities can be set in a policy. Some additional benefits of vSAN policy application include the ability to configure policies at any time and to switch policies “on-the-fly” for virtual machines.

Finally, all virtual machines created on vSAN will be thin-provisioned, so a design must deal with capacity management accordingly to avoid overcommitting resources. An understanding of the policies that can be set allows the administrator to avoid configuring policies that do not make sense.

5.6.5.1 vSAN Policy Capabilities

vSAN allows you to set several different policy attributes in a storage policy. These attributes can be used alone or combined to provide different service levels.

Before making design decisions, understand the policies and the objects to which they can be applied. The policy options are listed in the following table.

Table 8. vSAN Policy Options

Capability	Use Case	Value	Comments
Number of disk stripes per object	Performance	Default 1 Maximum 12	This is a standard RAID 0 stripe configuration used to increase performance for a virtual machine disk. This setting defines the number of HDDs on which each replica of a storage object is striped. If the value is higher than 1, increased performance can result. However, an increase in system resource usage might also result.
Flash read cache reservation (%)	Performance	Default 0 Maximum 100%	Flash capacity reserved as read cache for the storage is a percentage of the logical object size that will be reserved for that object. Use this setting only for workloads that must have read performance issues addressed. The downside is that other objects cannot use a reserved cache. VMware recommends not using these reservations unless it is necessary because unreserved flash is shared fairly among all objects.

Capability	Use Case	Value	Comments
Number of failures-to-tolerate	Redundancy	Default 1 Maximum 3 Note: Maximum value is 1 if disk size is greater than 16 TB	Defines the number of host, disk, or network failures a storage object can tolerate. The higher the value, the more failures can be tolerated. When the fault tolerance method is mirroring, for n failures tolerated, *n*+1 copies of the disk are created, and 2*n*+1 hosts or fault domains contributing storage are required. The higher *n* value indicates that more replicas of virtual machines are made, which can consume more disk space than expected. When the fault tolerance method is erasure coding, to tolerate 1 failure, 4 hosts (or fault domains contributing storage) are required, and to tolerate 2 failures, 6 hosts or fault domains are required.
Failure tolerance method (Only available with all-flash vSAN 6.2)	Performance or Capacity	Default: RAID 1 (Mirroring)	Defines the method used to tolerate failures. RAID 1 achieves failure tolerance using mirrors, which provides better performance. RAID 5/6 achieves failure tolerance using parity blocks, which provides better space efficiency. RAID 5/6 is only available on all-flash vSAN clusters, and when the number of failures-to-tolerate (FTT) is set to 1 or 2. A value of 1 FTT implies a RAID 5 configuration, and a value of 2 FTT implies a RAID 6 configuration.
IOPS limit for Object (Only available with vSAN 6.2)	Performance	Default: 0	Defines IOPS limit for a disk. IOPS is calculated as the number of I/Os using a weighted size. By default, the system uses a base size of 32 KB. Thus, by default, a 64 KB I/O represents 2 I/O. **Note** When calculating IOPS, read and write are regarded as equivalent and cache hit ratio or sequence is not taken into account. If IOPS of a disk exceeds the limit, I/O will be throttled. If the limit is set to 0, it means no limit is being applied.

Capability	Use Case	Value	Comments
Disable Object Checksum (Only available with vSAN 6.2)	Override Policy	Default: No	vSAN uses end-to-end checksum to secure the integrity of data by confirming that each copy of a file is exactly the same as the source file. The system checks the validity of the data during read/write operations, and if an error is detected, vSAN repairs the data or reports the error. If a checksum mismatch is detected, vSAN automatically repairs the data by overwriting the incorrect data with the correct data. Checksum calculation and error-correction are performed as background operations. This setting determines whether checksums are calculated for the data being written to the volume or not.
Force provisioning	Override policy	Default: No	Force provisioning allows for provisioning to occur even if the policy configured cannot be satisfied by the currently available cluster resources. This is useful if there is a planned expansion of the vSAN cluster, and provisioning of VMs must continue, because vSAN automatically tries to bring the object into compliance as resources become available.
Object space reservation (%)	Thick provisioning	Default 0 Maximum 100%	The percentage of the storage objects that are thick provisioned upon VM creation. The remainder of the storage objects are thin provisioned. This is useful if a predictable amount of storage will always be filled by objects, cutting back on repeatable disk growth operations for all but new or non-predictable storage use.

5.6.5.2 vSAN Default Policy

By default, policies are configured based on the application requirements. However, they are applied differently depending on the object. The following table lists the default policy options applied to the different objects in vSAN.

Table 9. Object Policy Defaults

Object	Policy	Comments
Virtual Machine Namespace	1 Failures-to-Tolerate	Configurable. However, changes are not recommended.

Object	Policy	Comments
Swap	1 Failures-to-Tolerate	Configurable. However, changes are not recommended.
Virtual Disks	User-Configured Storage Policy	Can be any storage policy configured on the system.
Virtual Disk Snapshots	Uses Virtual Disk Policy	Same as Virtual Disk policy by default. Changing is not recommended.

If there is no user-configured policy, the default system policy of one failure-to-tolerate and one disk stripe is used for virtual disks and virtual disk snapshots.

Policy defaults for the VM namespace and swap are set statically and are not configurable, so that there is appropriate protection for these critical virtual machine components. Policies must be configured based on the application's requirements. This feature gives vSAN its power, because it can adjust how a disk performs on demand based on the configured policies.

5.6.5.3 Application Requirements

Policy design starts with an assessment of business needs and application requirements, because policies allow any configuration to become as customized as needed. Using general policies is advisable if no specific use cases exist, or if a proof of concept (PoC) is under development.

Start by assessing the following different application requirements:

- I/O performance and profile of your workloads on a per-virtual-disk basis
- Working sets of your workloads
- Hot-add of additional cache requires repopulation of cache
- Specific application best practice (such as block size)

Aim to design policies for availability and performance in a conservative manner, so that space consumed and recoverability properties are balanced. In many cases, the default system policy is adequate, and no additional policies are required, unless there are specific requirements for performance or availability.

5.7 vSAN and Recoverability

vSAN enables providers to protect consumer data, even in the event of disk, host, network, or rack failures, with built-in distributed RAID and cache mirroring. Resiliency is enhanced by defining availability on a per-VM basis. The number of host, network, disk, or rack failures to tolerate in a vSAN cluster can be customized based on customer, application, or business requirements.

vSAN supports clustering technologies from both Oracle and Microsoft. With Oracle Real Application Cluster (RAC), customers can run multiple Oracle RDBM instances, accessing the vSAN datastore to deliver better performance, scalability, and resilience. Additionally, Windows Server Failover Clustering (WSFC) can provide protection against application and service failures.

Technically, there is a minimum requirement of three ESXi hosts in a vSAN cluster. Although VMware fully supports three-node configurations, there are limitations when compared with configurations of four or more nodes. In particular, in the event of a failure, there is no way for vSAN to rebuild the components on another host in the cluster to tolerate another failure. There is also an operational limitation because vSAN will not have the ability to migrate all data from a node during maintenance.

Business continuity and disaster recovery are paramount for ensuring that consumers' business critical environment, data, and online presence are available with minimal downtime. vSAN recoverability can be addressed through a range of solutions, from near-line capabilities such as snapshots, offline capabilities provided by vSphere storage integrated backup solutions (using APIs for data protection), or inter-data center solutions, such as stretched cluster.

5.8 Understanding How Failures Impact vSAN

vSAN was built with a native ability to handle failures. It can handle transient failures, such as a network misconfiguration, or a full device failure, such as a hard drive failing. vSAN classifies failures in two distinct ways:

- Absent failures
- Degraded failures

The way that vSAN handles a failure depends on its type. The following sections describe these two types of failures.

5.8.1 Absent Failures

Absent failure events are recognized in vSAN whenever I/O failures are detected with any of the following hardware devices:

- Physical network switches
- Network interfaces (NICs)
- ESXi hosts or fault domains

When an absent failure occurs, all data is not immediately resynchronized to another host in the environment. The resynchronization operation is in fact not initiated for 60 minutes from the detection and acknowledgement of the failure event.

The reason for this is that these types of failures can often be temporary. For instance, the network could be down, or a host might be rebooting for regular maintenance. Therefore, the system waits for these components to come back online before committing to a resynchronization of all the data on the node.

While the default delay is 60 minutes, this is an adjustable system parameter that can be increased or decreased based on business requirements. For example, you might want to adjust the parameter if a host takes longer than the set delay to reboot, or if the host will not be available during a maintenance window that is a longer period of time than the set delay.

See *Changing the default repair delay time for a host failure in VMware Virtual SAN* (http://kb.vmware.com/kb/2075456) for instructions on modifying the settings without having to restart the hosts.

5.8.2 Degraded Failures

Degraded failure events are recognized in vSAN whenever I/O failures are detected with any of the following hardware devices:

- Mechanical disks
- Flash-based devices
- Storage controllers

Unlike absent failure events, degraded failure events immediately activate the resynchronization of the data among the other hosts in the cluster. This function is not configurable because when the previously listed devices fail, the failure is not likely to be temporary. Therefore, there is no benefit to waiting for the devices to come back online.

When this type of event is detected, the resynchronization operation is performed for objects that are no longer in compliance with their policies due to the failure. It is an expensive operation because it creates new replicas of the data that existed on the failed device or component, using the remaining replica for the component as the source. This data exists elsewhere in the cluster, on other hosts or on other mechanical disks of the same host, based on the configured failures-to-tolerate (FTT). The replicas of individual objects are not all created in the same place. Rather, the replicas are distributed around the rest of the cluster wherever there is spare capacity. Thus, the entire cluster is used as a *hot spare.*

Regardless of when it is activated, the resynchronization operation contends with the I/O of virtual machines and the resource available in the cluster. The operation could have a detrimental impact on the overall capabilities of the vSAN cluster if not sized and designed correctly. From a performance perspective, the resynchronization operation essentially limits the IOPS available to virtual machines in the cluster because of the operations being performed to recover the affected objects and components.

From a data availability perspective, whenever the resynchronization operation is unable to be completed, due to an inadequate amount of collective spare capacity in the cluster, data accessibility could be at risk. For instance, consider a scenario with a three-node cluster configured with the default availability policy setting of FTT=1. When a host failure occurs, the remaining two hosts are excluded from the resynchronization operation because they are already hosting objects and components for the affected virtual machines. In this scenario, the resynchronization operation waits until the failed node is brought back online, or a new one is added to the cluster, before resuming the resynchronization operation, and restoring the compliance value for data availability.

Table 10. vSAN Resilience

Component	VM	User Experience	Restore process	VSAN Recovery process
HDD	Keep running	No impact	• HDD replacement • Add to vSAN cluster	Copy data on failure HDD to another (Instantly)
SSD	Keep running	No impact	• SSD replacement • Add to vSAN cluster	Copy data on all HDD in disk group to other HDDs (Instantly)
RAID Adapter	Keep running	No impact	• RAID adapter replacement • Reconfigure RAID • Add to vSAN cluster	Copy data on all HDDs under failure RAID Adapter to other HDDs (Instantly)
ESXi Host	Reboot VMs on failure host	A few minutes downtime (Reboot automatically)	• Replace host • Initial setup • Add to vSAN cluster	Copy data on all HDDs in the failure host (60 minutes later)
Network	Reboot VMs	A few minutes downtime (Reboot automatically)	• Restore network	Copy data on all HDDs in isolated hosts to other HDDs (60 minutes later)

5.9 Summary of Key vSAN Design Factors

There are a number of considerations that are needed, prior to configuring a vSphere cluster to participate as a vSAN. The following list shows the minimum key requirements to implement a vSAN enabled cluster.

Table 11. Key Virtual SAB Design Factors

Design Factor	Detail
Controller Queue Depth	Controller queue depth impacts the rebuild/resync times. A low controller queue depth may impact the availability of production virtual machines during rebuild/resync. A minimum queue depth of 256 is required for vSAN.
Number of disk groups	• The number of disk groups impacts fault isolation as well as rebuild/resync times. • Fault isolation: Configuring more than 1 disk group allows better tolerance against SSD failures, because data is spread across more disk groups. • Rebuild/resync times: Configuring more than 1 disk group allows faster rebuilds/resyncs.
Number of hard drives (HDDs) in a disk group	The number of HDDs in a disk group has an impact on the performance of vSAN. While a single HDD is the minimum requirement for a disk group, for better performance when there are more virtual machines, and better handling of rebuild/resync activities, VMware recommends configuring more than 1 HDD per SSD, per the guidance above.
Class of SSDs	The class of SSD you choose has a direct impact on the performance of your overall system.
Balanced compared with unbalanced cluster	An unbalanced cluster can impact vSAN performance, as well as the rebuild/resync times. A balanced cluster delivers more predictable performance, even during hardware failures. In addition, performance impact during resync/rebuild is minimal when the cluster is balanced.
1-GB vs 10-G Ethernet	The choice of 1-GbE vs 10-GbE Ethernet has an impact on the vSAN performance. For larger, higher performing and production workloads, 10 GbE is recommended.

vSAN Performance Testing

Testing the performance capabilities of the vSAN cluster helps to confirm that the configuration is sound. Many of the traditional methodologies used for testing storage subsystems in a vSphere environment are supported with vSAN. The tests differ when analyzing vSAN performance. Tools such as vSAN Observer can be used to provide a more in-depth level of detail.

For optimal performance, VMware recommends that the active working set of the virtual machine fit into the flash acceleration layer of vSAN. Measuring the active working set of a virtual machine is not straightforward, but a good conservative estimate is 10 percent of the used capacity, not taking into account capacity utilized for the Number of Failures-to-Tolerate redundancy policy assigned to the virtual machine.

6.1.1 Tools Used for Performance Testing

VMware recommends using any tool that allows for testing with multiple outstanding I/Os for use when testing vSAN. The Health Check plug-in interface now contains performance testing tools. This is by far the easiest way to test the vSAN environment.

Another tool that can be used is Iometer, because it gives the ability to test using a number of different parameters.

Note Tools such as dd and SIO are not recommended for performance testing of vSAN because they can be configured only to run tests that sequentially issue outstanding I/O (OIO) tests.

In addition, VMware has released a tool called the VMware I/O Analyzer that utilizes Iometer for testing. I/O Analyzer is supplied as an easy-to-deploy virtual appliance that automates storage performance testing and analysis through a unified web interface. It can be used to configure and perform storage tests and view graphical results for the tests. I/O Analyzer can be used to run Iometer-based workloads or use application trace replays.

6.1.1.1 Storage Policy Configuration for Performance Testing

It is also important to recognize that vSAN supports per-object based policies, which impact performance and availability. The policies have previously been discussed in this document and are equally applicable to performance testing. These include:

- VMware vSphere Flash Read Cache™ Reservation – VMware does not recommend using this policy during performance testing of server workloads. Any reservation will reserve a portion of the 70 percent allocation of Flash Read Cache to an object, whether it is needed or not. VMware recommends letting the vSAN algorithms handle read cache reservation automatically. This policy is used by default only for VMware Horizon® 6 when using linked clones with replica objects.
- Stripe Width – Use this policy to test the performance of a single virtual machine/VMDK. This allows the VMDK to be split into multiple components, and allows those components to be spread across the cluster of hosts. VMware recommends setting the stripe width equal to the number of nodes in the cluster, up to the maximum of 12. If testing and performing scale-out performance tests with multiple virtual machines or multiple VMDKs, this policy is not recommended.
- Object Space Reservation – This policy is recommended to encourage even distribution of components through a vSAN cluster, by forcing the vSAN algorithms to take into account the full size of the object when making placement decisions. Using this policy is similar to a lazy zero thick disk format, and has no impact on performance outside of the influence on component placement and distribution.
- Number of Failures-to-Tolerate – VMware recommends keeping and testing with a failure to tolerate setting that adheres to the availability needs of the environment. The default is 1. However, testing is recommended using both policies for comparison.

Eight Common Service Provider Use Cases

Cloud service providers can utilize either traditional infrastructure (servers with legacy storage), a hyper-converged platform, such as vSAN, or a combination of both. A hyper-converged platform is a platform that combines compute, storage, networking, and virtualization resources and management, enabling scale-out consumption. The tight integration of vSAN with vSphere enables VMware Cloud Providers to host public or private cloud offerings and easily utilize vSAN as the underlying storage infrastructure.

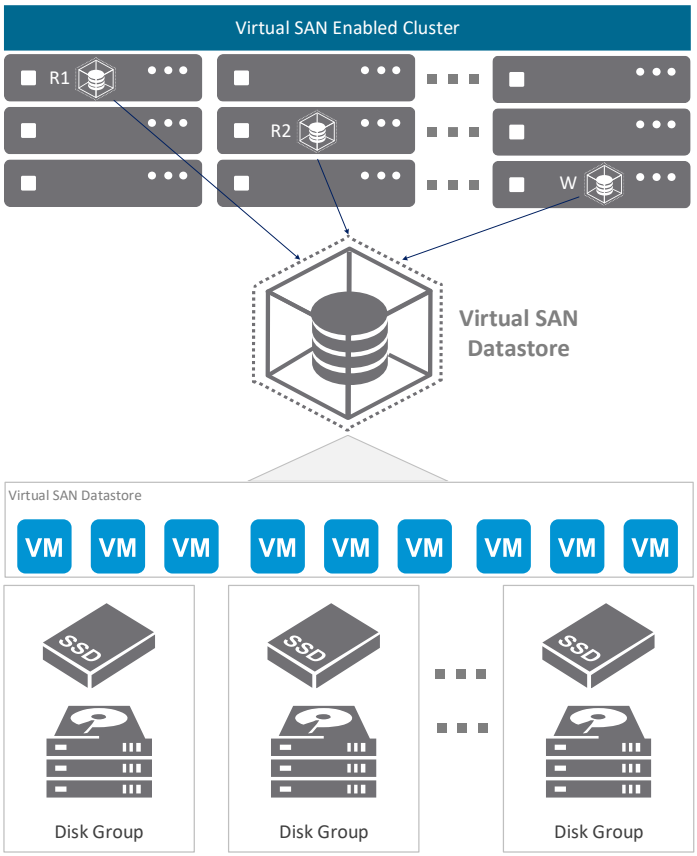
The customers of service providers that leverage vSAN predominately run mission-critical workloads in which stability, reliability, and predictability are extremely important. Many traditional infrastructures, over time, hinder VMware Cloud Providers' ability to remain competitive in such a price-sensitive market. vSAN enables VMware Cloud Providers launching new offerings to avoid the large capital expenditures typically associated with legacy storage arrays.

VMware Cloud Providers who offer consumers hybrid cloud-based virtualized solutions, such as vSphere as a Service, Desktop-as-a-Service, Tier 1 applications, and Database-as-a-Service, are ideal candidates for vSAN. Additionally, VMware Cloud Providers can leverage vSAN as the storage infrastructure for their vCloud Director cloud offering.

7.1 Local Data Center Site Deployment Model

Deploying vSAN within the boundaries of a single data center site is the most common design to date. In this use case, all nodes within the vSAN enabled clusters are installed at a single site, locally.

Figure 28. Local Single Data Center Deployment of vSAN



This single site model facilitates a number of different use cases for service providers, including, but not limited to:

- Tier 1, 2 and 3 production workload systems
- Management or cloud management clusters
- DMZ/isolated clusters
- Development platform clusters
- Backup/disaster recovery target
- Virtual desktop and EUC infrastructure

The following sections address the design considerations and recommendations for these use cases.

7.1.1 Tier 1 / Tier 2 / Tier 3 Workloads

A vSAN infrastructure is a highly effective solution for almost any virtualization workloads. vSAN provides the flexibility to configure the level of performance and redundancy required for workloads, such as:

- MS Exchange / SQL
- Oracle
- SAP

This allows for the policy to be changed as business needs demand, lowering costs, and simplifying operations for the organization as well as providing a building-block scale-up and out architecture.

When implementing a vSAN design for Tier 1 / Tier 2 / Tier 3 workloads, there are a number of specific design considerations that must be taken into account to be successful:

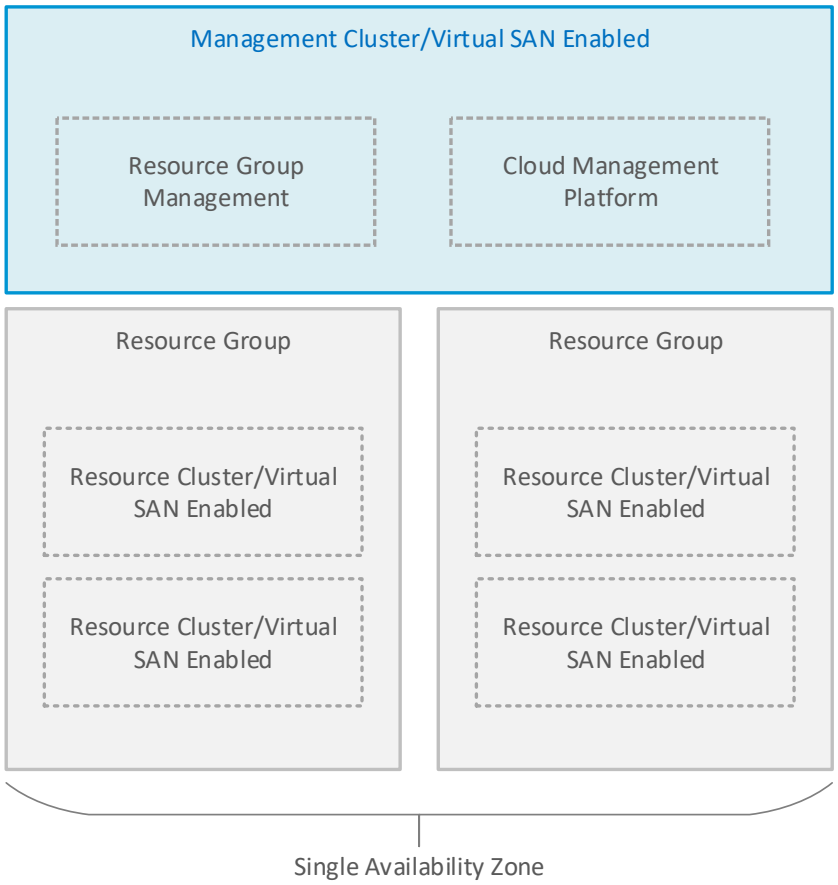
- Normally workloads are a balanced workload type, where the recommended designs are a mix between the need for performance and the need for capacity. Policy recommendations must take this into account to prevent a single policy from being used for all workloads, negating the benefit of software-defined storage. Tier 1 or business critical applications may have varying workloads however.
- When designing the disk group configuration, verify that there are enough disk groups configured to provide the level of availability and performance required by the environment and verify that the workloads selected are appropriate for use with vSAN, and will benefit from the policy-based configurations.
- Application needs and dependencies assessment must be completed before the design is implemented so that the policies meet the needs of the workloads to be hosted.

For more information on employing vSAN on various workload types, please refer to the product-specific papers in the reference section.

7.1.2 Management and Cloud Management Platform Clusters

Traditionally, providing dedicated storage for management-only clusters increased providers' infrastructure costs. It required the purchase of an additional storage array to provide the expected performance and availability for what is typically a high I/O performance and highly-available virtual environment. These costs, in turn, were inherently passed on to the consumer. With vSAN, these costs are significantly reduced, making it an ideal solution for a dedicated management environment. Additionally, leveraging vSAN for the management cluster eliminates single points of failure and is completely integrated with vCenter Server. See the following figure for a logical diagram of this architecture.

Figure 29. vSAN Based Management Cluster



For more information on implementing vSAN for management and cloud management platform clusters, see the *Leveraging VMware Virtual SAN for Highly Available Management Clusters* vCAT-SP paper (http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vcat/vmware-leveraging-virtual-san-for-ha-management-clusters.pdf).

7.1.3 DMZ/Isolated Clusters

vSAN storage is an effective security boundary for a DMZ isolated cluster configuration. By using only local storage presented by vSAN, the environment can be completely segregated into its own security zone. In this configuration, vSAN means that workloads do not share the same underlying storage, and security and storage policies are provided based on the workloads. This setup means that these critical systems are highly available and perform well in the environment.

Employing vSAN for use in DMZ, isolated or perimeter network clusters can help VMware Cloud Providers:

- Service providers today typically run fully isolated clusters except shared storage
- Buying a separate array for just DMZ is prohibitively expensive
- Smaller arrays often do not offer same performance or features
- Very common use case for vSAN, because it offers full isolation

When implementing a vSAN design for DMZ/isolated cluster configurations, there are specific design considerations that must be taken into account to be successful. The following considerations must be taken into account:

- Typically, DMZ/isolated cluster workloads are a balanced workload type, where the recommended designs are a mix between the need for performance and the need for capacity. Policy recommendations must take this into account to prevent a single policy being used for all workloads, negating the benefit of software-defined storage.
- Migrations to and from these hosts must be considered. If there are no other underlying storage connections, which is the recommendation for DMZ hosts, it might take a significant amount of time to migrate to and from the vSAN volume.

7.1.4 Development Platform Clusters

vSAN is a good use case for test/development/staging environments where storage cost is a consideration. By implementing vSAN storage, service providers can avoid purchases of expensive storage arrays. This lowers the total cost of ownership for the environment, and provides a faster time to provision environments for utilization for these critical business needs. In addition, vSAN:

- Can be programmatically accessed
- Can differentiate through policies instead of tiers
- Can provide fast provisioning and agility, with no impact on production
- Can run next to the traditional storage infrastructure
- Can be used to extend into an existing platform to meet *different* needs
- Can be front-ended by vSphere with VMware Integrated OpenStack, vCloud Director, or VMware vRealize Automation™
- The test/development/staging environment-specific prerequisites for vSAN are as follows:
    - It is a normal practice to have separate clusters for the different types of workloads (test, development or staging). It is a best practice to do an assessment for each workload to be able to design the cluster configuration for disk groups accordingly. For example, a staging environment might want to more closely mimic the production environment, and thus the policy must include availability considerations as well.
    - Availability and capacity is not as pertinent to the end configuration, because these are in many cases disposable and not subject to SLAs. Also, there is a high turnover rate for the virtual machines.

7.1.5 Backup/Disaster Recovery Target

vSAN storage can also provide a low-cost disaster recovery solution, enabled through features such as vSphere Data Protection and vSphere Replication that allow you to back up and replicate to any storage platform. This also allows the utilization of low-cost local storage to reduce costs but still provide a reliable solution.

When implementing a vSAN design for backup/disaster recovery target environments, there are specific design considerations to take into account, even if the target storage is not required to match the active site. The following considerations must be taken into account:

- Normally backup/disaster recovery target workloads are a capacity workload type, where the recommended design services the need for capacity on the storage volumes. In addition, availability is also important, so there will typically be more replicas to make sure that the environment is recoverable.

- Performance of the destination storage is important to consider. There is no single recommendation, however, the time to recover the service can depend on the response time for those virtual machines to be powered on.
- The storage policies are not guaranteed on the recovery site if the storage destination is something other than vSAN.

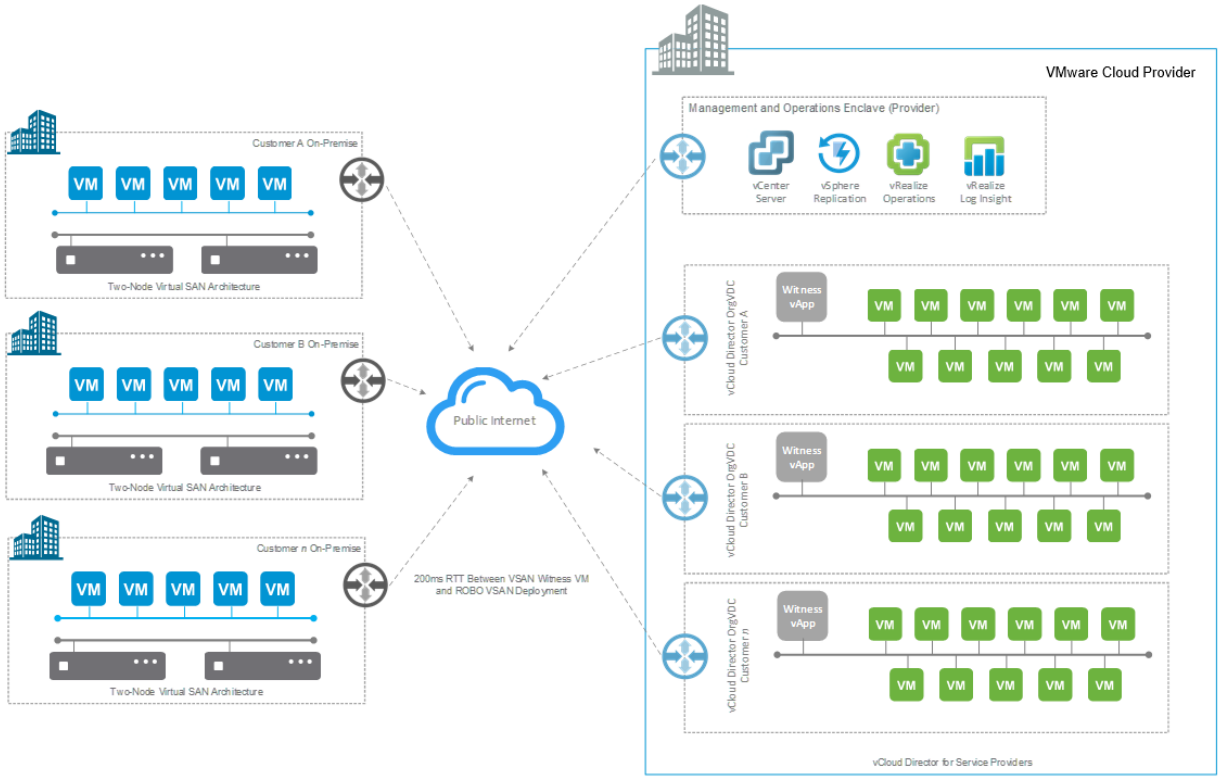
7.2 Remote and Branch Offices

Very similar to the local data center design described in the previous section is the Remote and Branch Offices design because it is a single-site deployment. Two-node clusters have traditionally not been possible with vSAN because there is a three host minimum to make sure that components can properly be protected. However, the former design requirements of minimum of a three-node configuration is not cost-effective for very small ROBO deployments. For this reason, VMware introduced a two-node design.

However, a two-node design can lead to split brain configurations when a failure occurs. Hence, a witness node at a third site is required to avoid that failure condition. The witness node is a VM that can be installed at the nearest data center site. It requires a round-trip latency of 200 ms.

vSAN 6.2 can be configured from the configuration wizard, allowing for you to setup a two-node cluster with a witness, allowing for smaller configurations of vSAN. This is popular with smaller environments.

Figure 30. ROBO Site Deployment with Third Site Witness Appliance



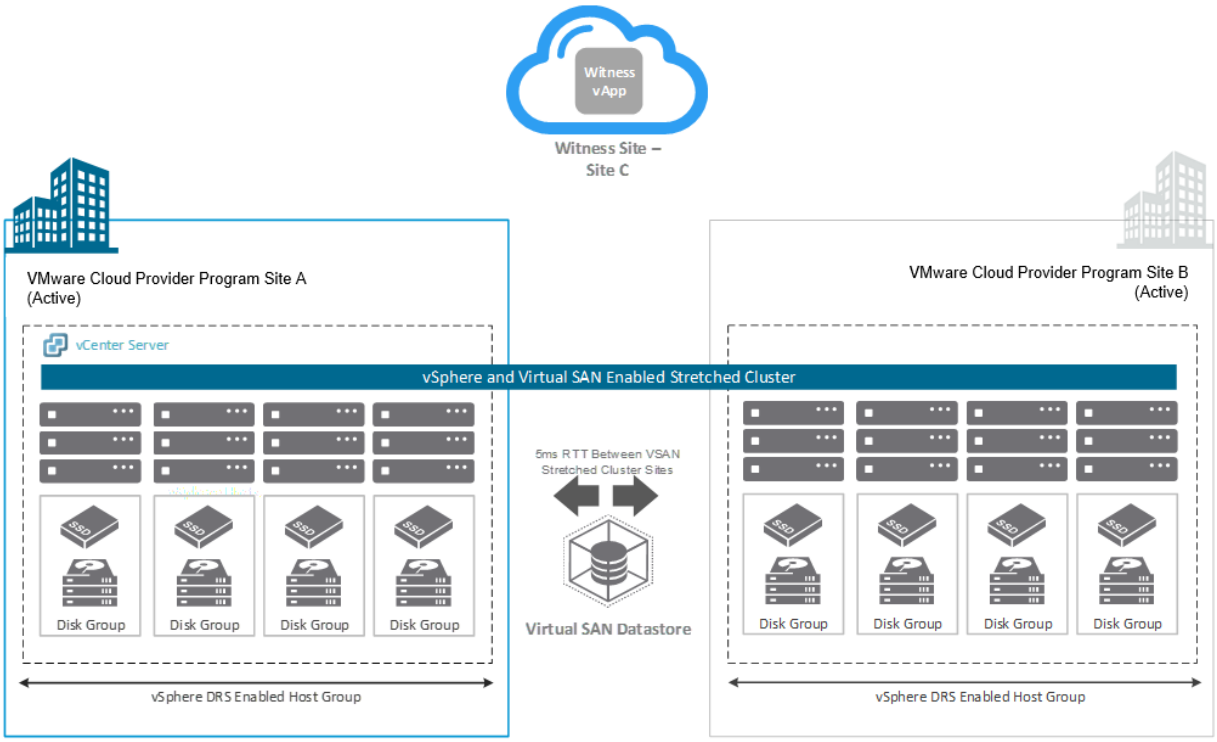
For more information relating to service provider use cases for ROBO deployments, see the vCAT-SP paper, *VMware vSAN Two-Node Architecture Service Provider Use Cases* (https://www.vmware.com/files/pdf/vcat/vmw-virtual-san-two-node-architecture-service-provider-use-cases.pdf).

7.3 vSAN Stretched Cluster Deployments

Stretched storage with vSAN allows you to split the vSAN cluster across two sites, so that if a site fails, providers are able to seamlessly fail over to the other site without any loss of data. vSAN in a stretched storage deployment accomplishes this by synchronously mirroring data across the two sites. The failover is initiated by a witness VM that resides in a central location, accessible from both sites. This is a specific implementation for environments where disaster avoidance and unscheduled downtime are key requirements.

A vSAN stretched cluster is a specific deployment where the provider sets up a vSAN cluster with two disparate Active/Active sites with an identical number of ESXi hosts distributed evenly between the two sites. The witness host resides at a third site and the sites are connected by way of a high-bandwidth, low-latency link. The third site, hosting the witness host, is connected to both of the Active/Active data sites. The sites can be a combination of the VMware Cloud Provider Program, customer, and third-party data centers. See the following figure for an example.

Figure 31. VMware Cloud Provider Program Stretched Cluster Example



In a vSAN stretched cluster implementation, each site is configured as a vSAN fault domain, and there is only one witness host in any configuration. Each site can be considered a fault domain and a maximum of three sites (two data, one witness) is supported. For deployments that manage multiple stretched clusters, each cluster must have its own unique witness host.

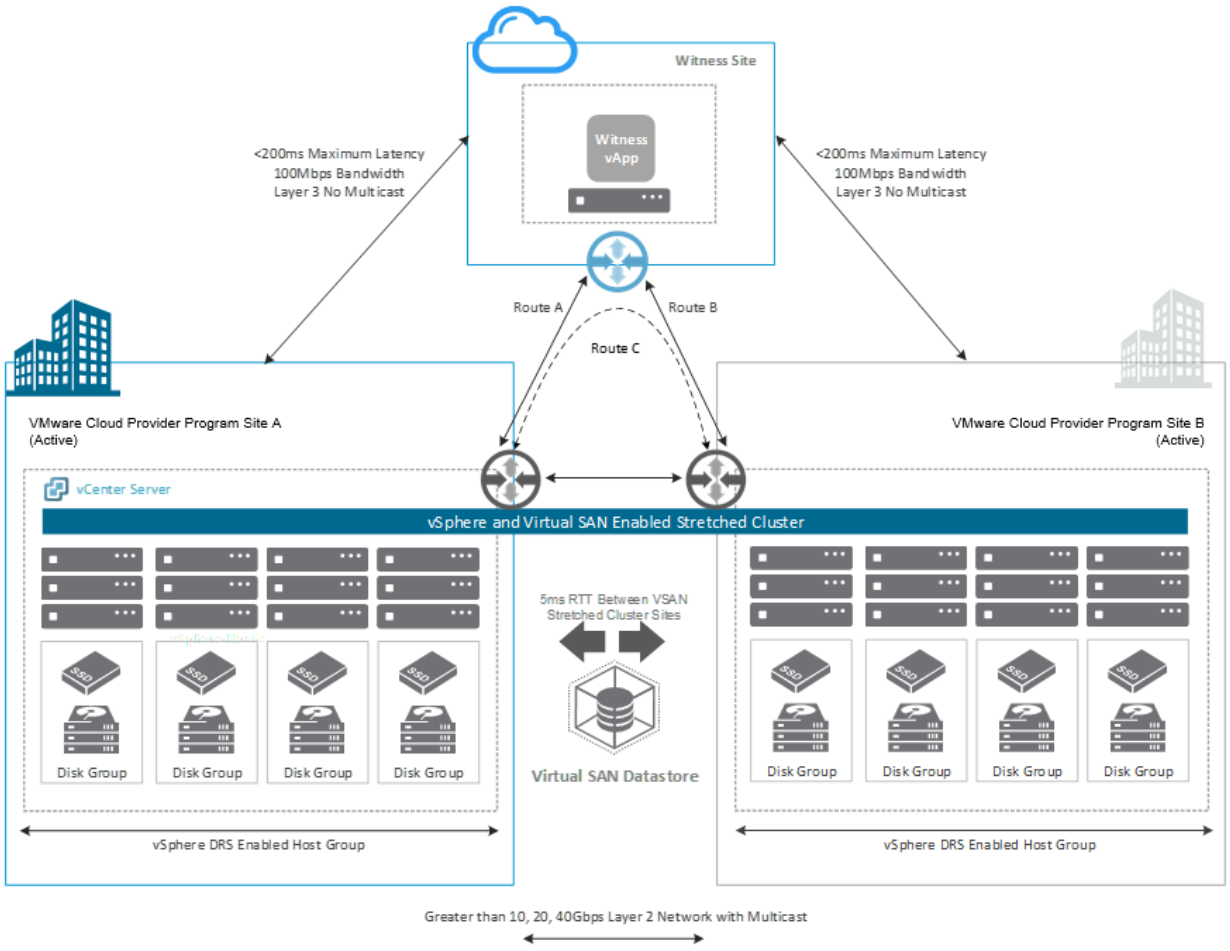
A virtual machine deployed on a vSAN stretched cluster has one copy of its data on site A, a second copy of its data on site B, while any witness components are placed on the witness host in site C. This configuration is achieved through fault domains and affinity rules. In the event of a complete site failure, there is a full copy of the virtual machine data available, as well as greater than 50 percent of the components. This allows the virtual machine to remain available on the vSAN datastore. If the virtual machine needs to be restarted on the other data site, vSphere HA accommodates this task.

Geographic distance is, in theory, not a concern when designing a vSAN stretched cluster. The key requirement is the latency between the respective sites. VMware requires a maximum latency of no more than 5 ms RTT (Round-Trip Time) between data sites and no more than 200 ms RTT between data sites and the witness host. As long as the latency requirements are met, there is no restriction on geographic distance.

As discussed earlier, vSAN stretched cluster requires three disparate sites, and each site must communicate on the management, vSAN, VM, and vSphere vMotion networks. Detailed networking design is beyond the scope of this document. However, to minimize uncertainty with the implementation, VMware recommends that providers implement a stretched L2 between the data sites, and a L3 configuration between the data sites and the witness site.

In the example illustrated in the following figure, the choice was made to use a virtual witness connected over L3 with static routes. The witness is deployed on a physical ESXi host with two preconfigured networks for the management and vSAN networks, respectively. The data sites are connected by way of stretched L2 which backs the management, vSAN, VM, and vSphere vMotion networks. All hosts in the cluster must be able to successfully communicate, and to facilitate this communication, static routes must be configured (per host) between the data hosts in Site A and B, and the witness host in Site C, for vSAN traffic to flow between the data sites and witness site.

Figure 32. Network Connectivity for Stretched Cluster



Ultimately, the success and design considerations of a specific vSAN stretched cluster implementation depend upon many factors ranging from choice of topology to the physical capabilities of a provider's networking infrastructure.

With vSAN 6.2, stretched clusters have been enhanced to simplify the creation of the configuration. A new graphical configuration wizard assists with the configuration as appropriate. vSAN stretched clustering is a specific configuration implemented in environments where disaster/downtime avoidance is a key requirement. However, the maximum number of hosts in a stretch cluster configuration remains at 31, where Site 1 contains 15 hosts, Site 2 contains 15 hosts, and site 3 contains the witness host or virtual appliances.

For detailed guidance on designing a vSAN stretch cluster, consult the *VMware Virtual SAN 6.1 Stretched Cluster & 2 Node Guide* at https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/VMware-Virtual-SAN-6.1-Stretched-Cluster-Guide.pdf.

7.3.1 vSAN Witness

As highlighted in the previous section, the vSAN witness is used in place of having a third site of hosts for the redundancy provided by vSAN. With both stretched cluster and two-node cluster configurations, the witness stores components of virtual machines and is used to monitor both of the sites.

The witness can be either an ESXi host that resides in another site and has reliable network connectivity to the vSAN cluster, or the vSAN witness appliance can be used. Either way, it is added to the cluster during the configuration of stretched clustering.

For more details, see the *VMware Virtual SAN 6.1 Stretched Cluster & 2 Node Guide* at https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/VMware-Virtual-SAN-6.1-Stretched-Cluster-Guide.pdf.

7.4 Hosted SDDC

For customers who want single-tenant, dedicated compute, network, and storage, a hosted vSphere infrastructure is the ideal solution. In this scenario, the entire infrastructure, physical and virtual, is owned and managed by the service provider. Consumers are able to access this infrastructure through a service portal where they have the ability to provision and manage virtual machines, on demand, through an intuitive graphical user interface.

As discussed earlier, vSAN is a hyper-converged offering which makes the hosted vSphere environment an ideal use case. First, it enables providers to stand up an infrastructure in a timely and predictable manner. There is no need to add external arrays or configure LUNs, so the time to deployment is greatly reduced and the process is highly repeatable. Second, vSAN is supported on all major server OEM platforms. Most service providers have a preferred server vendor and receive favorable pricing. Because vSAN is likely supported on their preferred platform, the existing procurement vehicles do not need to be altered.

In summary, service providers can leverage their current VMware skill sets, reduced time to deployment, lower CapEX/OpEX, and the ability to use existing processes and procedures for procuring equipment. End clients get increased agility, reduced risk, and reduced CapEx/OpEx. These benefits make vSAN the best hyper-converged offering for hosted vSphere environments.

7.5 Hosted Private SDDC Cloud

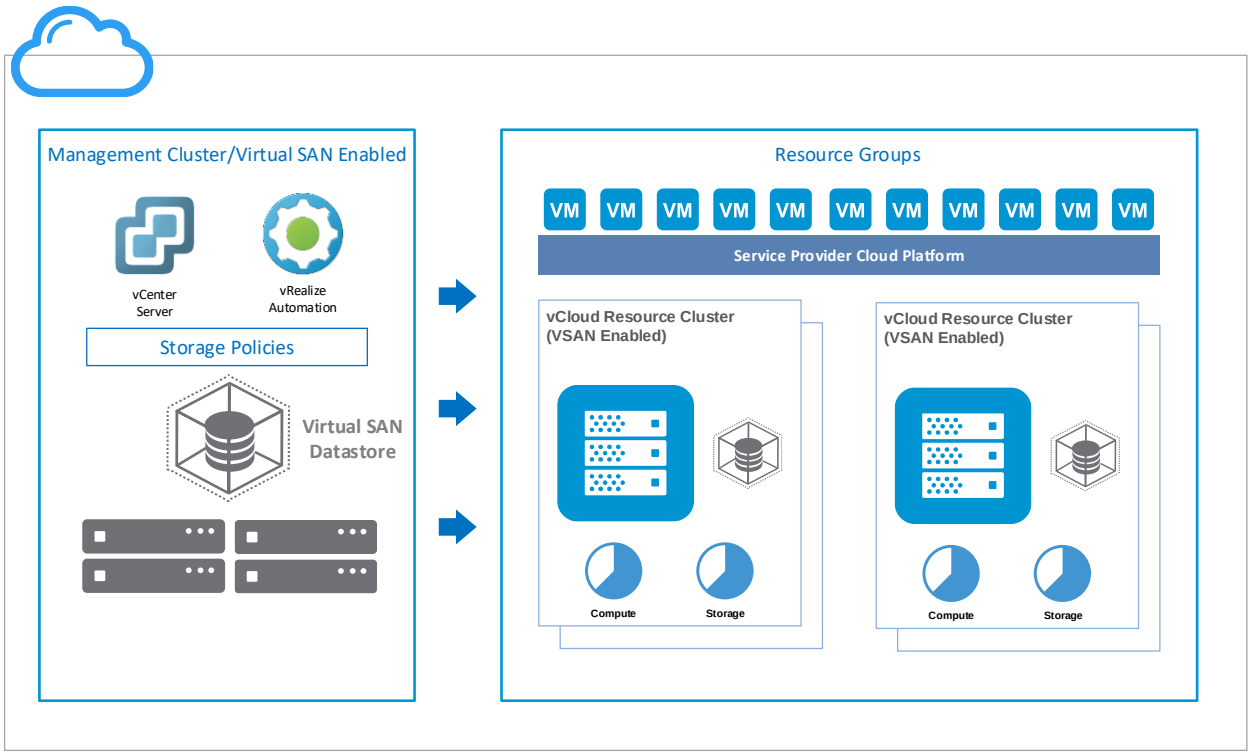
vRealize Automation gives providers another option for offering a hosted single-tenant virtualized solution to customers. A hosted vRealize Automation solution allows customers to manage their own resources using custom provisioning workflows, manage virtual machines, create snapshots, and distribute specific resources into resource pools and/or catalogs as needed. With a host vRealize Automation solution, customers can benefit from the agility of a private cloud management platform, without the overhead that is required to deploy and manage the private cloud management infrastructure.

vRealize Automation can be used to provision virtual machines and/or applications onto a vSAN datastore while leveraging the storage-based policy management capabilities of vSAN. This is extremely

valuable because storage policies can be configured on a per-application or per-virtual machine basis, allowing providers to assign policies based on a particular use case, such as tiered storage.

When implementing vSAN in a hosted Private Cloud model, vSAN is managed in a way similar to how a customer experiences an on-premises implementation of vSphere with vSAN. The vSphere Web Client provides the primary interface for any required maintenance and configuration of vSAN. The hosting solution can either be managed by the service provider or self-managed by the consumer.

Figure 33. vRealize Automation with vSAN Logical Architecture



7.6 Public Cloud (vCloud Director for Service Providers)

When leveraging vSAN with a vCloud Director based cloud infrastructure, the typical service provider architecture consists of two clusters based on vSAN. The first is a management cluster, which hosts all components needed for a vCloud Director environment, in conjunction with a resource cluster. This enables the provider to start out with a relatively low investment in hardware and quickly scale as new customers are added or when existing customers require new hardware in the cloud environment.

This design is in alignment with typical cloud architecture in which the management components are deployed in the management cluster and tenant workloads are hosted in the resource cluster. It enables the provider to offer different SLAs for management components and tenant workloads, provides separation of duties, and allows both clusters to easily scale by adding hosts where needed. The following figure shows this architecture.

When designing a vSAN backed vCloud Director infrastructure, VMware recommends not using a vSAN datastore for catalogs. All catalog media images are uploaded as file objects into the same directory structure by vCloud Director. If the same vSAN datastore is used as storage policy for different catalogs, they share one VM Home Namespace object with maximum size of 256 GB. In this case, use a third-party virtual storage appliance that provides NFS file services to provide catalogs that scale larger than 256 GB.

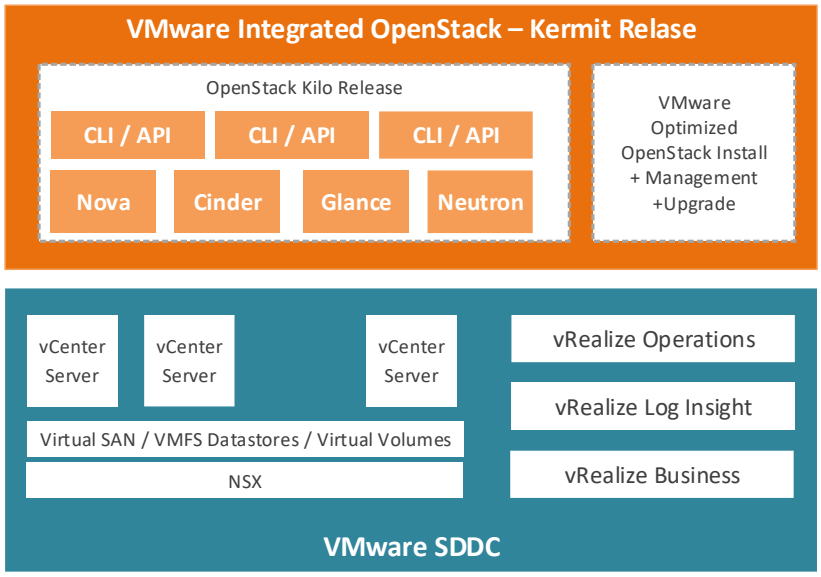
For more information on employing a vSAN platform as part of a vCloud Director for Service Providers platform, see the vCAT-SP paper, *Developing a Hyper-Converged Storage Strategy for VMware vCloud Director with VMware Virtual SAN* (http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vcat/vmware-developing-a-vcloud-director-storage-strategy-with-virtual-san.pdf).

7.7 VMware Integrated OpenStack

VMware Integrated OpenStack is a VMware supported OpenStack distribution, which is the fastest and most reliable route to running a production-grade OpenStack environment.

VMware Integrated OpenStack builds on your existing vSphere SDDC environments. It is a full OpenStack distribution, which has the OpenStack source codes optimized and hardened to run on VMware technologies and includes drivers for VMware technologies. VMware also provides OpenStack-aware cloud management platform integration such as vRealize Operations Management Pack for OpenStack and VMware vRealize Log Insight™ Content Pack for OpenStack, simplifying monitoring, troubleshooting, and business visibility for your OpenStack clouds.

Figure 34. VMware Integrated OpenStack



This is a fully validated architecture that is tested and supported for vSAN and VMware provides a single support contact for the entire infrastructure including supporting the OpenStack open source code. VMware Integrated OpenStack is available for use by VMware Cloud Providers.

Future releases of VMware Integrated OpenStack will be available regularly to stay in sync with the new open source OpenStack community versions. Similarly, VMware plans to continue to update VMware Integrated OpenStack to incorporate additional VMware SDDC functionality made available through new releases of products such as vSphere, VMware NSX, and vSAN.

7.8 Horizon and End User Computing

A vSAN infrastructure is particularly useful for virtual desktops by providing a scale-out model, using predictive repeatable infrastructure blocks. This lowers costs and simplifies operations for the organization. Typically, in a traditional virtual desktop infrastructure (VDI) environment, the storage costs are high. By implementing vSAN for VDI solutions, dramatic cost reductions can be found.

When implementing a vSAN design for VDI, there are specific design considerations which need to be taken into account to be successful:

- Normally VDI workloads are a performance-based workload type, where the recommended designs are built to have as high performance as possible for a large number of VMs. Availability and capacity is not as pertinent to the end configuration, because these are in many cases disposable, and destroyed after use. In addition, they have a predictable footprint and can be tuned to provide more efficient performance.
- When designing for floating linked clones' workloads, it is important to consider application placement/availability. Floating linked clones, created by VMware View® Composer™, are frequently refreshed resulting in the loss of any user-specific data and applications on the desktop. A network-based profile management system is required to maintain user settings. Applications must be either installed in the parent image or streamed over the network from a file share.
- Dedicated linked clone workloads have similar considerations to that of floating linked clones. Desktops must be refreshed/recomposed regularly to prevent linked clone growth. Profile management and application delivery mechanisms are required to maintain user-specific settings.
- Existing desktop management and patching practices can be used for maintaining full clone desktops. However, full clone desktops require a large disk footprint.

Conclusion

Achieving economies of scale means scaling storage resources in a consistent and predictable manner. Scaling the storage capacity in vSAN can be achieved by simply adding more mechanical disks (hybrid configuration) or SSDs (all-flash) to an existing disk group or by creating an entirely new disk group. Scaling the performance layer (flash layer) of vSAN is achieved in a similar way to scaling the storage capacity in the sense that you can quickly select the necessary disk devices and create a disk group.

With the combined resources of the VMware Cloud Provider Program and vSAN, an organization can achieve business-critical levels of performance and availability for a variety of workloads including databases, ERP systems, streaming content, and web services. All this can be achieved while accelerating the time to deployment, reducing costs, and simplifying operations.

Assumptions and Caveats

VMware and other third-party hardware and software information provided in this document is based on the current performance estimates and feature capabilities of the versions of code indicated. These are subject to change.

Reference Documents

10.1 Supporting Documentation

For more information, see the following configuration and administration guides, white papers, blogs, and best practices documents.

Document Title	Link or URL
VMware Virtual SAN 6.2 Design and Sizing Guide	http://www.vmware.com/files/pdf/products/vsan/virtual-san-6.2-design-and-sizing-guide.pdf
VMware Virtual SAN Health Check Plug-in Guide	http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/products/vsan/vmw-gdl-vsan-health-check.pdf
VMware vSphere 6.0 Configuration Maximums Guide	https://www.vmware.com/pdf/vsphere6/r60/vsphere-60-configuration-maximums.pdf
vsanSparse - Tech Note for Virtual SAN 6.0	https://www.vmware.com/files/pdf/products/vsan/Tech-Notes-Virtual-San6-Snapshots.pdf
VMware Virtual SAN 6.2 Space Efficiency Technologies	https://www.vmware.com/files/pdf/products/vsan/vmware-vsan-62-space-efficiency-technologies.pdf
VMware Virtual-SAN 6.0 Proof of Concept Guide	https://www.vmware.com/files/pdf/products/vsan/VMwareVirtual-SAN6-Proof-Of-Concept-Guide.pdf
VMware Virtual SAN Stretched Cluster Bandwidth Sizing Guidance	https://www.vmware.com/files/pdf/products/vsan/vmware-virtual-san-6.1-stretched-cluster-bandwidth-sizing.pdf
VMware Virtual SAN 6.2 Stretched Cluster & 2 Node Guide	http://www.vmware.com/files/pdf/products/vsan/VMware-Virtual-SAN-6.2-Stretched-Cluster-Guide.pdf
VMware Virtual SAN Layer 2 and Layer 3 Network Topologies	https://www.vmware.com/files/pdf/products/vsan/vmware-vsan-layer2-and-layer3-network-topologies.pdf
VMware Virtual SAN 6.0 Performance Scalability and Best Practices	http://www.vmware.com/files/pdf/products/vsan/VMware-Virtual-San6-Scalability-Performance-Paper.pdf

Document Title	Link or URL
An overview of VMware Virtual SAN caching algorithms	https://www.vmware.com/files/pdf/products/vsan/vmware-virtual-san-caching-whitepaper.pdf
Understanding Data Locality in VMware Virtual SAN	https://www.vmware.com/files/pdf/products/vsan/VMware-Virtual-SAN-Data-Locality.pdf
VMware Virtual SAN Health Check Guide	https://www.vmware.com/files/pdf/products/vsan/VMware-Virtual-SAN-Health-Check-Guide-6.1.pdf
VMware Virtual SAN Diagnostics and Troubleshooting Reference Manual	http://www.vmware.com/files/pdf/products/vsan/VSAN-Troubleshooting-Reference-Manual.pdf

10.2 Tools

The VMware vSAN TCO and Sizing Calculator tool (http://vsantco.vmware.com/)

10.3 Further Information

10.3.1 VMware Ready Nodes

http://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan

10.3.2 VMware Compatibility Guide

http://vmwa.re/vsanhclc

10.3.3 vSAN Community Page

https://communities.vmware.com/community/vmtn/vsan

10.3.4 Key Bloggers

http://cormachogan.com/vsan/

http://www.yellow-bricks.com/virtual-san/

http://www.virtuallyghetto.com/category/vsan

http://www.punchingclouds.com/tag/vsan/

http://blogs.vmware.com/vsphere/storage

http://www.thenicholson.com/vsan

10.3.5 Links to Existing Documentation

http://www.vmware.com/products/virtual-san/resources.html

https://www.vmware.com/support/virtual-san

10.3.6 VMware Support

https://my.vmware.com/web/vmware/login

http://kb.vmware.com/kb/2006985 - How to file a Support Request

http://kb.vmware.com/kb/2072796 - Collecting Virtual SAN support logs

10.3.7 Additional Reading

http://blogs.vmware.com/vsphere/files/2014/09/vsan55-sql-dvdstore-perf-v2.pdf

https://www.vmware.com/files/pdf/products/vsan/VMW-TMD-Virt-SAN-Dsn-Szing-Guid-Horizon-View.pdf

https://www.vmware.com/files/pdf/products/vsan/VMware-Virtual-SAN-Network-Design-Guide.pdf