VMware vSAN 6.5 Technical Overview

First Published On: 09-21-2016
Last Updated On: 11-12-2016

Table of Contents

1. Introduction
    1.1. Introduction
2. What's New in vSAN 6.5
    2.1. Witness Traffic Separation
    2.2. Two-Node Direct-Connect
    2.3. iSCSI Target Service
    2.4. Cloud Native Application Storage
    2.5. Full Featured PowerCLI Commandlets
    2.6. 512e Drive Support
3. Architecture
    3.1. Hardware and Deployment Options
    3.2. vSAN Embedded in vSphere
4. Simple Deployment and Operations
    4.1. Enabling vSAN
    4.2. Health Check
    4.3. Proactive Tests
    4.4. Capacity Reporting
    4.5. Performance Reporting
5. Policy Based Management and Automation
    5.1. Storage Policy Based Management
    5.2. Automation
6. Enterprise Availability
    6.1. Objects and Component Placement
    6.2. Rack Awareness
    6.3. Stretched Clusters
7. Space Efficiency
    7.1. Deduplication and Compression
    7.2. RAID-5/6 Erasure Coding
8. Quality of Service
    8.1. IOPS Limits
9. Summary
    9.1. Summary

1. Introduction

A brief introduction to VMware vSAN

1.1 Introduction

Introduction to vSAN

VMware vSAN is radically simple, enterprise-class storage for hyper-converged infrastructure (HCI). Uniquely embedded in the VMware vSphere hypervisor, vSAN delivers flash-optimized, high-performance storage. It leverages commodity x86 server components to drastically lower TCO by up to 50% and deliver all-flash solutions for as low as half the price of competitive hybrid HCI systems. Seamless integration with vSphere and the entire VMware stack makes it the simplest storage platform for business-critical workloads, virtual desktop infrastructure (VDI), remote office IT, disaster recovery, and DevOps infrastructures. Customers of all industries and sizes trust vSAN to run their most important applications.

Key Benefits

- **Radically Simple** – Configure with just a few clicks using the vSphere Web Client and automate management using storage policies. Eliminate expensive, complex, purpose-built storage array hardware and automate management of storage service levels through VM-centric policies.
- **High Performance** – Deliver up to 150,000 IOPs per host and over 6 Million IOPs per cluster with predictable sub-millisecond response time from a single, all-flash configuration. Built on an optimized I/O data path designed for flash speeds, vSAN delivers much better performance than virtual appliance and external device solutions.
- **Elastic Scalability** – Easily grow storage performance and capacity by adding new nodes or drives without disruption. Linearly scale capacity and performance from 2 to 64 physical vSphere hosts per cluster.
- **Lower TCO** – Lower storage costs by up to 50% by deploying standard x86 servers with local storage devices for low upfront investment and by shrinking data center footprint and operational overheads. Further improve total cost of ownership with features like deduplication, compression, erasure coding, iSCSI target services, and automation with PowerCLI.
- **Enterprise-Class Availability** – Enable maximum levels of data protection and application availability with built-in tolerance for disk and host failures, stretched cluster configurations, and compatibility with DR solutions such as VMware Site Recovery Manager and vSphere Replication.
- **Advanced Management** – Management for storage, compute and networking with advanced performance and capacity monitoring all within the vSphere Web Client.

Use cases for vSAN include mission-critical applications, test and development, remote and branch offices, disaster recovery sites, and virtual desktop infrastructure (VDI). vSAN supports a number of configuration options such as 2-node clusters for small implementations at remote offices to clusters as large as 64 nodes delivering over 6 million IOPS. Stretched clusters can also be configured to provide resiliency against site failure and workload migration with no downtime for disaster avoidance. VMware provides the market-leading HCI platform powered by vCenter Server, vSphere and vSAN.

2. What's New in vSAN 6.5

The new features and benefits of vSAN 6.5

2.1 Witness Traffic Separation

Separating vSAN Witness Traffic

New in vSphere 6.5 and vSAN 6.5 is the ability to separate witness network traffic from the network that connects the two physical hosts in a two-node cluster. This configuration option reduces complexity by eliminating the need to create and maintain static routes on the physical hosts. Security is also improved as the data network (between physical hosts) is completely separated from the WAN for witness traffic.

A vSAN Witness is a virtual appliance that is utilized in stretched clusters and two-node configurations. The witness acts as a "tie-breaker" in certain events such as a network partition between hosts in a 2-node cluster. The witness virtual appliance stores metadata such as witness components. It does not store data components such as VM Home and virtual disk (VMDK) files.

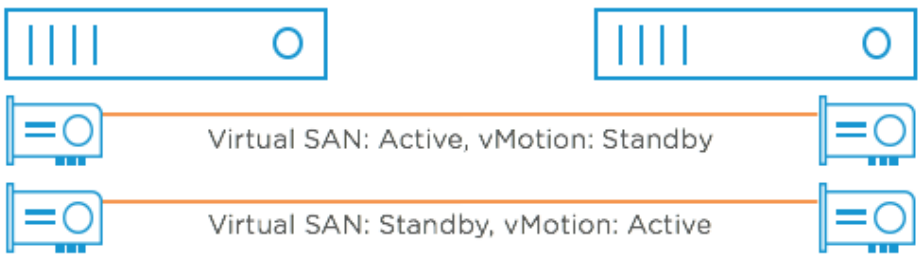
Two-node VSAN cluster configurations are commonly used for branch offices and remote locations running a small number of virtual machines. A witness virtual appliance must not run on either of the two physical nodes in the cluster as this would defeat the purpose of the witness. The recommended location for the witness virtual appliance(s) is a primary data center connected to the remote locations by a WAN. Having the witness virtual appliances centrally located facilitates ease of management.

2.2 Two-Node Direct-Connect

Two-Node Configuration with Crossover Cables

Physical Hosts in a two-node cluster can be connected using crossover cables. Organizations with existing 100Mbps or 1Gbps networking hardware can continue to utilize this equipment and enable 10Gbps connections for vSAN and vMotion. All-flash vSAN configurations (including two-node) require 10Gbps connectivity and vMotion performance is enhanced using faster connections between hosts. Connection reliability between the physical hosts is also improved with direct connections.

As with any storage fabric, redundant connections are recommended. In addition to providing resiliency against a NIC or cable failure, two connections between physical hosts enables administrators to separate vSAN from vMotion traffic. It is possible for vMotion to saturate a 10Gbps link when migrating multiple virtual machine simultaneously, e.g., when a host is put into maintenance mode. This scenario has the potential to impact vSAN performance while the migrations are taking place. The figure below shows one possible configuration that would prevent vSAN and vMotion from using the same connection unless there is a NIC or cable failure.



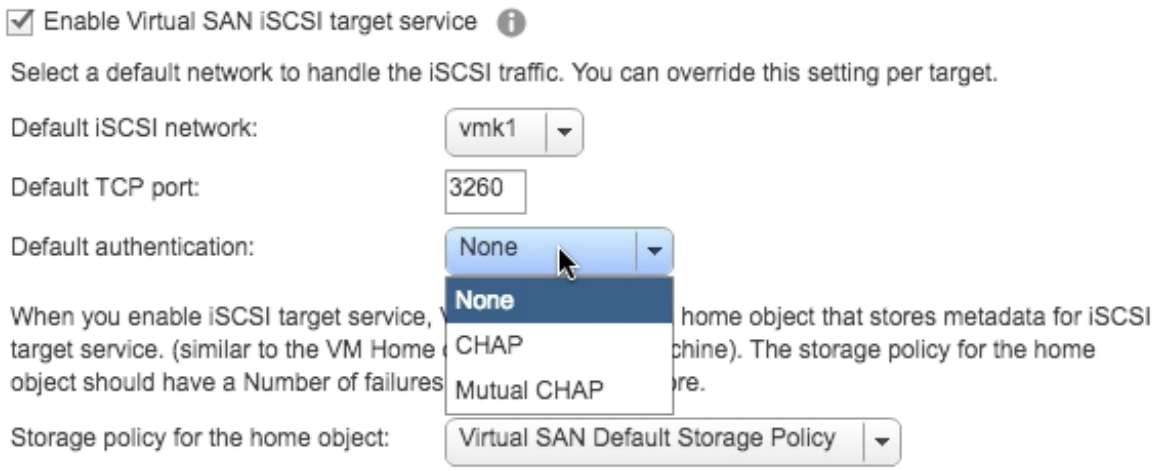
For more information on network configuration options and recommendations, see the vSphere documentation and the vSAN Network Design Guide.

2.3 iSCSI Target Service

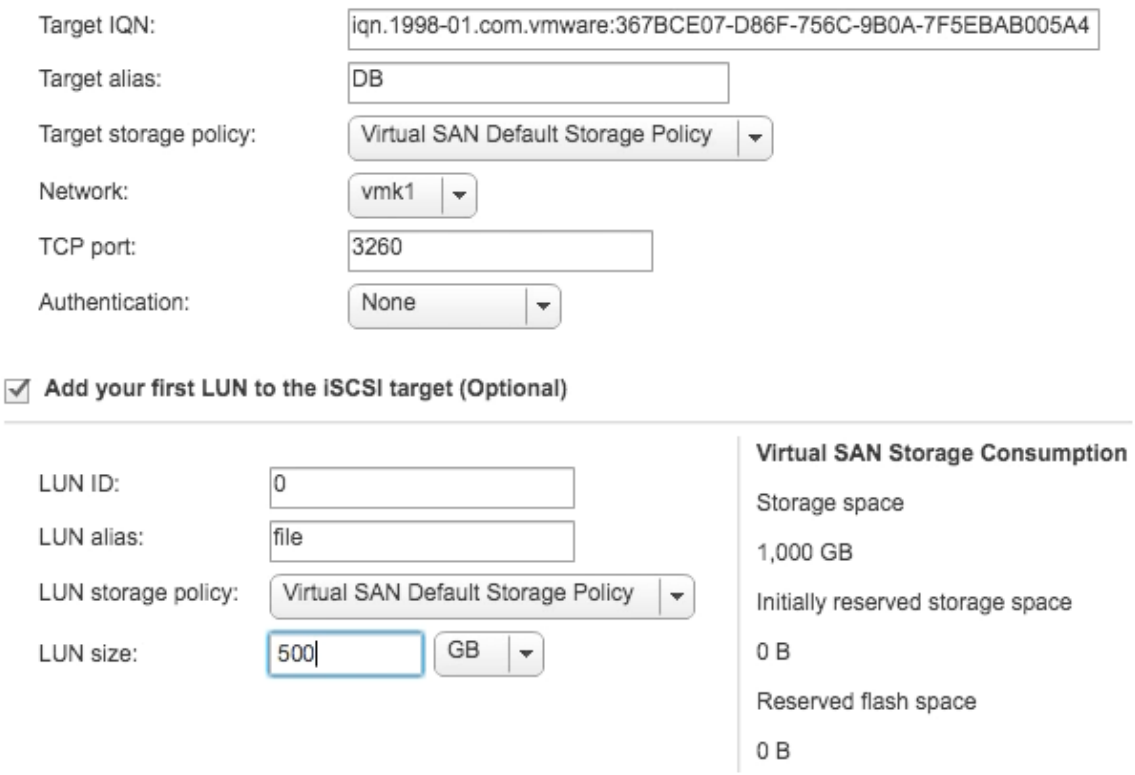
iSCSI Target Service

Block storage can be provided to physical workloads using the iSCSI protocol. The vSAN iSCSI target service provides flexibility and potentially avoids expenditure on purpose-built, external storage arrays. In addition to capital cost savings, the simplicity of vSAN lowers operational costs.

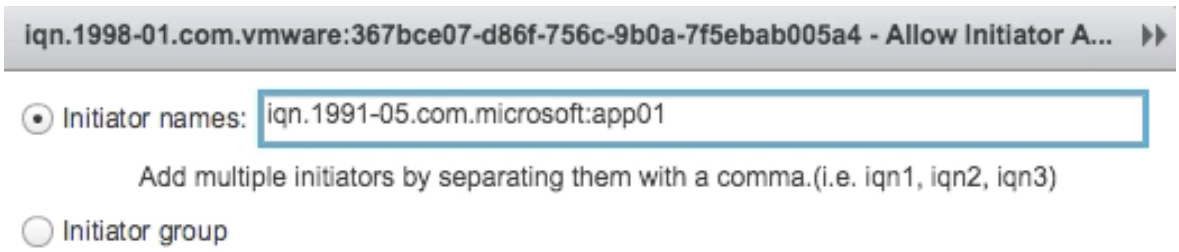
The vSAN iSCSI target service is enabled with just a few mouse clicks. CHAP and Mutual CHAP authentication is supported. vSAN objects that serve as iSCSI targets are managed with storage policies just like virtual machine objects.



After the iSCSI target service is enabled, iSCSI targets and LUNs can be created. The screen shot below shows target and LUN configuration options.



The last step is adding initiator names or an initiator group, which controls access to the target, as shown here.



In nearly all cases, it is best to run workloads in virtual machines to take full advantage of vSAN's simplicity, performance, and reliability. However, for those use cases that truly need block storage, it is now possible to utilize vSAN iSCSI Target Service.

2.4 Cloud Native Application Storage

Cloud Native Application Storage

New application architecture and development methods have emerged that are designed to run in today's mobile-cloud era. For example, "DevOps" is a term that describes how these next-generation applications are developed and operated. "Container" technologies such as Docker and Kubernetes are a couple of the many solutions that have emerged as options for deploying and orchestrating these applications. VMware is embracing these new application types with a number products and solutions. Here are a few examples:

Photon OS - a minimal Linux container host optimized to run on VMware platforms.

Lightwave - an open source project comprised of enterprise-grade identity and access management services.

Photon Controller - a distributed, multi-tenant ESXi host controller optimized for containers.

Cloud native applications naturally require persistent storage just the same as traditional applications. vSAN for Photon Controller enables the use of a vSAN cluster in cloud native application environments managed by Photon Controller. vSAN for Photon Controller includes developer-friendly APIs for storage provisioning and consumption. APIs and a graphical user interface (GUI) geared toward IT staff are also included for management and operations.

vSphere Integrated Containers Engine is a container runtime for vSphere, allowing developers familiar with Docker to develop in containers and deploy them alongside traditional virtual machine workloads on vSphere clusters. vSphere Integrated Containers Engine enables these workloads to be managed through the vSphere GUI in a way familiar to vSphere admins. Availability and performance features in vSphere and vSAN can be utilized by vSphere Integrated Containers Engine just the same as traditional virtual machine environments.

Docker Volume Driver enables vSphere users to create and manage Docker container data volumes on vSphere storage technologies such as VMFS, NFS, and vSAN. This driver makes it very simple to use containers with vSphere storage and provides the following key benefits:

- DevOps-friendly API for provisioning and policy configuration.
- Seamless movement of containers between vSphere hosts without moving data.
- Single platform to manage - run virtual machines and containers side-by-side on the same vSphere infrastructure.

vSAN along with the solutions above provides an ideal storage platform for developing, deploying, and managing cloud native applications.

2.5 Full Featured PowerCLI Commandlets

New and Improved PowerCLI Cmdlets

VMware's PowerCLI implementation is one of the most widely adopted extensions to the framework. It features a plethora of functions that abstract the vSphere API down to simple and powerful cmdlets including a number of cmdlets for vSAN. This makes it easy to automate a number of actions from simply enabling vSAN to deployment and configuration of a vSAN stretched cluster. Here are a few examples of what can be accomplished with vSAN and PowerCLI:

Assigning a Storage Policy to Multiple VMs with PowerCLI

Sparse Virtual Swap Files

Automated Deployments

With each new release of vSphere PowerCLI and vSAN APIs, the functionality becomes more robust. Cmdlets have been added and improved for these and many other operations:

- Enabling deduplication and compression
- Enabling the performance service
- Configuring the health check service
- Creating all-flash disk groups
- Fault domain management
- Stretched cluster configuration
- Selecting the data migration option for maintenance mode
- Retrieve capacity information

vSphere administrators and DevOps shops can utilize these new cmdlets to lower costs by enforcing standards, streamlining operations, and enabling automation for vSphere and vSAN environments.

2.6 512e Drive Support

512e Storage Devices

Disk drives have been using a native 512-byte sector size. Due to increasing demands for larger capacities, the storage industry introduced new formats that use 4KB physical sectors. These are commonly referred to as "4K native" drives or simply "4Kn" drives. Some 4Kn devices include firmware that emulates 512 byte (logical) sectors while the underlying (physical) sectors are 4K. These devices are referred to as "512B emulation" or "512e" drives.

vSphere 6.5 and vSAN 6.5 support the use of 512e drives. The latest information regarding support for these new drive types can be found in this VMware Knowledge Base Article: Support statement for 512e and 4K Native drives for VMware vSphere and VSAN (2091600)

3. Architecture

Choice of hardware platforms, network requirements, vSAN embedded in vSphere, and 2-node direct-connect configurations.

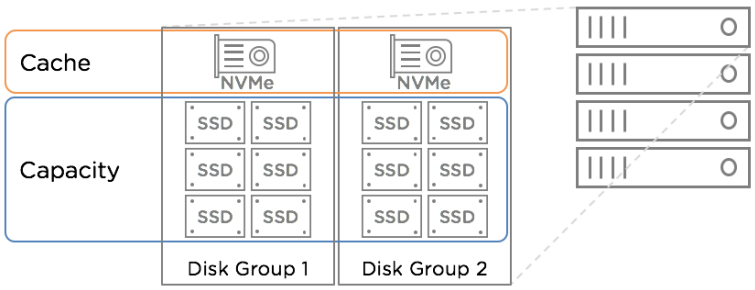
3.1 Hardware and Deployment Options

Servers with Local Storage

vSAN clusters consist of any number of physical server hosts from two to 64. Each host contains flash devices (all-flash configuration) or a combination of magnetic disks and flash devices (hybrid configuration) that contribute cache and capacity to the vSAN distributed datastore. Each host has one to five disk groups. Each disk group contains one cache device and one to seven capacity devices.

For all-flash configurations, the flash device(s) in the cache tier are used for write caching only as read performance from the capacity flash devices is more than sufficient. Two grades of flash devices are commonly used in an all-flash vSAN configuration: Lower capacity, higher endurance devices 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.

In a hybrid configuration, one flash device and one or more magnetic drives are configured as a disk group. A disk group can have up to seven magnetic drives for capacity. One or more disk groups are utilized in a vSphere host depending on the number of flash devices and magnetic drives contained in the host. Flash devices serve as read and write cache for the vSAN datastore while magnetic drives make up the capacity of the datastore. By default, vSAN will use 70% of the flash capacity as read cache and 30% as write cache.



Deployment Options

vSAN Ready Nodes are typically the easiest, most flexible approach when considering deployment methods. vSAN Ready Nodes are x86 servers, available from all of the leading server vendors that have been pre-configured, tested and certified for vSAN. Each Ready Node is optimally configured for vSAN with the required amount of CPU, memory, network, I/O controllers and storage devices.

Turn-key appliances such as Dell EMC VxRail provide a fully-integrated, preconfigured, and pre-tested VMware hyper-converged solution for a variety of applications and workloads. Simple deployment enables customers to be up and running in as little as 15 minutes.

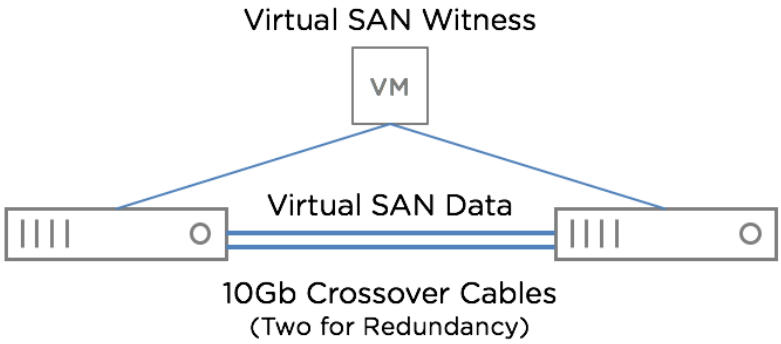
Custom configurations using jointly validated components from a number of OEM vendors is also an option. The vSAN Hardware Quick Reference Guide provides some sample server configurations as directional guidance and all components should be validated using the VMware Compatibility Guide for vSAN.

Network Considerations

All-flash vSAN configurations require 10Gb network connectivity. 1Gb connections are supported for hybrid configurations although 10Gb is recommended. Currently, vSAN requires the use of multicast on the network. Multicast communication is used for host discovery and to optimize network

bandwidth consumption for the metadata updates. This eliminates the computing resource and network bandwidth penalties that unicast imposes in order to send identical data to multiple recipients.

Support for using crossover cables in a 2-node cluster is new in vSAN 6.5. Utilizing crossover cables eliminates the need to procure and manage a 10Gb network switch for these hosts, which lowers costs – especially in scenarios such as remote office deployments. This configuration also improves availability and provides complete separation of vSAN witness traffic from data traffic.



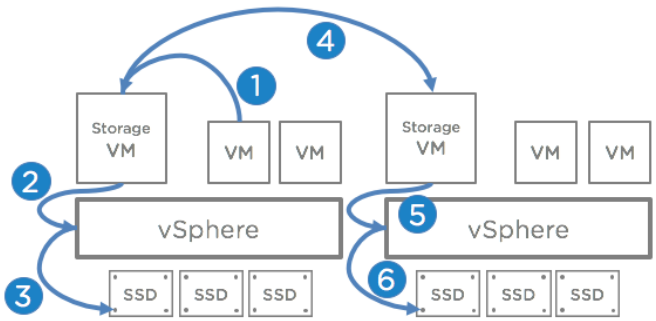
For more information on vSAN network requirements and configuration recommendations, see the vSAN Network Design Guide.

3.2 vSAN Embedded in vSphere

Storage Virtual Appliance Disadvantages

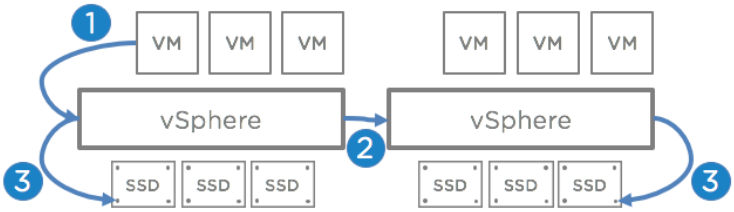
Storage in a hyper-converged infrastructure (HCI) requires compute resources that have been traditionally offloaded to dedicated storage arrays. Most HCI solutions require the deployment of storage virtual appliances to some or all of the hosts in the cluster to provide storage services to each host. These virtual appliances typically require CPU and/or memory reservations to avoid resource contention, which can result in performance degradation. Running a virtual appliance on every host in the cluster reduces the overall amount of compute resources available to run regular virtual machine workloads. Consolidation ratios will likely be lower and total cost of ownership rises when these storage virtual appliances are present and competing for the same resources as regular virtual machine workloads.

Storage virtual appliances can also introduce additional latency, which negatively affects performance. This is due to the number of steps required to handle and replicate write operations as shown in the figure below.



vSAN Embedded in the vSphere Hypervisor

vSAN does not require the deployment of storage virtual appliances or the installation of a vSphere Installation Bundle (VIB) on every host in the cluster. vSAN is embedded in the vSphere kernel and typically consumes less than 10% of the compute resources on each host. vSAN does not compete with other virtual machines for resources and the IO path is shorter.



A shorter IO path and the absence of resource-intensive storage virtual appliances enables vSAN to provide extreme performance with minimal overhead. Higher consolidation ratios translate into lower total costs of ownership.

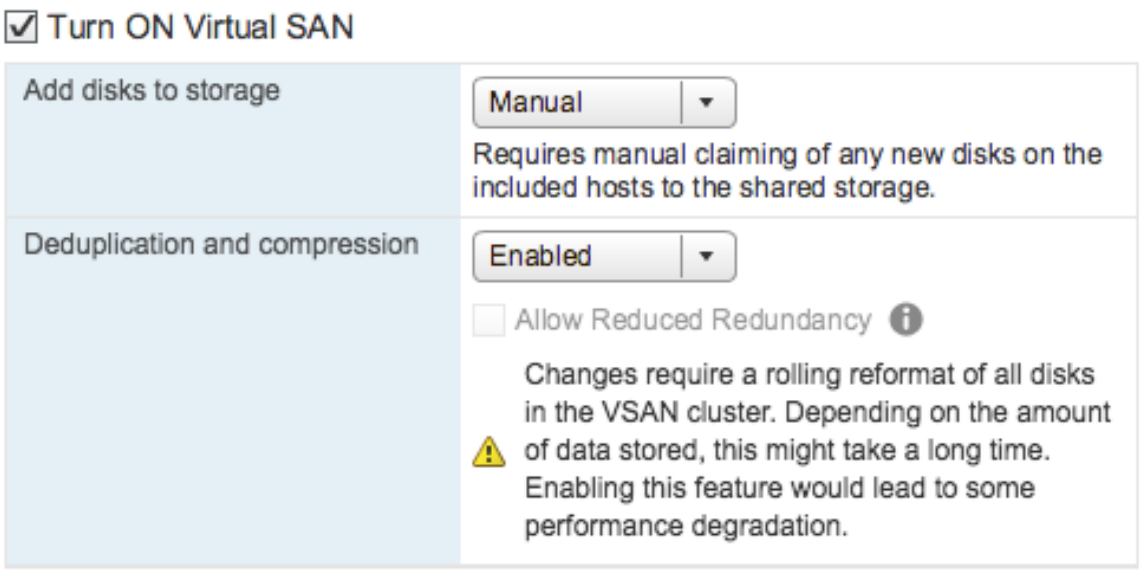
4. Simple Deployment and Operations

Enabling vSAN, health check, proactive tests, capacity and performance.

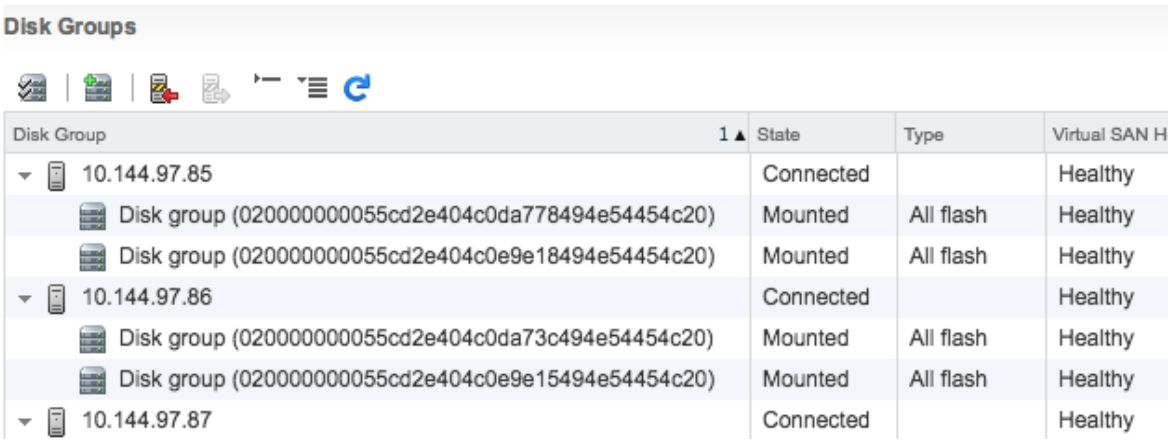
4.1 Enabling vSAN

Enabling vSAN

vSAN is built into vSphere. vSAN is enabled with just a few mouse clicks. There is no requirement to install additional software and/or deploy virtual storage appliances to every host in the cluster. Simply click the Enable vSAN checkbox to start the process. Deduplication and compression can also be enabled at that time.



The next step is claiming local storage devices in each host for the vSAN cache and capacity tiers. One or more disk groups are created in each host. Each disk group contains one cache device (flash) and one or more capacity devices (flash or magnetic). vSAN pools these local storage device together to create a pool of shared storage.



The process of enabling vSAN takes only a better of minutes. This is a tribute to the simplicity of vSAN - especially when you compare it to other enterprise-class hyper-converged and traditional storage systems, which typically take much longer to set up.

4.2 Health Check

Health Service

vSAN 6.2 features a comprehensive health service that actively tests and monitors a number of items such as hardware compatibility, network connectivity, cluster health, and capacity consumption. The health service is enabled by default and configured to check the health of the vSAN environment every 60 minutes.

Virtual SAN Health (Last checked: Today at 7:12 PM)

Test Result	Test Name
Passed	Hardware compatibility
Passed	Network
Passed	Physical disk
Passed	Data
Passed	Cluster
Passed	Limits
Passed	Performance service

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

Passed	Network
Passed	All hosts have a Virtual SAN vmknic configured
Passed	All hosts have matching multicast settings
Passed	All hosts have matching subnets
Passed	Basic (unicast) connectivity check (normal ping)
Passed	Hosts disconnected from VC
Passed	Hosts with connectivity issues
Passed	Hosts with Virtual SAN disabled
Passed	MTU check (ping with large packet size)
Passed	Multicast assessment based on other checks
Passed	Unexpected Virtual SAN cluster members
Passed	Virtual SAN cluster partition

If an issue is detected, 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 (HCL) will trigger a warning. In addition to providing details about the warning, vSAN Health also has an"Ask VMware" button, which brings up the relevant VMware Knowledge Base article.

SCSI Controller on Virtual SAN HCL Ask VMware

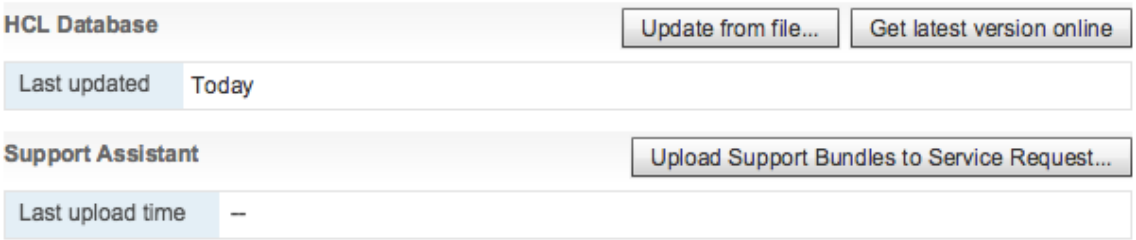
Checks if the controller is compatible with the VMware Compatibility Guide

Controller List

Host	Device	Display Name	On HCL	PCI ID
10.144.97.88	vmhba2	LSI Logic / Symbios Logic LSI2008	Warning	1000,00...
10.144.97.87	vmhba2	LSI Logic / Symbios Logic LSI2008	Warning	1000,00...

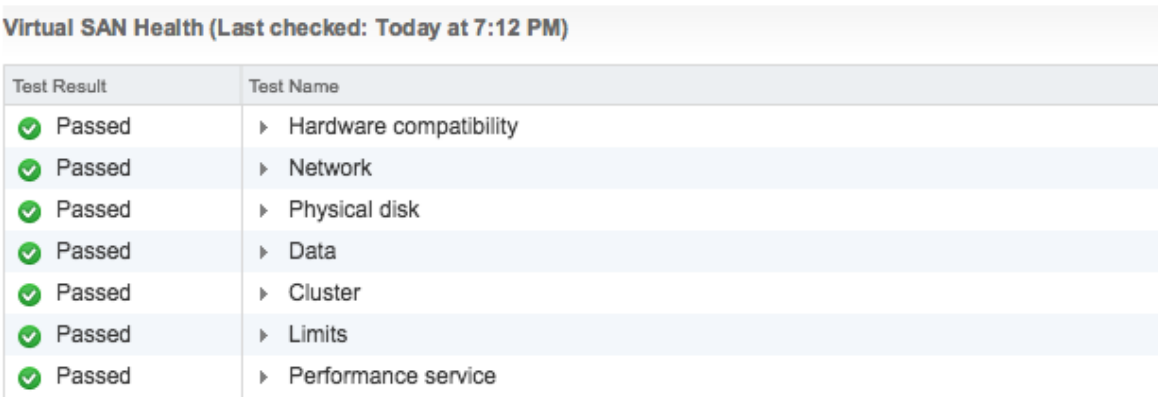
vSphere and vSAN support a wide variety of hardware configurations. The list ofhardware components and corresponding drivers that are supported with vSAN can be found in the VMware Compatibility Guide. It is very important to use only hardware, firmware, and drivers found in this guide to ensure stability and performance. The list of certified hardware, firmware, and drivers is contained in a hardware compatibility list (HCL) database. vSAN makes it 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 to enable offline updates.

If an issue does arise that requires the assistance of VMware Support, it is 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.



Health Service

vSAN 6.2 features a comprehensive health service that actively tests and monitors a number of items such as hardware compatibility, network connectivity, cluster health, and capacity consumption. The health service is enabled by default and configured to check the health of the vSAN environment every 60 minutes.



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

Passed	Network
Passed	All hosts have a Virtual SAN vmknic configured
Passed	All hosts have matching multicast settings
Passed	All hosts have matching subnets
Passed	Basic (unicast) connectivity check (normal ping)
Passed	Hosts disconnected from VC
Passed	Hosts with connectivity issues
Passed	Hosts with Virtual SAN disabled
Passed	MTU check (ping with large packet size)
Passed	Multicast assessment based on other checks
Passed	Unexpected Virtual SAN cluster members
Passed	Virtual SAN cluster partition

If an issue is detected, 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 (HCL) will trigger a warning. In addition to providing details about the warning, vSAN Health also has an"Ask VMware" button, which brings up the relevant VMware Knowledge Base article.

SCSI Controller on Virtual SAN HCL Ask VMware

Checks if the controller is compatible with the VMware Compatibility Guide

Controller List

Host	Device	Display Name	On HCL	PCI ID
10.144.97.88	vmhba2	LSI Logic / Symbios Logic LSI2008	Warning	1000,00...
10.144.97.87	vmhba2	LSI Logic / Symbios Logic LSI2008	Warning	1000,00...

vSphere and vSAN support a wide variety of hardware configurations. The list ofhardware 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 ensure stability and performance. The list of certified hardware, firmware, and drivers is contained in a hardware compatibility list (HCL) database. vSAN makes it 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 to enable offline updates.

If an issue does arise that requires the assistance of VMware Support, it is 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.

HCL Database Update from file... Get latest version online

Last updated Today

Support Assistant Upload Support Bundles to Service Request...

Last upload time –

Figure x. vSAN HCL Database and Support Assistant

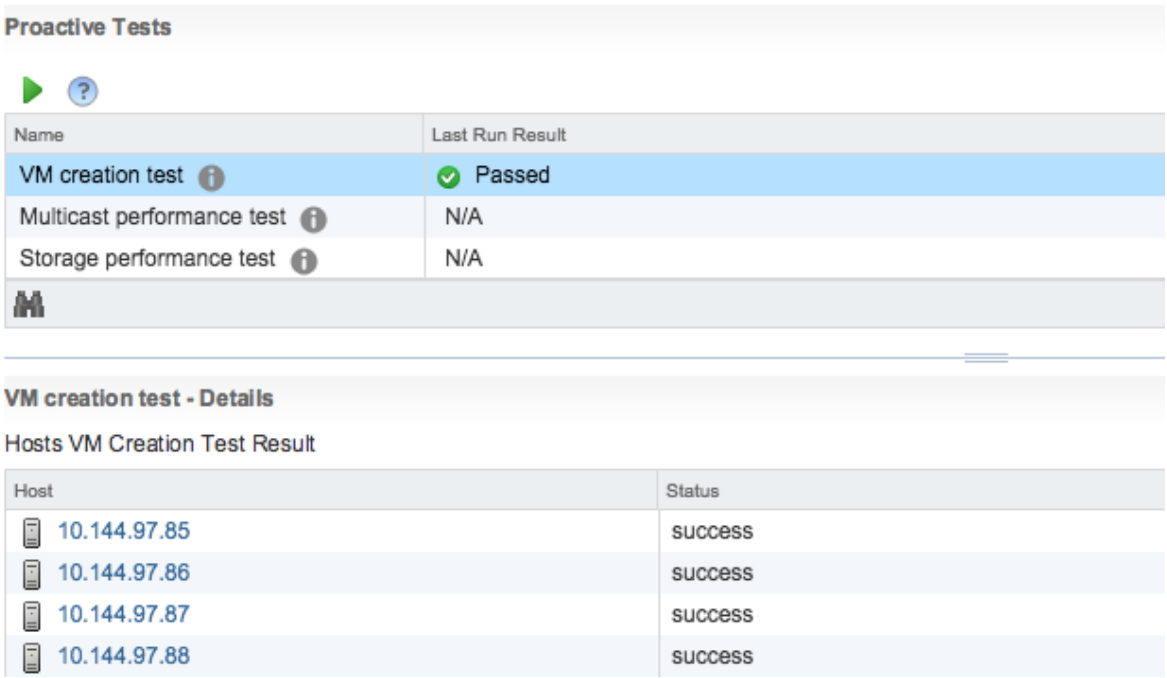
4.3 Proactive Tests

Proactive Tests

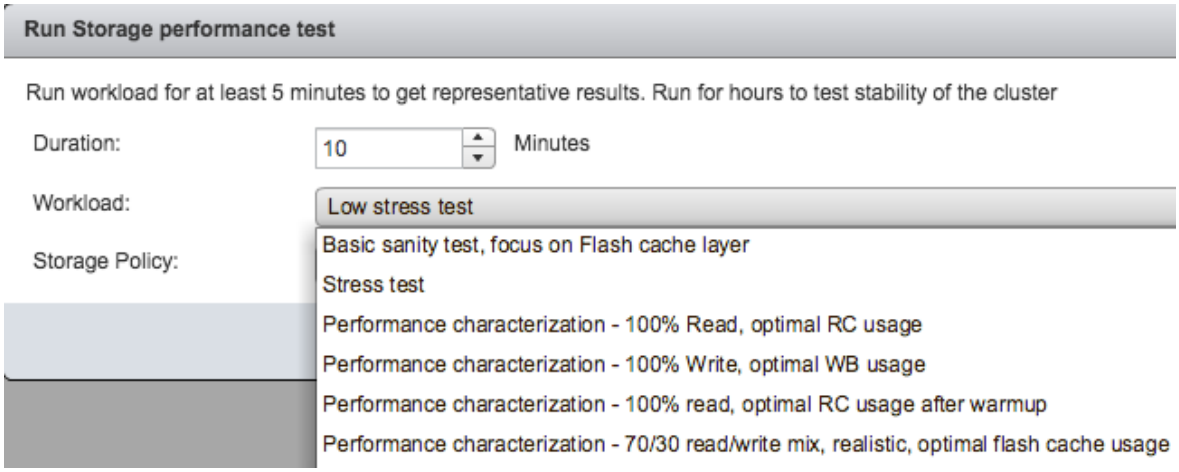
vSAN proactive tests enable administrators verify vSAN configuration, stability, and performance to minimize risk and confirm that the datastore is ready for production use. Three proactive tests are available:

- VM creation test
- Multicast performance test
- Storage performance test.

The VM creation test creates and deletes a small virtual machine on each host confirming basic functionality. The multicast test verifies multicast is working properly on each host and performance meets vSAN requirements. The following figure shows the results of a VM creation test.



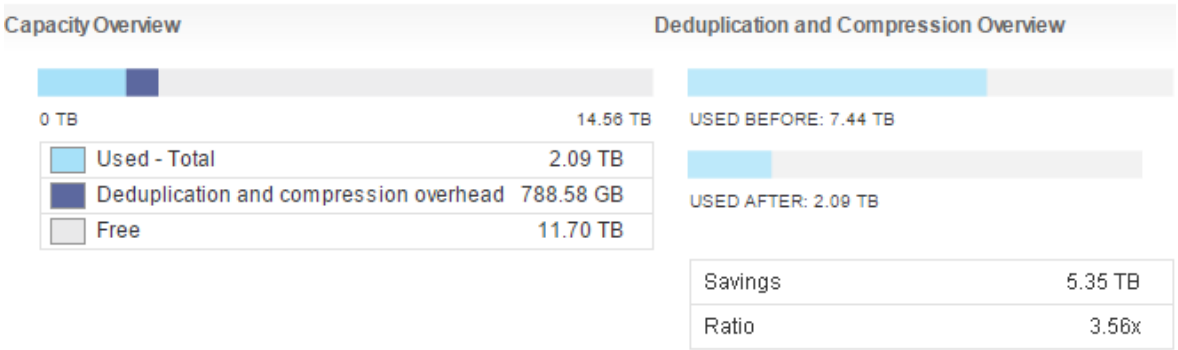
The storage performance test is used to check the stability of the vSAN cluster under heavy I/O load. There are a number of workload profiles that can be selected for this test as shown below. Keep in mind the storage performance test can affect other workloads and tasks. This test is intended to run before production virtual machine workloads are provisioned on vSAN.



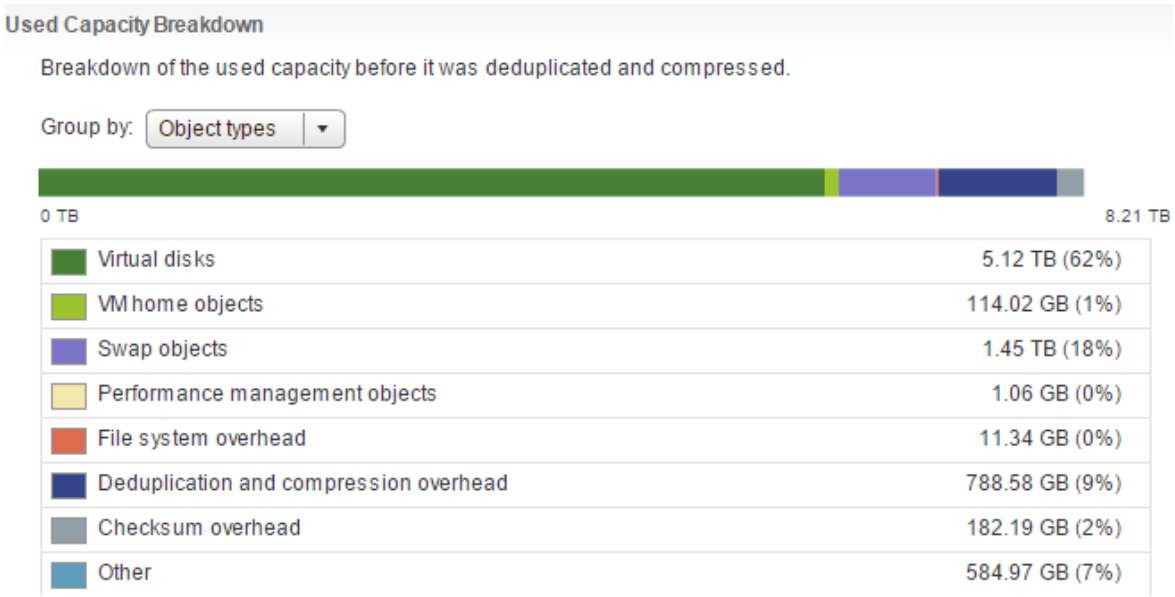
4.4 Capacity Reporting

Capacity Reporting

Capacity overviews are available in the vSAN user interface making it easy for administrators to see used and free space at a glance. Deduplication and compression information is also displayed.



Information is also available showing how much capacity various object types are consuming. Note that percentages are of used capacity, not of total capacity.



This list provides more details on the object types in the Used Capacity Breakdown chart:

- Virtual disks: Virtual disk consumption before deduplication and compression
- VM home objects: VM home object consumption before deduplication and compression
- Swap objects: Capacity utilized by virtual machine swap files
- Performance management objects: When the vSAN performance service is enabled, this is the amount of capacity used to store the performance data
- File system overhead: Capacity required by the vSAN file system metadata
- Deduplication and compression overhead: deduplication and compression metadata, such as hash, translation, and allocation maps
- Checksum overhead: Capacity used to store checksum information
- Other: Virtual machine templates, unregistered virtual machines, ISO files, and so on that are consuming vSAN capacity

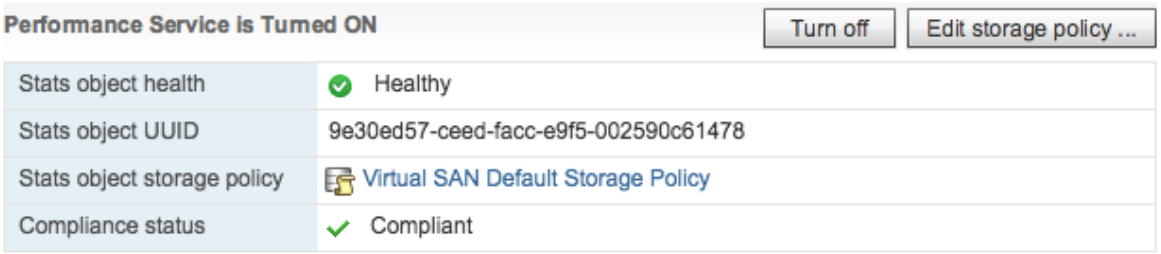
4.5 Performance Reporting

Performance Service

A healthy vSAN environment is one that is performing well. vSAN includes a number of graphs and data points that provide performance information at the cluster, host, virtual machine, and virtual disk levels. Time Range can be modified to show information from the last 1-24 hours or a custom date and time range.

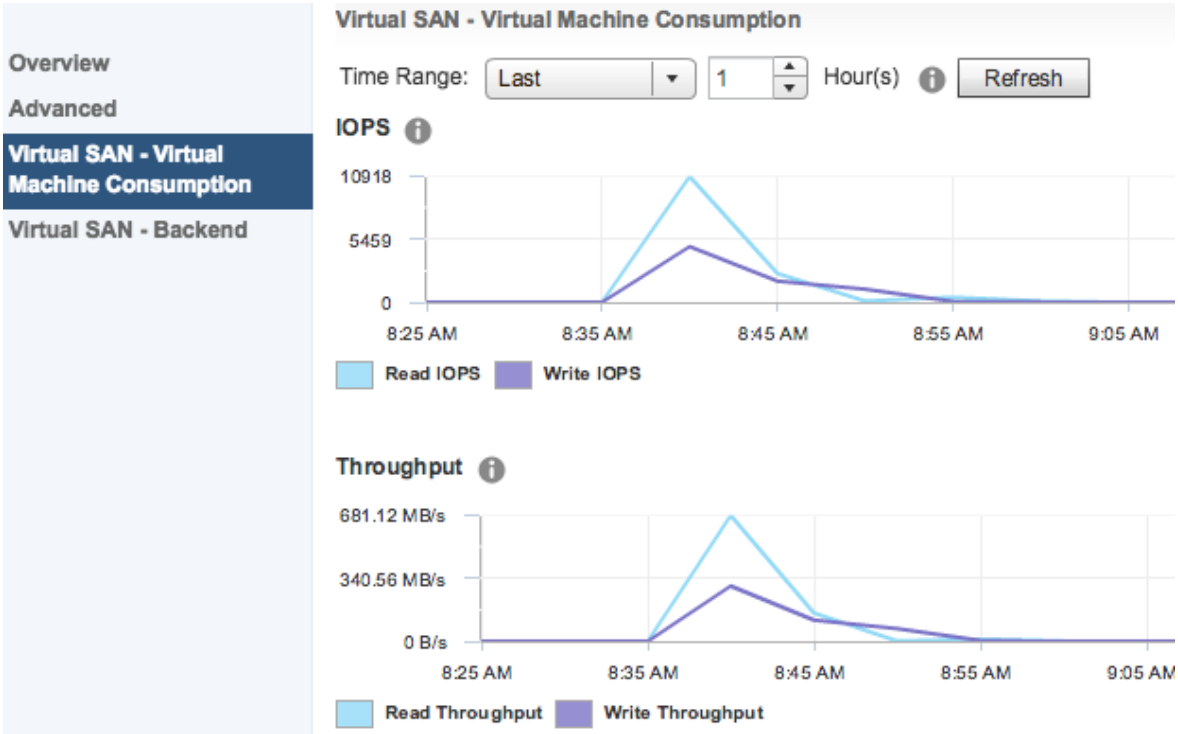
The performance service is enabled at the cluster level. The performance history database is stored as a vSAN object independent of vCenter. A storage policy is assigned to the object to control space consumption, availability, and performance of that object. If the object becomes unavailable, performance history for the cluster cannot be viewed until access to the object is restored.

The performance service is turned off by default. A few mouse clicks are all that is needed to enable the service.



Cluster Metrics

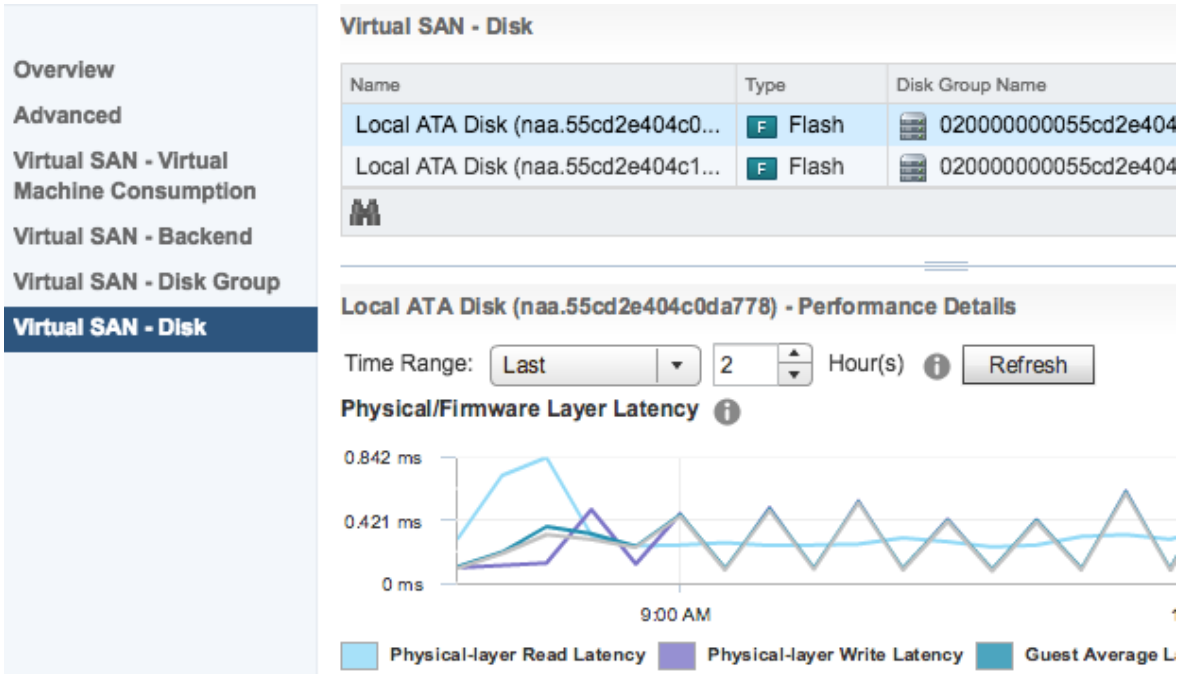
At the cluster level, the performance monitoring service shows performance metrics for virtual machines running on vSAN. These metrics provide quick visibility to the entire vSAN cluster. A number of graphs are included such as read and write IOPs, read and write throughput, read and write latency, and congestion.



Backend consumption stems from activities such as metadata updates, component builds, etc. For example, a virtual machine with a number of failures to tolerate set to 1 with a failure tolerance method of RAID-1 (Mirroring). For every write IO to a virtual disk, two are produced on the backend - a write to each component replica residing on two different hosts.

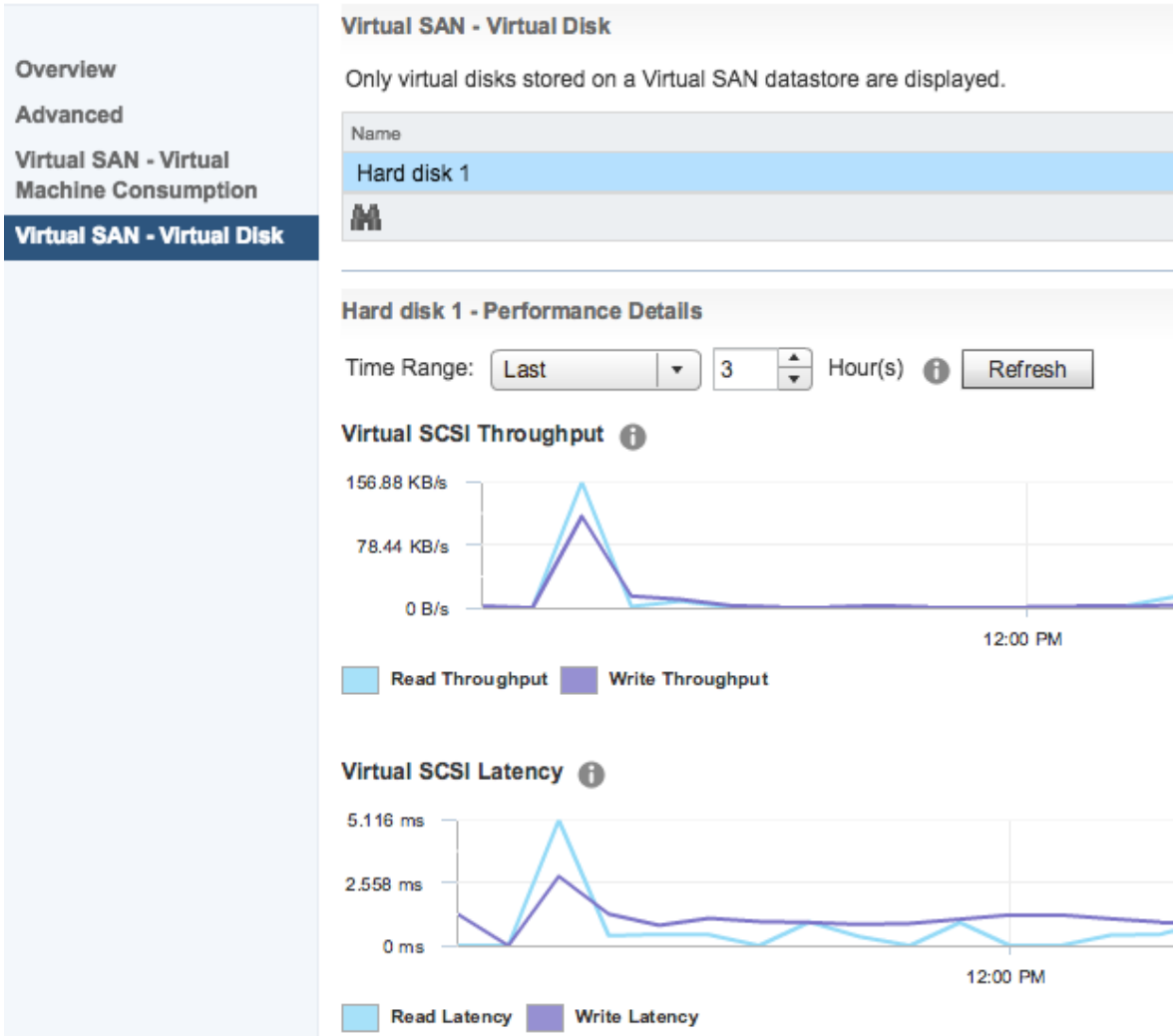
Host Metrics

In addition to virtual machine consumption and backend performance metrics, disk group and individual disk performance information is available at the host level. Seeing metrics for individual disks eases the process of troubleshooting issues such as failed storage devices.



Virtual Machine Metrics

vSAN performance information for individual virtual machines and virtual disks. Metrics include IOPS, throughput, and latency. The figure below shows virtual disk-level Virtual SCSI throughput and latencies for reads and writes.

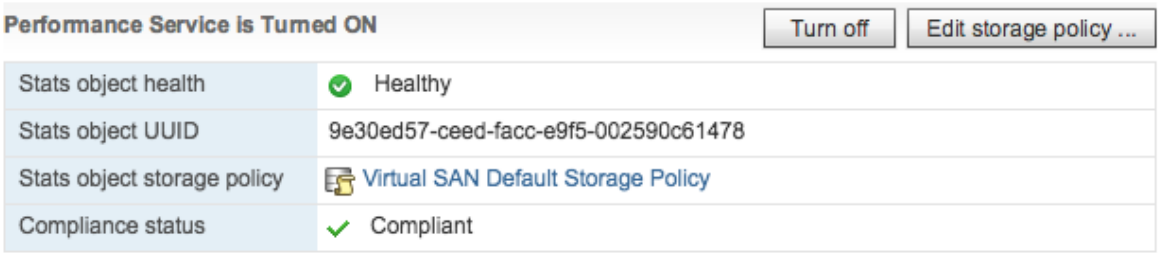


Performance Service

A healthy vSAN environment is one that is performing well. vSAN includes a number of graphs and data points that provide performance information at the cluster, host, virtual machine, and virtual disk levels. Time Range can be modified to show information from the last 1-24 hours or a custom date and time range.

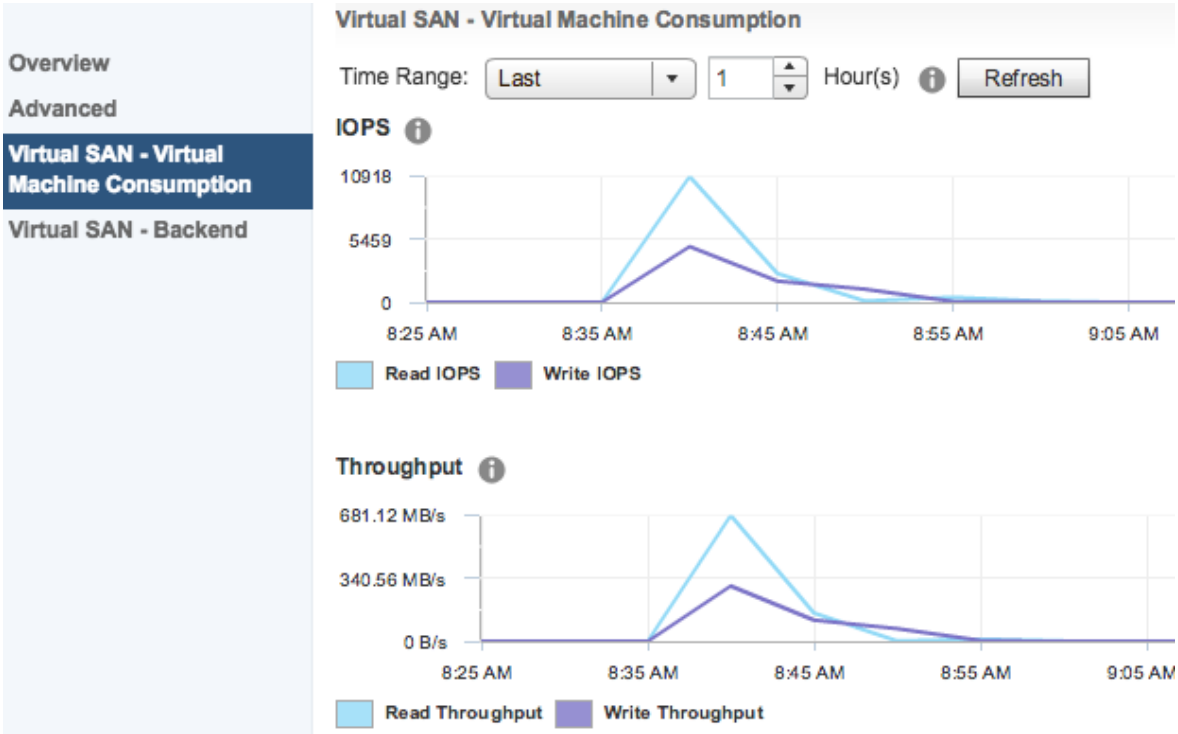
The performance service is enabled at the cluster level. The performance history database is stored as a vSAN object independent of vCenter. A storage policy is assigned to the object to control space consumption, availability, and performance of that object. If the object becomes unavailable, performance history for the cluster cannot be viewed until access to the object is restored.

The performance service is turned off by default. A few mouse clicks are all that is needed to enable the service.



Cluster Metrics

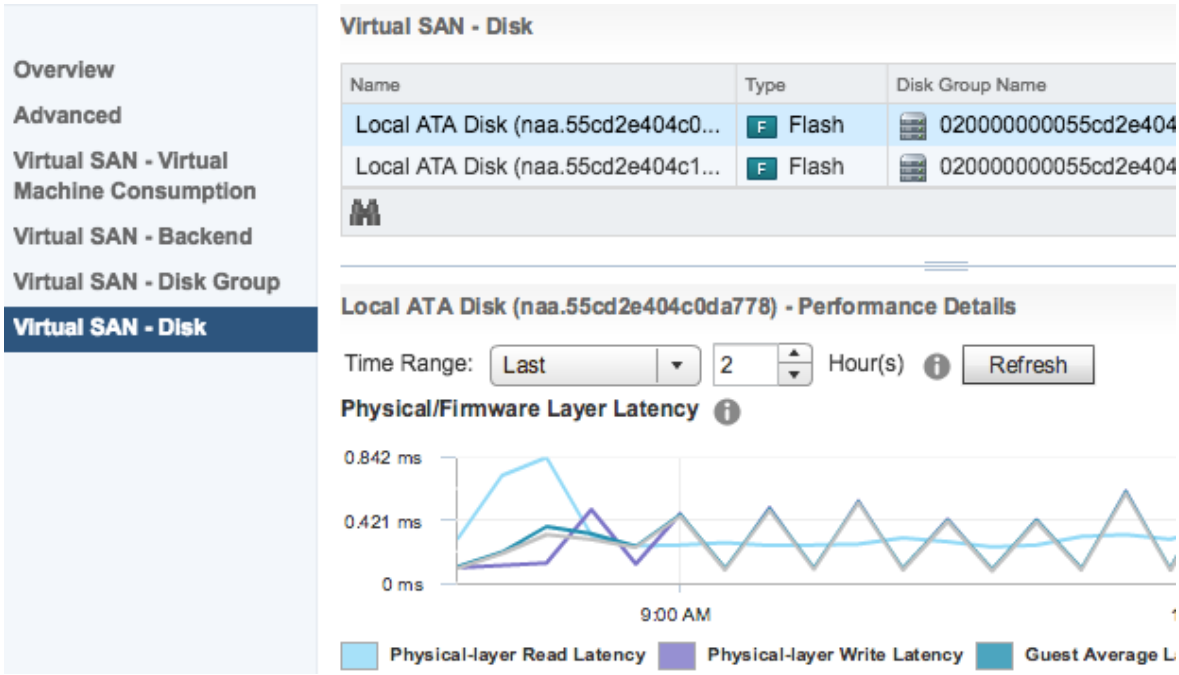
At the cluster level, the performance monitoring service shows performance metrics for virtual machines running on vSAN. These metrics provide quick visibility to the entire vSAN cluster. A number of graphs are included such as read and write IOPs, read and write throughput, read and write latency, and congestion.



Backend consumption stems from activities such as metadata updates, component builds, etc. For example, a virtual machine with a number of failures to tolerate set to 1 with a failure tolerance method of RAID-1 (Mirroring). For every write IO to a virtual disk, two are produced on the backend - a write to each component replica residing on two different hosts.

Host Metrics

In addition to virtual machine consumption and backend performance metrics, disk group and individual disk performance information is available at the host level. Seeing metrics for individual disks eases the process of troubleshooting issues such as failed storage devices.



Virtual Machine Metrics

vSAN performance information for individual virtual machines and virtual disks. Metrics include IOPS, throughput, and latency. The figure below shows virtual disk-level Virtual SCSI throughput and latencies for reads and writes.

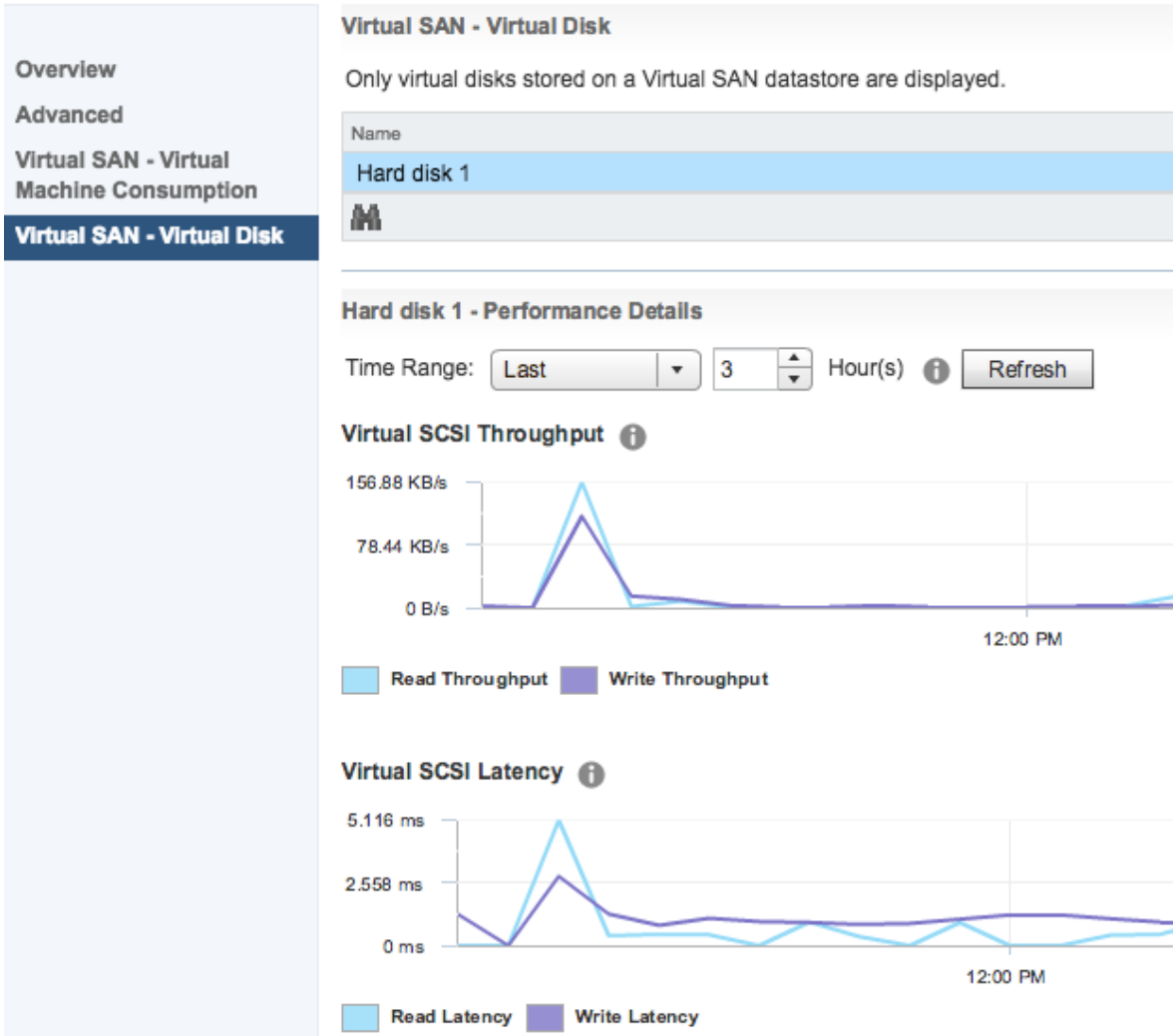


Figure x. Virtual Disk Performance

5. Policy Based Management and Automation

Creating and modifying a storage policy, component placement, and automation (PowerCLI, Python, APIs)

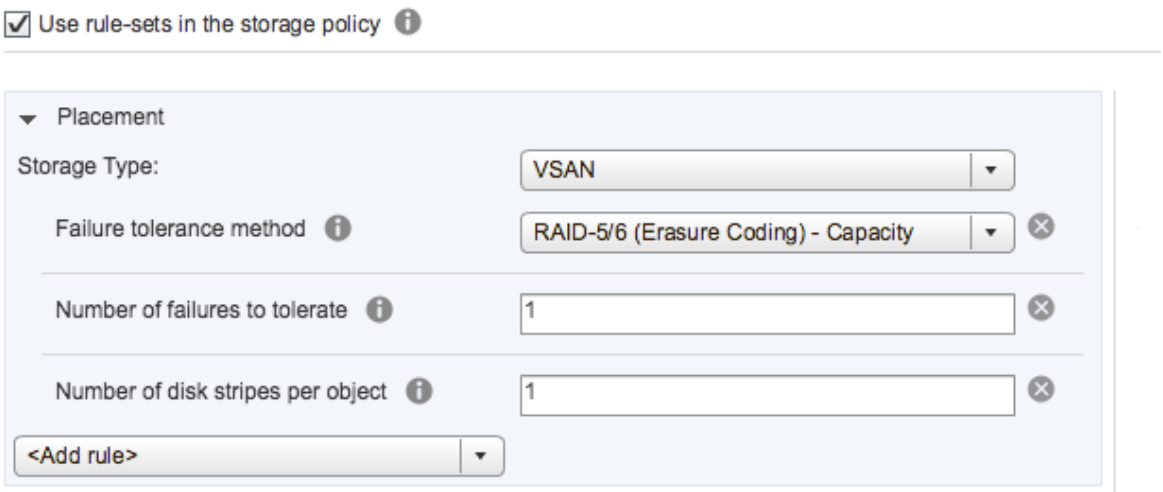
5.1 Storage Policy Based Management

Traditional Storage Management

Traditional storage models utilize LUNs or volumes. A LUN or a volume is commonly configured with a specific disk configuration such as RAID to provide a specific level of performance and availability. The challenge with this model is each LUN or volume is confined to providing only one level of service regardless of the workloads that it contains. This leads to the provisioning of numerous LUNs or volumes in an attempt to provide the right levels or storage services to each workload. Maintaining a large number of LUNs or volumes leads to management complexity. Deployment and management of workloads and storage in traditional storage environments can be time consuming and error prone.

Storage Policy Based Management

Storage Policy Based Management (SPBM) enables precise control of the storage services. Similar to other storage solutions, vSAN provides services such as availability level, striping for performance, and the ability to limit IOPS. Policies that contain one or more rules are created using the vSphere Web Client.



These policies are assigned to virtual machines and individual objects such as a virtual disk. Storage policies can easily be changed and/or reassigned if application requirements change. These changes are performed with no downtime and without the need to migrate (Storage vMotion) virtual machines from one LUN or volume to another. This approach makes it possible to assign and modify service levels based on specific application needs even though the virtual machines reside on the same datastore.

5.2 Automation

Automation

vSAN features an extensive management API and multiple software development kits (SDKs) to provide IT organizations options for rapid provisioning and automation. Administrators and developers can orchestrate all aspects of installation, configuration, lifecycle management, monitoring, and troubleshooting of vSAN environments. This is especially useful in large environments and geographically disbursed organizations to speed up deployment times, reduce operational costs, maintain standards, and orchestrate common workflows.

SDKs are available for several programming languages including .NET, Perl, and Python. They are available for download from VMware Developer Center and include libraries, documentation, and code samples. For example, this Python script can be used to generate vSAN capacity information: vSAN Capacity - Total and Free from vCenter

vSAN APIs can also be accessed through vSphere PowerCLI cmdlets. IT administrators can automate common tasks such as assigning storage policies and checking storage policy compliance. Consider a repeatable task such as deploying or upgrading two-node vSAN clusters at 100 retail store locations. Performing each one manually would take a considerable amount of time. There is also a higher risk of error leading to non-standard configurations and possibly downtime. vSphere PowerCLI can instead be used to ensure all of the vSAN clusters are deployed with the same configuration. Lifecycle management, such as applying patches and upgrades, is also much easier when these tasks are automated.

This video demonstrates a number of operations from creating a new cluster to configuring vSAN using just a few lines of vSphere PowerCLI code:

vSphere PowerCLI: Creating a Cluster and Configuring vSAN

6. Enterprise Availability

Resiliency against disk, host, rack, and entire site failures.

6.1 Objects and Component Placement

Objects

vSAN (VSAN) is an object datastore with a mostly flat hierarchy of objects and containers (folders). Items that make up a virtual machine (VM) are represented by objects. These are the most common object types you will find on a VSAN datastore:

- VM Home, which contains virtual machine configuration files and logs such as the VMX and NVRAM files
- VM Swap
- Virtual Disk (VMDK)
- Delta Disk (snapshot)
- Memory Delta, which is present when the checkbox to snapshot a VM's memory is checked

There are a few other objects that might be found on a VSAN datastore such as the VSAN performance service database and VMDKs that belong to iSCSI targets.

Components

Each object consists of one or more components. The number of components that make up an object depends primarily on a couple things: The size of the objects and the storage policy assigned to the object. The maximum size of a component is 255GB. If an object is larger than 255GB, it is split up into multiple components. The image below shows a 600GB virtual disk split up into three components.

Type	Component State	Host	Cache Disk Name
RAID 0			
Component	Active	10.144.97.88	Local ATA Disk (naa.55cd2e404c0da
Component	Active	10.144.97.88	Local ATA Disk (naa.55cd2e404c0da
Component	Active	10.144.97.88	Local ATA Disk (naa.55cd2e404c0da

In most cases, a VM will have a storage policy assigned that contains availability rules such as Number of Failures to Tolerate and Failure Tolerance Method. These rules will also affect the number of components that make up an object. As an example, let's take that same 600GB virtual disk and apply the vSAN Default Storage Policy, which uses the RAID-1 mirroring failure tolerance method and has the number of failures to tolerate set to one. The 600GB object with three components will be mirrored on another host. This configuration provides two full copies of the data distributed across two hosts so that the loss of a disk or an entire host can be tolerated. Below is an image showing the six components (three on each host). A seventh component, the witness, is created by VSAN to "break the tie" and achieve quorum in the event of a network partition between the hosts. The witness object is place on a third host.

Type	Component State	Host	Cache Disk Name
RAID 1			
RAID 0			
Component	Active	10.144.97.88	Local ATA Disk (naa.55cd2e404c0da
Component	Active	10.144.97.88	Local ATA Disk (naa.55cd2e404c0da
Component	Active	10.144.97.88	Local ATA Disk (naa.55cd2e404c0da
RAID 0			
Component	Active	10.144.97.87	Local ATA Disk (naa.55cd2e404c0da
Component	Active	10.144.97.87	Local ATA Disk (naa.55cd2e404c0da
Component	Active	10.144.97.87	Local ATA Disk (naa.55cd2e404c0da
Witness	Active	10.144.97.85	Local ATA Disk (naa.55cd2e404c0e9

In this last example of component placement, we take the same 600GB virtual disk and apply a storage policy with RAID-5 erasure coding (Failures to Tolerate = 1). The object now consists of four components - three data components and a parity component - distributed across the four hosts in the cluster. If disk or host containing any one of these components is offline, the data is still accessible. If one of these components are permanently lost, vSAN can rebuild the lost data or parity component from the other three surviving components.

Type	Component State	Host 1 ▲	Cache Disk Name
RAID 5			
Component	Active	10.144.97.85	Local ATA Disk (naa.55cd2e404c0e9
Component	Active	10.144.97.86	Local ATA Disk (naa.55cd2e404c0da
Component	Active	10.144.97.87	Local ATA Disk (naa.55cd2e404c0da
Component	Active	10.144.97.88	Local ATA Disk (naa.55cd2e404c0da

vSAN requires a minimum number of hosts depending on the failure tolerance method and number of failures to tolerate (FTT) configurations. For example, a minimum of three hosts are needed for FTT=1 with RAID-1 mirroring. A minimum of four hosts are required for FTT=1 with RAID-5 erasure coding. More implementation details and recommendations can be found in the vSAN Design and Sizing Guide.

6.2 Rack Awareness

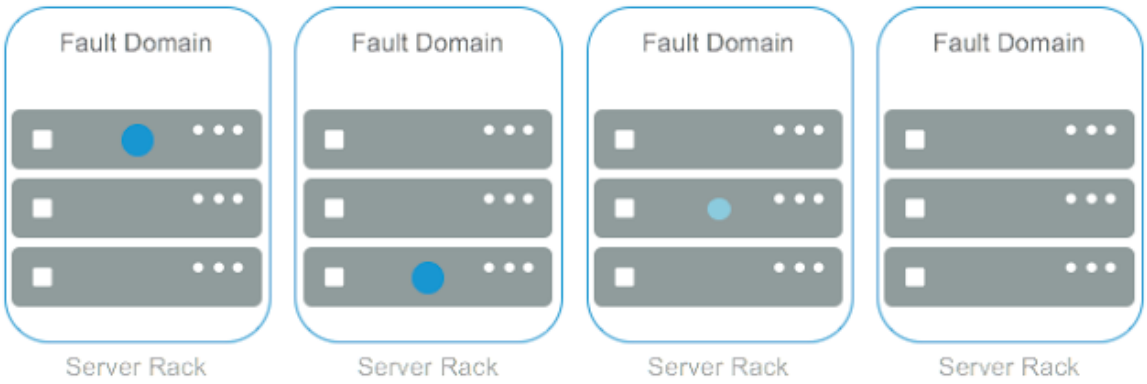
Fault Domains

"Fault domain" is a term that comes up fairly often in availability discussions. In IT, a fault domain usually refers to a group of servers, storage, and/or networking components that would be impacted collectively by an outage. A very common example of this is a server rack. If a top-of-rack (TOR) switch or the power distribution unit (PDU) for a server rack would fail, it would take all of the servers in that rack offline even though the servers themselves are functioning properly. That server rack is considered a fault domain.

Rack Awareness

While the failure of a disk or entire host can be tolerated, what if all of these servers are in the same rack and the TOR switch goes offline? Answer: All hosts are isolated from each other and none of the

objects are accessible. To mitigate this risk, the servers in a vSAN cluster should be spread across server racks and fault domains must be configured in the vSAN user interface. After fault domains are configured, vSAN will redistribute the components across server racks to eliminate the risk of a rack failure taking multiple objects offline. This feature is commonly referred to as "Rack Awareness". The diagram below shows what this might look like with a 12-node vSAN cluster spread across four server racks.



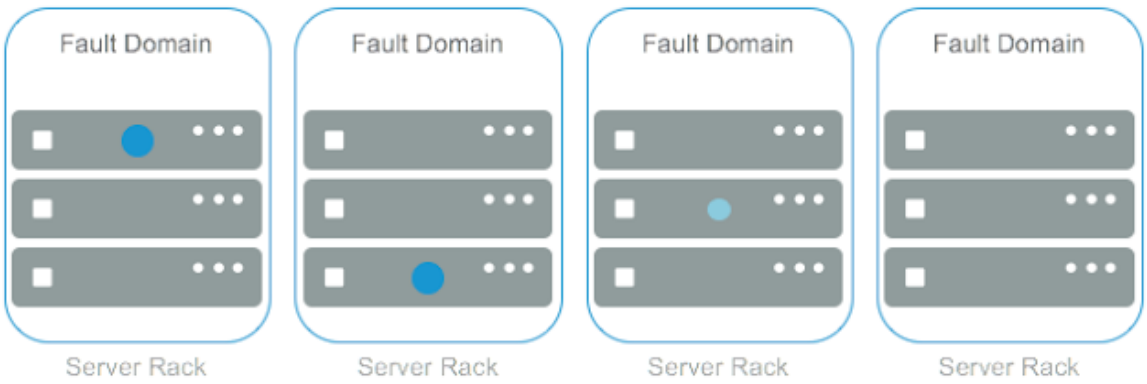
This feature could also be used to create a highly available vSAN acr

Fault Domains

"Fault domain" is a term that comes up fairly often in availability discussions. In IT, a fault domain usually refers to a group of servers, storage, and/or networking components that would be impacted collectively by an outage. A very common example of this is a server rack. If a top-of-rack (TOR) switch or the power distribution unit (PDU) for a server rack would fail, it would take all of the servers in that rack offline even though the servers themselves are functioning properly. That server rack is considered a fault domain.

Rack Awareness

While the failure of a disk or entire host can be tolerated, what if all of these servers are in the same rack and the TOR switch goes offline? Answer: All hosts are isolated from each other and none of the objects are accessible. To mitigate this risk, the servers in a vSAN cluster should be spread across server racks and fault domains must be configured in the vSAN user interface. After fault domains are configured, vSAN will redistribute the components across server racks to eliminate the risk of a rack failure taking multiple objects offline. This feature is commonly referred to as "Rack Awareness". The diagram below shows what this might look like with a 12-node vSAN cluster spread across four server racks.



Configuring vSAN fault domains is quite simple as demonstrated in this video (no audio): vSAN Fault Domains

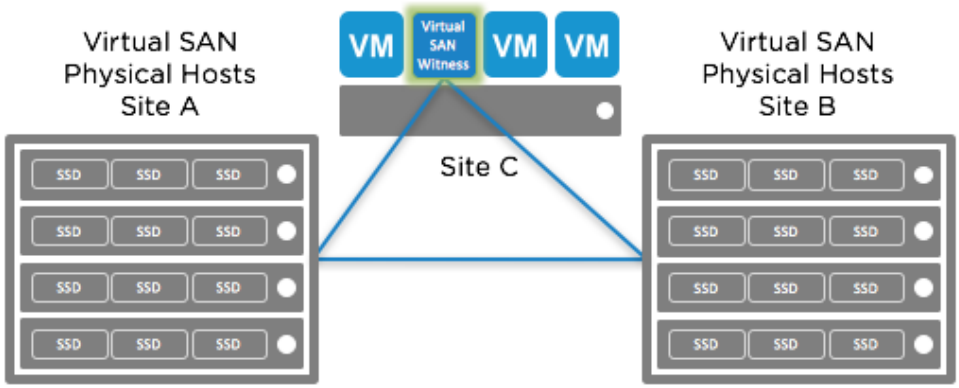
6.3 Stretched Clusters

Stretched Clusters

vSAN Stretched Clusters provide organizations with the capability to deploy a vSAN cluster across two locations. These locations can be opposite sides of the same data center, two buildings on the same campus, or geographically disbursed between two cities. It is important to note that this technology does have bandwidth and latency requirements as detailed in the vSAN Stretched Cluster Guide and 2-Node Guide.

A stretched cluster provides resiliency against larger scale outages and disasters by keeping two copies of the data - one at each location. If a failure occurs at either location, all data is available at the other location. vSphere HA restarts any virtual machines affected by the outage using the copy of the data at the surviving location. In the case of disaster avoidance such as an impending storm or rising flood waters, virtual machines can be migrated from one location to the other with no downtime using vMotion.

Since vSAN is a clustering technology, a witness is required to achieve quorum in the case of a "split-brain" scenario where the two locations lose network connectivity. A vSAN witness is simply a virtual machine running ESXi. The witness is deployed at a third location (separate from the two primary data locations) to avoid being affected by any issues that could occur at either of the main sites.



The witness does not store data such as virtual disk (VMDK) objects. Only witness objects are stored in the witness virtual appliance. If any one of the three sites goes offline, there is still more than 50% of each object's components online to achieve quorum and maintain availability.

7. Space Efficiency

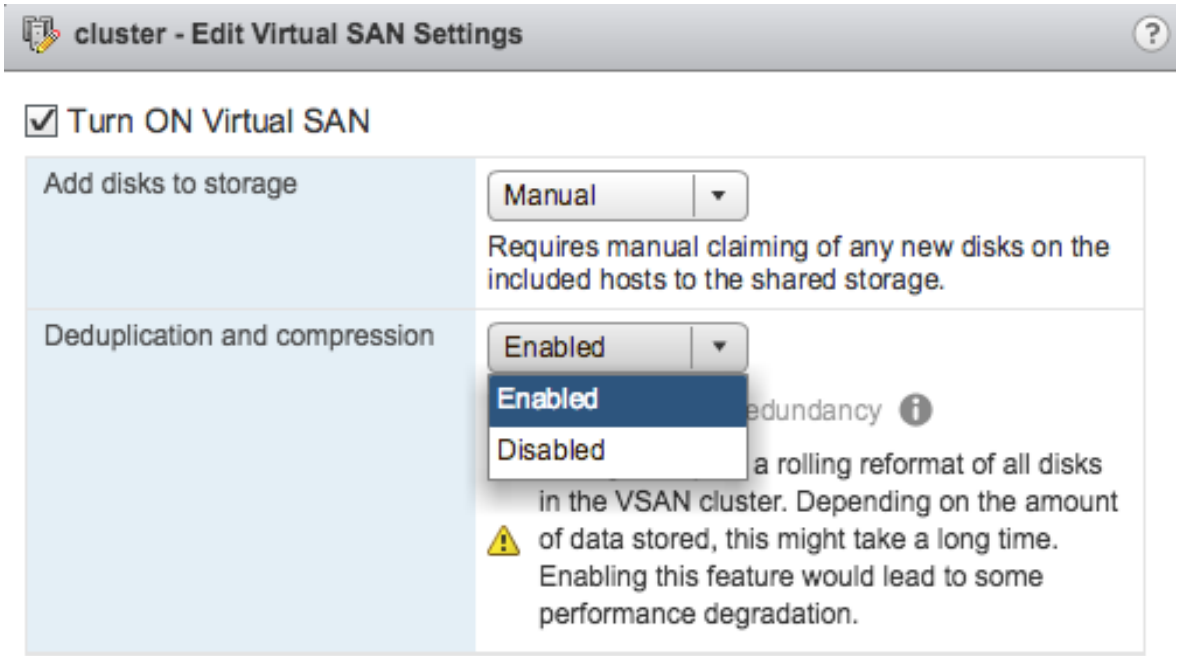
Maximize usable capacity with space efficiency features such as deduplication, compression, and erasure coding.

7.1 Deduplication and Compression

Deduplication and Compression

Enabling deduplication and compression can reduce the amount of physical storage consumed by as much as 7x, resulting in a lower total cost of ownership (TCO). Environments with highly-redundant data such as full-clone virtual desktops and homogenous server operating systems will naturally benefit the most from deduplication. Likewise, compression will offer more favorable results with data that compresses well such as text, bitmap, and program files. Data that is already compressed such as certain graphics formats and video files, as well as files that are encrypted, will yield little or no reduction in storage consumption from compression. In other words, deduplication and compression results will vary based on the types of data stored in an all-flash vSANenvironment.

Deduplication and compression is a single cluster-wide setting that is disabled by default and can be enabled using a simple drop-down menu. Note that a rolling format of all disks in the vSAN cluster is required, which can take a considerable amount of time. However, this process does not incur virtual machine downtime and can be done online, usually during an upgrade. Deduplication and compression are enabled as a unit. It is not possible to enable deduplication or compression individually.



Deduplication and compression are implemented after write acknowledgement to minimize impact to performance. Deduplication occurs when data is de-staged from the cache tier to the capacity tier of an all-flash vSAN datastore. The deduplication algorithm utilizes a 4K-fixed block size and is performed within each disk group. 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 compression algorithm is applied after deduplication has occurred just before the data is written tothe capacity tier. Considering the additional compute resource and allocation map overhead ofcompression, vSAN will only store compressed data if a unique 4K block can be reduced to 2K orless. Otherwise, the block is written uncompressed to avoid the use of additional resources whencompressing and decompressing these blocks which would provide little benefit.

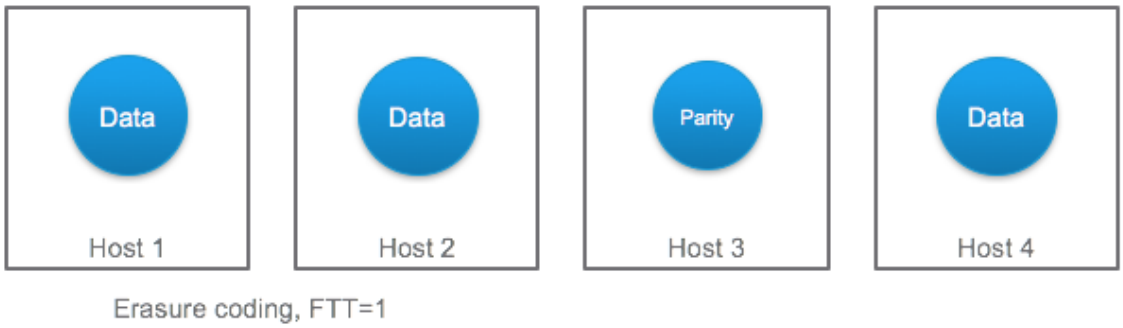
The processes of deduplication and compression on any storage platform incur overhead and potentially impact performance in terms of latency and maximum IOPS. vSAN is no exception.

However, considering deduplication and compression are only supported in all-flash vSAN configurations, these effects are predictable in the majority of use cases. The extreme performance and low latency of flash devices easily outweigh the additional resource requirements of deduplication and compression. The space efficiency generated by deduplication and compression lowers the cost-per-usable-GB of all-flash configurations.

7.2 RAID-5/6 Erasure Coding

RAID-5/6 Erasure Coding

RAID-5/6 erasure coding is a space efficiency feature optimized for all-flash configurations. Erasure coding provides the same levels of redundancy as mirroring, but with a reduced capacity requirement. In general, erasure coding is a method of taking data, breaking it into multiple pieces and spreading it across multiple devices, while adding parity data so it may be recreated in the event one of the pieces is corrupted or lost.



Erasure coding, FTT=1

Unlike deduplication and compression, which offer variable levels of space efficiency, erasure coding guarantees capacity reduction over a mirroring data protection method at the same failure tolerance level. As an example, let's consider a 100GB virtual disk. Surviving one disk or host failure requires 2 copies of data at 2x the capacity, i.e., 200GB. If RAID-5 erasure coding is used to protect the object, the 100GB virtual disk will consume 133GB of raw capacity - a 33% reduction in consumed capacity versus RAID-1 mirroring.

While erasure coding provides significant capacity savings over mirroring, understand that erasure coding requires additional processing overhead. This is common among any storage platform today. Erasure coding is only supported in all-flash vSAN configurations. Therefore, performance impact is negligible in most use cases due to the inherent performance of flash devices. Also note that RAID-5 erasure coding (FTT=1) requires a minimum of four hosts and RAID-6 erasure coding requires a minimum of six hosts.

8. Quality of Service

Quality of Service

8.1 IOPS Limits

Limiting IOPS

vSAN has the ability to limit the number of IOPS a virtual machine or virtual disk generates. There are situations where it is 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 impact other workloads or tenants in the same environment.

An example of a possible noisy neighbor scenario is month-end reporting. Management requests delivery of these reports on the second day of each month so the reports are generated on the first day of each month. The virtual machines that run the reporting application and database are dormant most of the time. Running the reports take just a few hours, but this generates very high levels of storage I/O. The performance of other workloads in the environment are impacted while the reports are running. To remedy this issue, an administrator creates a storage policy with an IOPS limit rule and assigns the policy to the virtual machines running the reporting application and database. The IOPS limit eliminates the performance impact to the other virtual machines. The reports take longer, but they are still finished in plenty of time for delivery the next day.

Keep in mind storage policies can be dynamically created, modified, and assigned to virtual machines. If an IOPS limit is proving to be too restrictive, simply modify the existing policy or create a new policy with a different IOPS limit and assign it to the virtual machines. The new or updated policy will take effect just moments after the change is made.

9. Summary

Summary

9.1 Summary

Summary

vSAN is optimized for modern all-flash storage with space efficiency features such as deduplication, compression, and erasure coding that lower TCO while delivering incredible performance. Hardware expenditures are further reduced for 2-node cluster configurations using directly connected crossover cables. This lowers network switch costs, reduces complexity, and improves reliability especially for use cases such as remote offices. The vSAN iSCSI target service enables physical servers and application cluster workloads to utilize a vSAN datastore. The performance and health services make it easy to verify vSAN configurations and closely monitor key metrics including 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 machines and per-virtual disk basis. All of these services are precisely managed using VM-centric storage policies. vSAN scales to 64 nodes per cluster with up to 150k IOPS per node using the latest hardware innovations such as NVMe. Deployment options include vSAN Ready Nodes, Dell EMC VxRail turnkey appliances, and build-your-own. All of these utilize components certified by VMware and the leading hardware OEMs to simplify hyper-converged infrastructure deployment. This provides organizations a vast number of options to run any app, any scale on an enterprise-class platform powered by vSAN.