vSAN Network Design

Update 1

VMware vSphere 8.0
VMware vSAN 8.0

vmware®

You can find the most up-to-date technical documentation on the VMware website at:

https://docs.vmware.com/

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

Contents

About vSAN Network Design

1

The *vSAN Network Design* guide describes network requirements, network design, and configuration practices for deploying a highly available and scalable vSAN cluster.

vSAN is a distributed storage solution. As with any distributed solution, the network is an important component of the design. For best results, you must adhere to the guidance provided in this document as improper networking hardware and designs can lead to unfavorable results.

At VMware, we value inclusion. To foster this principle within our customer, partner, and internal community, we create content using inclusive language.

Intended Audience

This guide is intended for anyone who is designing, deploying, and managing a vSAN cluster. The information in this guide is written for experienced network administrators who are familiar with network design and configuration, virtual machine management, and virtual data center operations. This guide also assumes familiarity with VMware vSphere, including VMware ESXi, vCenter Server, and the vSphere Client.

Related Documents

In addition to this guide, you can refer to the following guides to know more about vSAN networking:

- *vSAN Planning and Deployment Guide*, to know more about creating vSAN clusters
- *Administering VMware vSAN*, to configure a vSAN cluster and learn more about vSAN features
- *vSAN Monitoring and Troubleshooting Guide*, to monitor and troubleshoot vSAN clusters

Updated Information

2

This document is updated with each release of the product or when necessary.

This table provides the update history of *vSAN Network Design*.

Revision	Description
12 JUN 2023	Updated Network I/O Control Configuration Example, vSphere RDMA, and Chapter 11 Network I/O Control.
11 OCT 2022	Initial release.

What is vSAN Network

3

You can use vSAN to provision the shared storage within vSphere. vSAN aggregates local or direct-attached storage devices of a host cluster and creates a single storage pool shared across all hosts in the vSAN cluster.

vSAN is a distributed and shared storage solution that depends on a highly available, properly configured network for vSAN storage traffic. A high performing and available network is crucial to a successful vSAN deployment. This guide provides recommendations on how to design and configure a vSAN network.

vSAN has a distributed architecture that relies on a high-performing, scalable, and resilient network. All host nodes within a vSAN cluster communicate over the IP network. All the hosts must maintain IP unicast connectivity, so they can communicate over a Layer 2 or Layer 3 network. For more information on the unicast communication, see Chapter 5 Using Unicast in vSAN Network.

Releases prior to vSAN version 6.6 require IP multicast. If possible, always use the latest version of vSAN. For more information on multicast, see Chapter 14 Using Multicast in vSAN Network.

vSAN Networking Terms and Definitions

vSAN introduces specific terms and definitions that are important to understand. Before you get start designing your vSAN network, review the key vSAN networking terms and definitions.

Terms	Definitions
CLOM	The Cluster-Level Object Manager (CLOM) is responsible for ensuring that an object's configuration matches its storage policy. The CLOM checks whether enough disk groups are available to satisfy that policy. It decides where to place components and witnesses in a cluster.
CMMDS	The Cluster Monitoring, Membership, and Directory Service (CMMDS) is responsible for the recovery and maintenance of a cluster of networked node members. It manages the inventory of items such as host nodes, devices, and networks. It also stores metadata information, such as policies and RAID configuration for vSAN objects.

Terms	Definitions
DOM	The Distributed Object Manager (DOM) is responsible for creating the components and distributing them across the cluster. After a DOM object is created, one of the nodes (host) is nominated as the DOM owner for that object. This host handles all IOPS to that DOM object by locating the respective child components across the cluster and redirecting the I/O to respective components over the vSAN network. DOM objects include vdisk, snapshot, vmnamespace, vmswap, vmem, and so on.
LSOM	The Log-Structured Object Manager (LSOM) is responsible for locally storing the data on the vSAN file system as vSAN Component or LSOM-Object (data component or witness component).
NIC Teaming	Network Interface Card (NIC) teaming can be defined as two or more network adapters (NICs) that are set up as a "team" for high availability and load balancing.
NIOC	Network I/O Control (NIOC) determines the bandwidth that different network traffic types are given on a vSphere distributed switch. The bandwidth distribution is a user configurable parameter. When NIOC is enabled, distributed switch traffic is divided into predefined network resource pools: Fault Tolerance traffic, iSCSI traffic, vMotion traffic, management traffic, vSphere Replication traffic, NFS traffic, and virtual machine traffic.
Objects and Components	Each object is composed of a set of components, determined by capabilities that are in use in the VM Storage Policy. A vSAN datastore contains several object types: ■ **VM Home Namespace** - The VM Home Namespace is a virtual machine home directory where all virtual machine configuration files are stored. This includes files such as .vmx, log files, vmdks, and snapshot delta description files. ■ **VMDK** - VMDK is a virtual machine disk or .vmdk file that stores the contents of the virtual machine's hard disk drive. ■ **VM Swap Object** - VM Swap Objects are created when a virtual machine is powered on. ■ **Snapshot Delta VMDKs** - Snapshot Delta VMDKs are created when virtual machine snapshots are taken. ■ **Memory Object** - Memory Objects are created when the snapshot memory option is selected when creating or suspending a virtual machine.
RDT	The Reliable Data Transport (RDT) protocol is used for communication between hosts over the vSAN VMkernel ports. It uses TCP at the transport layer and is responsible to create and destroy TCP connections (sockets) on demand. It is optimized to send large files.

Terms	Definitions
SPBM	Storage Policy-Based Management (SPBM) provides a storage policy framework that serves as a single unified control panel across a broad range of data services and storage solutions. This framework helps you to align storage with application demands of your virtual machines.
VASA	The vSphere Storage APIs for Storage Awareness (VASA) is a set of application program interfaces (APIs) that enables vCenter Server to recognize the capabilities of storage arrays. VASA providers communicate with vCenter Server to determine the storage topology, capability, and state information which supports policy-based management, operations management, and DRS functionality.
VLAN	A VLAN enables a single physical LAN segment to be further segmented so that groups of ports are isolated from one another as if they were on physically different segments.
Witness Component	A witness is a component that contains only metadata and does not contain any actual application data. It serves as a tiebreaker when a decision must be made regarding the availability of the surviving datastore components, after a potential failure. A witness consumes approximately 2 MB of space for metadata on the vSAN datastore when using on-disk format 1.0, and 4 MB for the on-disk format for version 2.0 and later.

Understanding vSAN Networking

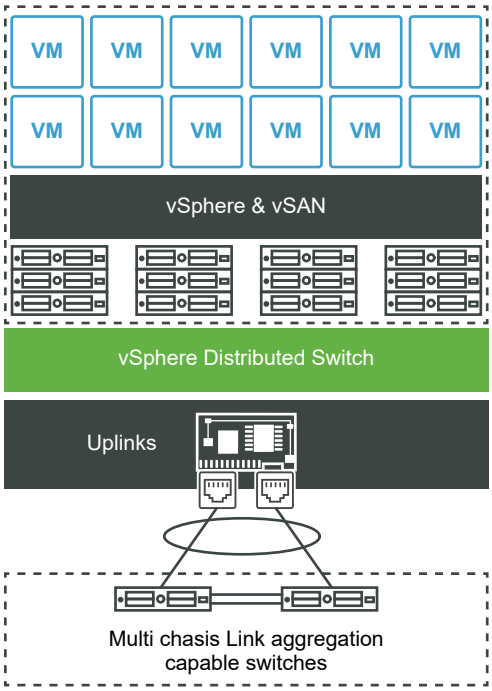
4

A vSAN network facilitates the communication between cluster hosts, and must guarantee fast performance, high availability, and bandwidth.

vSAN uses the network to communicate between the ESXi hosts and for virtual machine disk I/O.

Virtual machines (VMs) on vSAN datastores are made up of a set of objects, and each object can be made up of one or more components. These components are distributed across multiple hosts for resilience to drive and host failures. vSAN maintains and updates these components using the vSAN network.

The following diagram provides a high-level overview of the vSAN network:



This chapter includes the following topics:

- vSAN Network Characteristics
- ESXi Traffic Types
- Network Requirements for vSAN

vSAN Network Characteristics

vSAN is network-dependent. Understanding and configuring the right vSAN network settings is key to avoiding performance and stability issues.

A reliable and robust vSAN network has the following characteristics:

Unicast

vSAN 6.6 and later releases support unicast communication. Unicast traffic is a one-to-one transmission of IP packets from one point in the network to another point. Unicast transmits the heartbeat sent from the primary host to all other hosts each second. This ensures that the hosts are active and indicates the participation of hosts in the vSAN cluster. You can design a simple unicast network for vSAN. For more information on the unicast communication, see Chapter 5 Using Unicast in vSAN Network.

Multicast

Releases earlier than vSAN 6.6 use IP multicast communication as a discovery protocol to identify the nodes trying to join a vSAN cluster.

Note If possible, always use the latest version of vSAN.

IP multicast relies on communication protocols used by hosts, clients, and network devices to participate in multicast-based communications. For more information on the multicast communication, see Chapter 14 Using Multicast in vSAN Network.

Layer 2 and Layer 3 Network

All hosts in the vSAN cluster must be connected through a Layer 2 or Layer 3 network. vSAN releases earlier than vSAN 6.0 support only Layer 2 networking, whereas subsequent releases include support for both Layer 2 and Layer 3 protocols. Use a Layer 2 or Layer 3 network to provide communication between the data sites and the witness site. For more information on Layer 2 and Layer 3 network topologies, see Standard Deployments.

VMkernel Network

Each ESXi host in a vSAN cluster must have a network adapter for vSAN communication. All the intra-cluster node communication happens through the vSAN VMkernel port. VMkernel ports provide Layer 2 and Layer 3 services to each vSAN host and hosted virtual machines.

vSAN Network Traffic

Several different traffic types are available in the vSAN network, such as the storage traffic and the unicast traffic. The compute and storage of a virtual machine can be on the same host or on different hosts in the cluster. A VM that is not configured to tolerate a failure might be running on one host, and accessing a VM object or component that resides on a different host. This implies that all I/O from the VM passes through the network. The storage traffic constitutes most of the traffic in a vSAN cluster.

The cluster-related communication between all the ESXi hosts creates traffic in the vSAN cluster. This unicast traffic also contributes to the vSAN network traffic.

Virtual Switch

vSAN supports the following types of virtual switches:

- The Standard Virtual Switch provides connectivity from VMs and VMkernel ports to external networks. This switch is local to each ESXi host.
- A vSphere Distributed Switch provides central control of the virtual switch administration across multiple ESXi hosts. A distributed switch also provides networking features such as Network I/O Control (NIOC) that can help you set Quality of Service (QoS) levels on vSphere or virtual network. vSAN includes vSphere Distributed Switch irrespective of the vCenter Server version.

Bandwidth

vSAN traffic can share physical network adapters with other system traffic types, such as vSphere vMotion traffic, vSphere HA traffic, and virtual machine traffic. It also provides more bandwidth for shared network configurations where vSAN, vSphere management, vSphere vMotion traffic, and so on, are on the same physical network. To guarantee the amount of bandwidth required for vSAN, use vSphere Network I/O Control in the distributed switch.

In vSphere Network I/O Control, you can configure reservation and shares for the vSAN outgoing traffic:

- Set a reservation so that Network I/O Control guarantees that a minimum bandwidth is available on the physical adapter for vSAN.
- Set the share value to 100 so that when the physical adapter assigned for vSAN becomes saturated, certain bandwidth is available to vSAN. 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.

For information about using Network I/O Control to configure bandwidth allocation for vSAN traffic, see the *vSphere Networking* documentation.

ESXi Traffic Types

ESXi hosts use different network traffic types to support vSAN.

Following are the different traffic types that you need to set up for vSAN.

Table 4-1. Network Traffic Types

Traffic Types	Description
Management network	The management network is the primary network interface that uses a VMkernel TCP/IP stack to facilitate the host connectivity and management. It can also handle the system traffic such as vMotion, iSCSI, Network File System (NFS), Fiber Channel over Ethernet (FCoE), and fault tolerance.
Virtual Machine network	With virtual networking, you can network virtual machines and build complex networks within a single ESXi host or across multiple ESXi hosts.
vMotion network	Traffic type that facilitates migration of VM from one host to another. Migration with vMotion requires correctly configured network interfaces on source and target hosts. Ensure that the vMotion network is distinct from the vSAN network.
vSAN network	A vSAN cluster requires the VMkernel network for the exchange of data. Each ESXi host in the vSAN cluster must have a VMkernel network adapter for the vSAN traffic. For more information, refer to "Manually Enabling vSAN" in *vSAN Planning and Deployment*.

Network Requirements for vSAN

vSAN is a distributed storage solution that depends on the network for communication between hosts. Before deployment, ensure that your vSAN environment has all the networking requirements.

Physical NIC Requirements

Network Interface Cards (NICs) used in vSAN hosts must meet certain requirements. vSAN works on 10 Gbps, 25 Gbps, 40 Gbps, 50 Gbps, and 100 Gbps networks.

Ensure your hosts meet the minimum NIC requirements for vSAN Original Storage Architecture (OSA) or vSAN Express Storage Architecture (ESA).

Table 4-2. vSAN OSA Minimum NIC Requirements and Recommendations

Topology or Deployment Mode	Architecture	Support for 1 GbE NIC	Support for 10 GbE NIC	Support for NICs Greater than 10 GbE	Inter-Node Latency	Inter-Site Link Bandwidth or Latency	Latency Between Nodes and vSAN Witness Hosts	Bandwidth Between Nodes and vSAN Witness Hosts
Standard Cluster	Hybrid Cluster	Yes (Minimum)	Yes (Recommended)	Yes	Less than 1 ms RTT.	NA	NA	NA
	All-Flash Cluster	No	Yes	Yes (Recommended)				
Stretched Cluster	Hybrid or All-Flash Cluster	No	Yes (Minimum)	Yes	Less than 1 ms RTT within each site.	Recommended is 10 GbE (Workload Dependent) and 5 ms RTT or less.	Less than 200 ms RTT. Up to 10 hosts per site.	2 Mbps per 1000 components (Maximum of 100 Mbps with 45 k components).
							Less than 100 ms RTT. 11–15 hosts per site.	
Two-Node Cluster	Hybrid Cluster	Yes (Up to 10 VMs)	Yes (Recommended)	Yes	Less than 1 ms RTT within the same site.	Recommended is 10 GbE and 5 ms RTT or less.	Less than 500 ms RTT.	2 Mbps per 1000 components (Maximum of 1.5 Mbps).
	All-Flash Cluster	No	Yes (Minimum)					

Table 4-3. vSAN ESA Minimum NIC Requirements and Recommendations

Deployment Type	Support for 1 GbE NIC	Support for 10 GbE NIC	Support for NICs Greater than 10 GbE	Inter-Node Latency	Inter-Site Link Bandwidth or Latency	Latency Between Nodes and vSAN Witness Hosts	Bandwidth Between Nodes and vSAN Witness Hosts
Standard Cluster	No	No	Yes (25 GbE minimum)	Less than 1 ms RTT.	NA	NA	NA
Stretched Cluster	No	No	Yes (25 GbE minimum)	Less than 1 ms RTT within each site.	Recommended is 25 GbE (Workload Dependent) and 5 ms RTT or less.	Less than 200 ms RTT. Up to 10 hosts per site.	2 Mbps per 1000 components (Maximum of 100 Mbps with 45 k components).

Table 4-3. vSAN ESA Minimum NIC Requirements and Recommendations (continued)

Deployment Type	Support for 1 GbE NIC	Support for 10 GbE NIC	Support for NICs Greater than 10 GbE	Inter-Node Latency	Inter-Site Link Bandwidth or Latency	Latency Between Nodes and vSAN Witness Hosts	Bandwidth Between Nodes and vSAN Witness Hosts
						Less than 100 ms RTT. 11–15 hosts per site.	
Two-Node Cluster	No	No	Yes (25 GbE minimum)	Less than 1 ms RTT within the same site.	Recommended is 25 GbE and 5 ms RTT or less.	Less than 500 ms RTT.	2 Mbps per 1000 components (Maximum of 1.5 Mbps).

Note These NIC requirements assume that the packet loss is not more than 0.0001% in the hyper-converged environments. There can be a drastic impact on the vSAN performance, if any of these requirements are exceeded.

For more information about the stretched cluster NIC requirements, see *vSAN Stretched Cluster Guide*.

Bandwidth and Latency Requirements

To ensure high performance and availability, vSAN clusters must meet certain bandwidth and network latency requirements.

The bandwidth requirements between the primary and secondary sites of a vSAN stretched cluster depend on the vSAN workload, amount of data, and the way you want to handle failures. For more information, see *VMware vSAN Design and Sizing Guide*.

Table 4-4. Bandwidth and Latency Requirements

Site Communication	Bandwidth	Latency
Site to Site	vSAN OSA: minimum of 10 Gbps vSAN ESA: minimum of 25 Gbps	Less than 5 ms latency RTT.
Site to Witness	2 Mbps per 1000 vSAN components	■ Less than 500 ms latency RTT for 1 host per site. ■ Less than 200 ms latency RTT for up to 10 hosts per site. ■ Less than 100 ms latency RTT for 11-15 hosts per site.

Layer 2 and Layer 3 Support

VMware recommends Layer 2 connectivity between all vSAN hosts sharing the subnet.

vSAN also supports deployments using routed Layer 3 connectivity between vSAN hosts. You must consider the number of hops and additional latency incurred while the traffic gets routed.

Table 4-5. Layer 2 and Layer 3 Support

Cluster Type	L2 Supported	L3 Supported	Considerations
Hybrid Cluster	Yes	Yes	L2 is recommended and L3 is supported.
All-Flash Cluster	Yes	Yes	L2 is recommended and L3 is supported.
Stretched Cluster Data	Yes	Yes	Both L2 and L3 between data sites are supported.
Stretched Cluster Witness	No	Yes	L3 is supported. L2 between data and witness sites is not supported.
Two-Node Cluster	Yes	Yes	Both L2 and L3 between data sites are supported.

Routing and Switching Requirements

All three sites in a stretched cluster communicate across the management network and across the vSAN network. The VMs in all data sites communicate across a common virtual machine network.

Following are the vSAN stretched cluster routing requirements:

Table 4-6. Routing Requirements

Site Communication	Deployment Model	Layer	Routing
Site to Site	Default	Layer 2	Not required
Site to Site	Default	Layer 3	Static routes are required.
Site to Witness	Default	Layer 3	Static routes are required.
Site to Witness	Witness Traffic Separation	Layer 3	Static routes are required when using an interface other than the Management (vmk0) interface.
Site to Witness	Witness Traffic Separation	Layer 2 for two-host cluster	Static routes are not required.

Virtual Switch Requirements

You can create a vSAN network with either vSphere Standard Switch or vSphere Distributed Switch. Use a distributed switch to prioritize bandwidth for vSAN traffic. vSAN uses a distributed switch with all the vCenter Server versions.

The following table compares the advantages and benefits of a distributed switch over a standard switch:

Table 4-7. Virtual Switch Types

Design Requirement	Option 1 - vSphere Distributed Switch	Option 2 - vSphere Standard Switch	Description
Availability	No impact	No impact	You can use either of the options
Manageability	Positive impact	Negative impact	The distributed switch is centrally managed across all hosts, unlike the standard switch which is managed on each host individually.
Performance	Positive impact	Negative impact	The distributed switch has added controls, such as Network I/O Control, which you can use to guarantee performance for vSAN traffic.
Recoverability	Positive impact	Negative impact	The distributed switch configuration can be backed up and restored, the standard switch does not have this functionality.
Security	Positive impact	Negative impact	The distributed switch has added built-in security controls to help protect traffic.

vSAN Network Port Requirements

vSAN deployments require specific network ports and settings to provide access and services.

vSAN sends messages on certain ports on each host in the cluster. Verify that the host firewalls allow traffic on these ports. For the list of all supported vSAN ports and protocols, see the VMware Ports and Protocols portal at https://ports.vmware.com/.

Firewall Considerations

When you enable vSAN on a cluster, all required ports are added to ESXi firewall rules and configured automatically. There is no need for an administrator to open any firewall ports or enable any firewall services manually.

You can view open ports for incoming and outgoing connections. Select the ESXi host, and click **Configure > Security Profile**.

Network Firewall Requirements

When you configure the network firewall, consider which version of vSAN you are deploying.

When you enable vSAN on a cluster, all required ports are added to ESXifirewall rules and configured automatically. You do not need to open any firewall ports or enable any firewall services manually. You can view open ports for incoming and outgoing connections in the ESXi host security profile (**Configure > Security Profile**).

vsanEncryption Firewall Rule

If your cluster uses vSAN encryption, consider the communication between hosts and the KMS server.

vSAN encryption requires an external Key Management Server (KMS). vCenter Server obtains the key IDs from the KMS, and distributes them to the ESXi hosts. KMS servers and ESXi hosts communicate directly with each other. KMS servers might use different port numbers, so the vsanEncryption firewall rule enables you to simplify communication between each vSAN host and the KMS server. This allows a vSAN host to communicate directly to any port on a KMS server (TCP port 0 through 65535).

When a host establishes communication to a KMS server, the following operations occur.

- The KMS server IP is added to the vsanEncryption rule and the firewall rule is enabled.
- Communication between vSAN node and KMS server is established during the exchange.
- After communication between the vSAN node and the KMS server ends, the IP address is removed from vsanEncryption rule, and the firewall rule is deactivated again.

vSAN hosts can communicate with multiple KMS hosts using the same rule.

Using Unicast in vSAN Network

5

Unicast traffic refers to a one-to-one transmission from one point in the network to another. vSAN version 6.6 and later uses unicast to simplify network design and deployment.

All ESXi hosts use the unicast traffic, and the vCenter Server becomes the source for the cluster membership. The vSAN nodes are automatically updated with the latest host membership list that vCenter provides. vSAN communicates using unicast for CMMDS updates.

Releases earlier than vSAN version 6.6 rely on multicast to enable the heartbeat and to exchange metadata between hosts in the cluster. If some hosts in your vSAN cluster are running earlier versions of the software, a multicast network is still required. The switch to unicast network from multicast provides better performance and network support. For more information on multicast, see Chapter 14 Using Multicast in vSAN Network.

This chapter includes the following topics:

Pre-Version 5 Disk Group Behavior

The availability of a single version 5 disk group in vSAN version 6.6 disk group triggers the cluster to communicate permanently in the unicast mode.

vSAN version 6.6 clusters automatically revert to multicast communication in the following situations:

- All cluster hosts are running vSAN version 6.5 or lower.
- All disk groups are using on-disk version 3 or earlier.
- A non-vSAN 6.6 host such as vSAN 6.2 or vSAN 6.5 is added to the cluster.

For example, if a host running vSAN 6.5 or earlier is added to an existing vSAN 6.6 cluster, the cluster reverts to multicast mode and includes the 6.5 host as a valid node. To avoid this behavior, use the latest version for both ESXi hosts and on-disk format. To ensure that vSAN cluster continues communicating in unicast mode and does not revert to multicast, upgrade the disk groups on the vSAN 6.6 hosts to on-disk version 5.0.

Note Avoid having a mixed mode cluster where vSAN version 6.5 or earlier are available in the same cluster along with vSAN version 6.6 or later.

Version 5 Disk Group Behavior

The presence of a single version 5 disk group in a vSAN version 6.6 cluster triggers the cluster to communicate permanently in unicast mode.

In an environment where a vSAN 6.6 cluster is already using an on-disk version 5 and a vSAN 6.5 node is added to the cluster, the following events occur:

- The vSAN 6.5 node forms its own network partition.
- The vSAN 6.5 node continues to communicate in multicast mode but is unable to communicate with vSAN 6.6 nodes as they use unicast mode.

A cluster summary warning appears on the on-disk format showing that one node is at an earlier version. You can upgrade the node to the latest version. You cannot upgrade disk format versions when a cluster is in a mixed mode.

DHCP Support on Unicast Network

vCenter Server deployed on a vSAN 6.6 cluster can use IP addresses from Dynamic Host Configuration Protocol (DHCP) without reservations.

You can use DHCP with reservations as the assigned IP addresses are tied to the MAC addresses of VMkernel ports.

IPv6 Support on Unicast Network

vSAN 6.6 supports IPv6 with unicast communications.

With IPv6, the link-local address is automatically configured on any interface using the link-local prefix. By default, vSAN does not add the link local address of a node to other neighboring cluster nodes. As a result, vSAN 6.6 does not support IPv6 link local addresses for unicast communications.

Query Unicast with ESXCLI

You can run ESXCLI commands to determine the unicast configuration.

View the Communication Modes

Using `esxcli vsan cluster get` command, you can view the CMMDS mode (unicast or multicast) of the vSAN cluster node.

Procedure

- Run the `esxcli vsan cluster get` command.

Results

```
Cluster Information
  Enabled: true
  Current Local Time: 2020-04-09T18:19:52Z
  Local Node UUID: 5e8e3dc3-43ab-5452-795b-a03d6f88f022
  Local Node Type: NORMAL
  Local Node State: AGENT
  Local Node Health State: HEALTHY
  Sub-Cluster Master UUID: 5e8e3d3f-3015-9075-49b6-a03d6f88d426
  Sub-Cluster Backup UUID: 5e8e3daf-e5e0-ddb6-a523-a03d6f88dd4a
  Sub-Cluster UUID: 5282f9f3-d892-3748-de48-e2408dc34f72
  Sub-Cluster Membership Entry Revision: 11
  Sub_cluster Member Count: 5
  Sub-Cluster Member UUIDs: 5e8e3d3f-3015-9075-49b6-a03d6f88d426, 5e8e3daf-e5e0-ddb6-a523-
a03d6f88dd4a,
  5e8e3d73-6d1c-0b81-1305-a03d6f888d22, 5e8e3d33-5825-ee5c-013c-a03d6f88ea4c,
5e8e3dc3-43ab-5452-795b-a03d6f88f022
  Sub-Cluster Member HostNames: testbed-1.vmware.com, testbed2.vmware.com,
  testbed3.vmware.com, testbed4.vmware.com, testbed5.vmware.com
  Sub-Cluster Membership UUID: 0f438e5e-d400-1bb2-f4d1-a03d6f88d426
  Unicast Mode Enabled: true
  Maintenance Mode State: OFF
  Config Generation: ed845022-5c08-48d0-aa1d-6b62c0022222 7 2020-04-08T22:44:14.889
```

Verify the vSAN Cluster Hosts

Use the `esxcli vsan cluster unicastagent list` command to verify whether the vSAN cluster hosts are operating in unicast mode.

Procedure

- Run the `esxcli vsan cluster unicastagent list` command.

Results

```
NodeUuid                              IsWitness Supports Unicast IP Address  Port  Iface Name
Cert Thumbprint  SubClusterUuid
------------------------------------ --------- ---------------- ----------  ----- ----------
5e8e3d73-6d1c-0b81-1305-a03d6f888d22         0      true
10.198.95.10    12321
43:80:B7:A1:3F:D1:64:07:8C:58:01:2B:CE:A2:F5:DE:D6:B1:41:AB
5e8e3daf-e5e0-ddb6-a523-a03d6f88dd4a         0      true
10.198.94.240   12321
```

```
FE:39:D7:A5:EF:80:D6:41:CD:13:70:BD:88:2D:38:6C:A0:1D:36:69
5e8e3d3f-3015-9075-49b6-a03d6f88d426         0      true
10.198.94.244   12321
72:A3:80:36:F7:5D:8F:CE:B0:26:02:96:00:23:7D:8E:C5:8C:0B:E1
5e8e3d33-5825-ee5c-013c-a03d6f88ea4c         0      true
10.198.95.11    12321
5A:55:74:E8:5F:40:2F:2B:09:B5:42:29:FF:1C:95:41:AB:28:E0:57
```

The output includes the vSAN node UUID, IPv4 address, IPv6 address, UDP port with which vSAN node communicates, and whether the node is a data host (0) or a witness host (1). You can use this output to identify the vSAN cluster nodes that are operating in unicast mode and view the other hosts in the cluster. vCenter Server maintains the output list.

View the vSAN Network Information

Use the `esxcli vsan network list` command to view the vSAN network information such as the VMkernel interface that vSAN uses for communication, the unicast port (12321), and the traffic type (vSAN or witness) associated with the vSAN interface.

Procedure

- Run the `esxcli vsan network list` command.

Results

```
Interface
  VmkNic Name: vmk1
  IP Protocol: IP
  Interface UUID: e290be58-15fe-61e5-1043-246e962c24d0
  Agent Group Multicast Address: 224.2.3.4
  Agent Group IPv6 Multicast Address: ff19::2:3:4
  Agent Group Multicast Port: 23451
  Master Group Multicast Address: 224.1.2.3
  Master Group IPv6 Multicast Address: ff19::1:2:3
  Master Group Multicast Port: 12345
  Host Unicast Channel Bound Port: 12321
  Multicast TTL: 5
  Traffic Type: vsan
```

This output also displays the multicast information.

Intra-Cluster Traffic

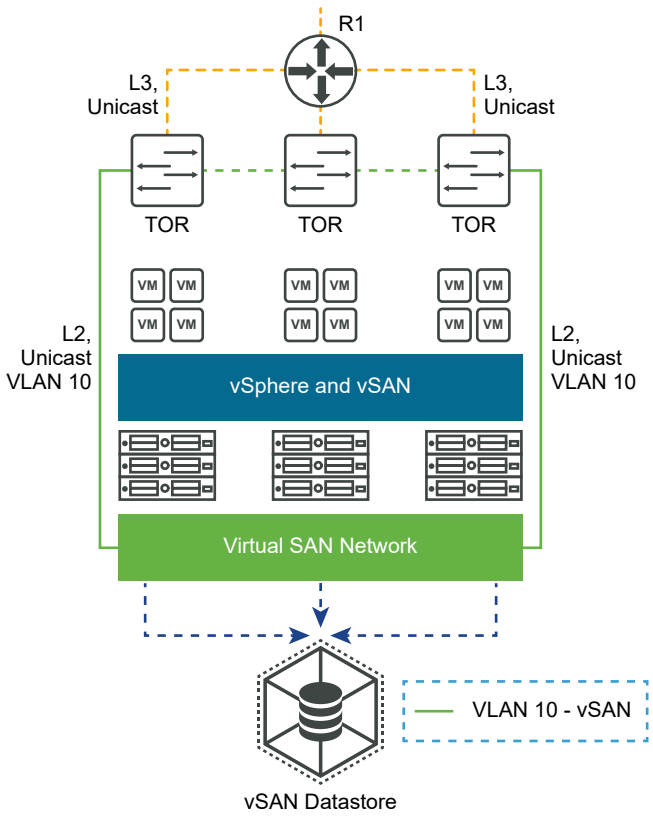
In unicast mode, the primary node addresses all the cluster nodes as it sends the same message to all the vSAN nodes in a cluster.

For example, if N is the number of vSAN nodes, then the primary node sends the messages N number of times. This results in a slight increase of vSAN CMMDS traffic. You might not notice this slight increase of traffic during normal, steady-state operations.

Intra-Cluster Traffic in a Single Rack

If all the nodes in a vSAN cluster are connected to the same top of the rack (TOR) switch, then the total increase in traffic is only between the primary node and the switch.

If a vSAN cluster spans more than one TOR switch, traffic between the switch expands. If a cluster spans many racks, multiple TORs form Fault Domains (FD) for the rack awareness. The primary node sends N messages to the racks or fault domains, where N is the number of hosts in each fault domain.

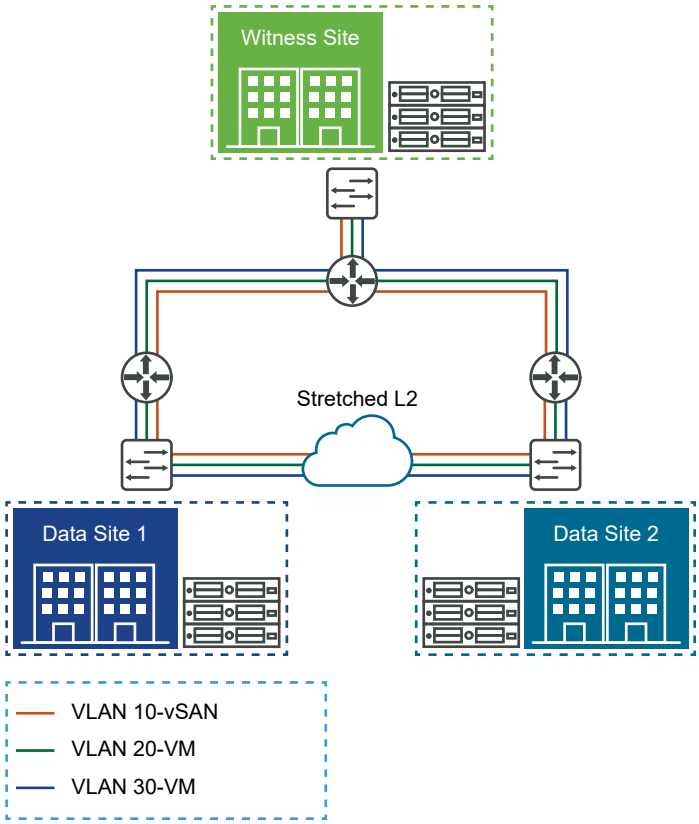


Intra-Cluster Traffic in a Stretched Cluster

In a stretched cluster, the primary node is located at the preferred site.

In a fault domain, CMMDS data must be communicated from the secondary site to the preferred site. To calculate the traffic in a stretched cluster, you must multiply the number of nodes in a secondary site with the CMMDS node size (in MB) to the number of nodes in the secondary site.

Traffic in a stretched cluster = number of nodes in the secondary site * CMMDS node size (in MB) * number of nodes in the secondary site.



With the unicast traffic, there is no change in the witness site traffic requirements.

Configuring IP Network Transport 6

Transport protocols provide communication services across the network. These services include the TCP/IP stack and flow control.

This chapter includes the following topics:

vSphere TCP/IP Stacks

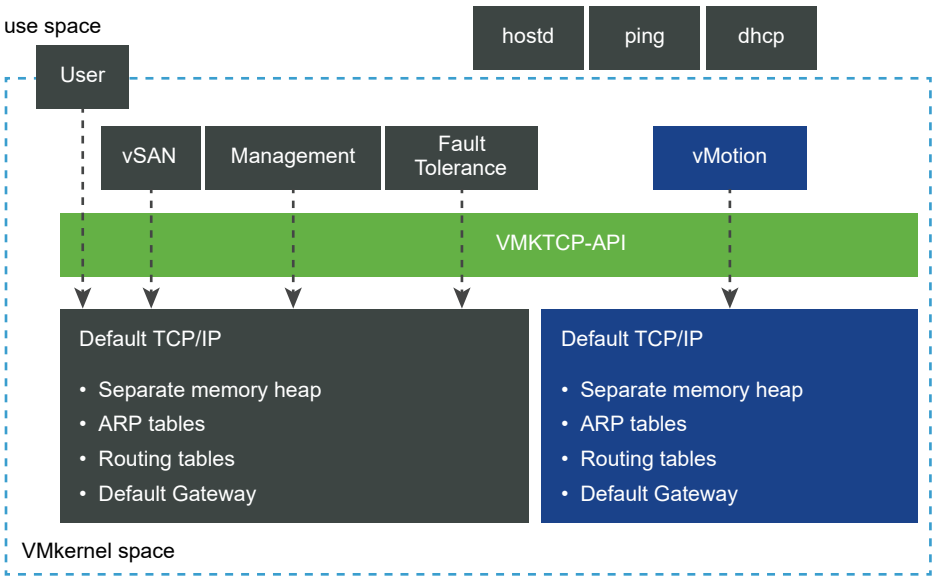
vSphere does not include a dedicated TCP/IP stack for the vSAN traffic service. You can add the vSAN VMkernel network interface to the default TCP/IP stack and define static routes for all hosts in the vSAN cluster.

vSphere does not support the creation of a custom vSAN TCP/IP stack. You can ensure vSAN traffic in Layer 3 network topologies leaves over the vSAN VMkernel network interface. Add the vSAN VMkernel network interface to the default TCP/IP stack and define static routes for all hosts in the vSAN cluster.

Note vSAN does not have its own TCP/IP stack. Use static routes to route vSAN traffic across L3 networks.

vSphere 6.0 introduced a new TCP/IP stack architecture, which can use multiple TPC/IP stacks to manage different VMkernel network interfaces. With this architecture, you can configure traffic services such as vMotion, management, and fault tolerance on isolated TCP/IP stacks, which can use multiple default gateways.

For network traffic isolation and security requirements, deploy the different traffic services onto different network segments or VLANs. This prevents the different traffic services from traversing through the same default gateway.



When you configure the traffic services on separate TCP/IP stacks, deploy each traffic service type onto its own network segment. The network segments are accessed through a physical network adapter with VLAN segmentation. Map each segment to different VMkernel network interfaces with the respective traffic services enabled.

TCP/IP Stacks Available in vSphere

vSphere provides TCP/IP stacks that support vSAN traffic requirements.

- **Default TCP/IP Stack**. Manage the host-related traffic services. This stack shares a single default gateway between all configured network services.
- **vMotion TCP/IP Stack**. Isolates vMotion traffic onto its own stack. The use of this stack completely removes or deactivates vMotion traffic from the default TCP/IP stack.
- **Provisioning TCP/IP Stack**. Isolates some virtual machine-related operations, such as cold migrations, cloning, snapshot, or NFC traffic.

You can select a different TCP/IP stack during the creation of a VMkernel interface.

Environments with isolated network requirements for the vSphere traffic services cannot use the same default gateway to direct traffic. Using different TCP/IP stacks simplifies management for traffic isolation, because you can use different default gateways and avoid adding static routes. Use this technique when you must route vSAN traffic to another network that is not accessible over the default gateway.

vSphere RDMA

vSAN 7.0 Update 2 and later supports Remote Direct Memory Access (RDMA) communication.

RDMA allows direct memory access from the memory of one computer to the memory of another computer without involving the operating system or CPU. The transfer of memory is offloaded to the RDMA-capable Host Channel Adapters (HCA).

vSAN supports the RoCE v2 protocol. RoCE v2 requires a network configured for lossless operation.

Each vSAN host must have a vSAN certified RDMA-capable NIC, as listed in the vSAN section of the VMware Compatibility Guide. Use only the same model network adapters from the same vendor on each end of the connection.

Note vSphere RDMA is not supported stretched or two-node vSAN clusters.

All hosts in the cluster must support RDMA. If any host loses RDMA support, the entire vSAN cluster switches to TCP.

vSAN with RDMA supports NIC failover, but does not support LACP or IP-hash-based NIC teaming.

IPv6 Support

vSAN 6.2 and later supports IPv6.

vSAN supports the following IP versions.

- IPv4
- IPv6 (vSAN 6.2 and later)
- Mixed IPv4/IPv6 (vSAN 6.2 and later)

In releases earlier than vSAN 6.2, only IPv4 is supported. Use mixed mode when migrating your vSAN cluster from IPv4 to IPv6.

IPv6 multicast is also supported. However, some restrictions are applicable with IPv6 and IGMP snooping on Cisco ACI. For this reason, do not implement IPv6 for vSAN using Cisco ACI.

For more information about using IPv6, consult with your network vendor.

Static Routes

You can use static routes to allow vSAN network interfaces from hosts on one subnet to reach the hosts on another network.

Most organizations separate the vSAN network from the management network, so the vSAN network does not have a default gateway. In an L3 deployment, hosts that are on different subnets or different L2 segments cannot reach each other over the default gateway, which is typically associated with the management network.

Use *static routes* to allow the vSAN network interfaces from hosts on one subnet to reach the vSAN networks on hosts on the other network. Static routes instruct a host how to reach a particular network over an interface, rather than using the default gateway.

The following example shows how to add an IPv4 static route to an ESXi host. Specify the gateway (-g) and the network (-n) you want to reach through that gateway:

```
esxcli network ip route ipv4 add –g 172.16.10.253 -n 192.168.10.0/24
```

When the static routes have been added, vSAN traffic connectivity is available across all networks, assuming the physical infrastructure allows it. Run the `vmkping` command to test and confirm communication between the different networks by pinging the IP address or the default gateway of the remote network. You also can check different size packets (-s) and prevent fragmentation (-d) of the packet.

```
vmkping –I vmk3 192.168.10.253
```

Jumbo Frames

vSAN fully supports jumbo frames on the vSAN network.

Jumbo frames are Ethernet frames with more than 1500 bytes of payload. Jumbo frames typically carry up to 9000 bytes of payload, but variations exist.

Using jumbo frames can reduce CPU utilization and improve throughput.

You must decide whether these gains outweigh the overhead of implementing jumbo frames throughout the network. In data centers where jumbo frames are already enabled in the network infrastructure, you can use them for vSAN. The operational cost of configuring jumbo frames throughout the network might outweigh the limited CPU and performance benefits.

Using VMware NSX with vSAN

7

vSAN and VMware NSX can be deployed and coexist in the same vSphere infrastructure.

NSX does not support the configuration of the vSAN data network over an NSX-managed VXLAN or Geneve overlay.

vSAN and NSX are compatible. vSAN and NSX are not dependent on each other to deliver their functionalities, resources, and services.

However, you cannot place the vSAN network traffic on an NSX-managed VxLAN/Geneve overlay. NSX does not support the configuration of the vSAN data network traffic over an NSX-managed VxLAN/Geneve overlay.

One reason VMkernel traffic is not supported over the NSX-managed VxLAN overlay is to avoid any circular dependency between the VMkernel networks and the VxLAN overlay that they support. The logical networks delivered with the NSX-managed VxLAN overlay are used by virtual machines, which require network mobility and flexibility.

When you implement LACP/LAG in NSX, the biggest issue with LAG is when it is used in a Cisco Nexus environment that defines the LAGs as virtual port channels (vPCs). Having a VPC implies you cannot run any dynamic routing protocol from edge devices to the physical Cisco switches, because Cisco does not support this.

Using Congestion Control and Flow Control

8

Use flow control to manage the rate of data transfer between senders and receivers on the vSAN network. Congestion control handles congestion in the network.

Flow Control

You can use flow control to manage the rate of data transfer between two devices.

Flow control is configured when two physically connected devices perform auto-negotiation.

An overwhelmed network node might send a pause frame to halt the transmission of the sender for a specified period. A frame with a multicast destination address sent to a switch is forwarded out through all other ports of the switch. Pause frames have a special multicast destination address that distinguishes them from other multicast traffic. A compliant switch does not forward a pause frame. Frames sent to this range are meant to be acted upon only within the switch. Pause frames have a limited duration, and expire after a time interval. Two computers that are connected through a switch never send pause frames to each other, but can send pause frames to a switch.

One reason to use pause frames is to support network interface controllers (NICs) that do not have enough buffering to handle full-speed reception. This problem is uncommon with advances in bus speeds and memory sizes.

Congestion Control

Congestion control helps you control the traffic on the network.

Congestion control applies mainly to packet switching networks. Network congestion within a switch might be caused by overloaded inter-switch links. If inter-switch links overload the capability on the physical layer, the switch introduces pause frames to protect itself.

Priority Flow Control

Priority-based flow control (PFC) helps you eliminate frame loss due to congestion.

Priority-based flow control (IEEE 802.1Qbb) is achieved by a mechanism similar to pause frames, but operates on individual priorities. PFC is also called Class-Based Flow Control (CBFC) or Per Priority Pause (PPP).

Flow Control and Congestion Control

Flow control is an end-to-end mechanism that controls the traffic between a sender and a receiver. Flow control occurs in the data link layer and the transport layer.

Congestion control is used by a network to control congestion in the network. This problem is not as common in modern networks with advances in bus speeds and memory sizes. A more likely scenario is network congestion within a switch. Congestion Control is handled by the network layer and the transport layer.

Flow Control Design Considerations

By default, flow control is enabled on all network interfaces in ESXi hosts.

Flow control configuration on a NIC is done by the driver. When a NIC is overwhelmed by network traffic, the NIC sends pause frames.

Flow control mechanisms such as pause frames can trigger overall latency in the VM guest I/O due to increased latency at the vSAN network layer. Some network drivers provide module options that configure flow control functionality within the driver. Some network drivers enable you to modify the configuration options using the `ethtool` command-line utility on the console of the ESXi host. Use module options or `ethtool`, depending on the implementation details of a given driver.

For information about configuring flow control on ESXi hosts, see VMware KB 1013413.

In deployments with 1 Gbps, leave flow control enabled on ESXi network interfaces (default). If pause frames are a problem, carefully plan disabling flow control in conjunction with Hardware Vendor Support or VMware Global Support Services.

To learn how you can recognize the presence of pause frames being sent from a receiver to an ESXi host, see Chapter 13 Troubleshooting the vSAN Network. A large number of pause frames in an environment usually indicates an underlying network or transport issue to investigate.

Basic NIC Teaming, Failover, and Load Balancing

9

Many vSAN environments require some level of network redundancy.

You can use NIC teaming to achieve network redundancy. You can configure two or more network adapters (NICs) as a team for high availability and load balancing. Basic NIC teaming is available with vSphere networking, and these techniques can affect vSAN design and architecture.

Several NIC teaming options are available. Avoid NIC teaming policies that require physical switch configurations, or that require an understanding of networking concepts such as Link Aggregation. Best results are achieved with a basic, simple, and reliable setup.

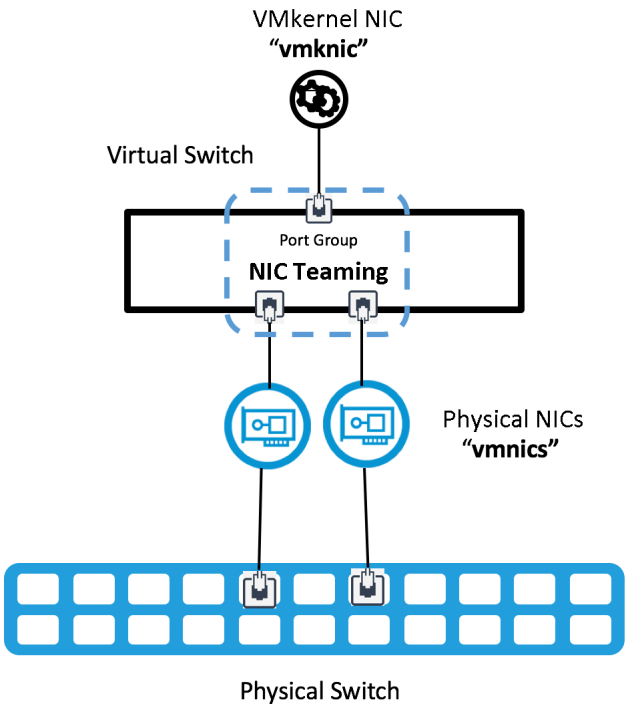
If you are not sure about NIC teaming options, use an Active/Standby configuration with explicit failover.

This chapter includes the following topics:

Basic NIC Teaming

Basic NIC teaming uses multiple physical uplinks, one vmknic, and a single switch.

vSphere NIC teaming uses multiple uplink adapters, called vmnics, which are associated with a single virtual switch to form a team. This is the most basic option, and you can configure it using a standard vSphere standard switch or a vSphere distributed switch.



Failover and Redundancy

vSAN can use the basic NIC teaming and failover policy provided by vSphere.

NIC teaming on a vSwitch can have multiple active uplinks, or an Active/Standby uplink configuration. Basic NIC teaming does not require any special configuration at the physical switch layer.

Note vSAN does not use NIC teaming for load balancing.

A typical NIC teaming configuration has the following settings. When working on distributed switches, edit the settings of the distributed port group used for vSAN traffic.

- Load balancing: Route based on originating virtual port
- Network failure detection: Link status only
- Notify switches: Yes
- Failback: Yes

Load Balancing vSAN traffic.

- Load balancing: Route based on originating virtual port
- Network failure detection: Link status only
- Notify switches: Yes
- Failback: Yes

Configure Load Balancing for NIC Teams

Several load-balancing techniques are available for NIC teaming, and each technique has its pros and cons.

Route Based on Originating Virtual Port

In Active/Active or Active/Passive configurations, use **Route based on originating virtual port** for basic NIC teaming. When this policy is in effect, only one physical NIC is used per VMkernel port.

Pros

- This is the simplest NIC teaming method that requires minimal physical switch configuration.
- This method requires only a single port for vSAN traffic, which simplifies troubleshooting.

Cons

- A single VMkernel interface is limited to a single physical NIC's bandwidth. As typical vSAN environments use one VMkernel adapter, only one physical NIC in the team is used.

Route Based on Physical NIC Load

Route Based on Physical NIC Load is based on **Route Based on Originating Virtual Port**, where the virtual switch monitors the actual load of the uplinks and takes steps to reduce load on overloaded uplinks. This load-balancing method is available only with a vSphere Distributed Switch, not on vSphere Standard Switches.

The distributed switch calculates uplinks for each VMkernel port by using the port ID and the number of uplinks in the NIC team. The distributed switch checks the uplinks every 30 seconds, and if the load exceeds 75 percent, the port ID of the VMkernel port with the highest I/O is moved to a different uplink.

Pros

- No physical switch configuration is required.
- Although vSAN has one VMkernel port, the same uplinks can be shared by other VMkernel ports or network services. vSAN can benefit by using different uplinks from other contending services, such as vMotion or management.

Cons

- As vSAN typically only has one VMkernel port configured, its effectiveness is limited.
- The ESXi VMkernel reevaluates the traffic load after each time interval, which can result in processing overhead.

Settings: Network Failure Detection

Use the default setting: **Link status only**. Do not use Beacon probing for link failure detection. Beacon probing requires at least three physical NICs to avoid split-brain scenarios. For more details, see VMware KB 1005577.

Settings: Notify Switches

Use the default setting: **Yes**. Physical switches have MAC address forwarding tables to associate each MAC address with a physical switch port. When a frame comes in, the switch determines the destination MAC address in the table and decides the correct physical port.

If a NIC failover occurs, the ESXi host must notify the network switches that something has changed, or the physical switch might continue to use the old information and send the frames to the wrong port.

When you set Notify Switches to **Yes**, if one physical NIC fails and traffic is rerouted to a different physical NIC in the team, the virtual switch sends notifications over the network to update the lookup tables on physical switches.

This setting does not catch VLAN misconfigurations, or uplink losses that occur further upstream in the network. The vSAN network partitions health check can detect these issues.

Settings: Failback

This option determines how a physical adapter is returned to active duty after recovering from a failure. A failover event triggers the network traffic to move from one NIC to another. When a **link up** state is detected on the originating NIC, traffic automatically reverts to the original network adapter when Failback is set to **Yes**. When Failback is set to **No**, a manual failback is required.

Setting Failback to **No** can be useful in some situations. For example, after a physical switch port recovers from a failure, the port might be active but can take several seconds to begin forwarding traffic. Automatic Failback has been known to cause problems in certain environments that use the Spanning Tree Protocol. For more information about Spanning Tree Protocol (STP), see VMware KB 1003804.

Setting Failover Order

Failover order determines which links are active during normal operations, and which links are active in the event of a failover. Different supported configurations are possible for the vSAN network.

Active/Standby uplinks: If a failure occurs on an Active/Standby setup, the NIC driver notifies vSphere of a link down event on Uplink 1. The standby Uplink 2 becomes active, and traffic resumes on Uplink 2.

Active/Active uplinks: If you set the failover order to Active/Active, the virtual port used by vSAN traffic cannot use both physical ports at the same time.

If your NIC teaming configuration for both Uplink 1 and Uplink 2 is active, there is no need for the standby uplink to become active.

Note When using an Active/Active configuration, ensure that Failback is set to **No**. For more information, see VMware KB 2072928.

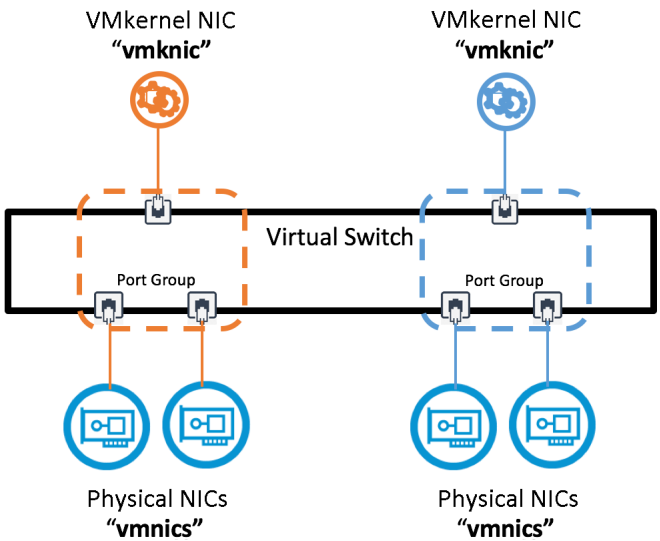
Advanced NIC Teaming

10

You can use advanced NIC teaming methods with multiple VMkernel adapters to configure the vSAN network. If you use Link Aggregation Protocol (LAG/LACP), the vSAN network can be configured with a single VMkernel adapter.

You can use advanced NIC teaming to implement an air gap, so a failure that occurs on one network path does not impact the other network path. If any part of one network path fails, the other network path can carry the traffic. Configure multiple VMkernel NICs for vSAN on different subnets, such as another VLAN or separate physical network fabric.

vSphere and vSAN do not support multiple VMkernel adapters (vmknics) on the same subnet. For more details, see VMware KB 2010877.



This chapter includes the following topics:

- Link Aggregation Group Overview
- Understanding Network Air Gaps
- Pros and Cons of Air Gap Network Configurations with vSAN
- NIC Teaming Configuration Examples

Link Aggregation Group Overview

By using the LACP protocol, a network device can negotiate an automatic bundling of links by sending LACP packets to a peer.

A Link Aggregation Group (LAG) is defined by the IEEE 802.1AX-2008 standard, which states that Link Aggregation allows one or more links to be aggregated together to form a Link Aggregation Group.

LAG can be configured as either static (manual) or dynamic by using LACP to negotiate the LAG formation. LACP can be configured as follows:

Active

Devices immediately send LACP messages when the port comes up. End devices with LACP enabled (for example, ESXi hosts and physical switches) send and receive frames called LACP messages to each other to negotiate the creation of a LAG.

Passive

Devices place a port into a passive negotiating state, in which the port only responds to received LACP messages, but do not initiate negotiation.

Note If the host and switch are both in passive mode, the LAG does not initialize, because an active part is required to trigger the linking. At least one must be Active.

In vSphere 5.5 and later releases, this functionality is called **Enhanced LACP**. This functionality is only supported on vSphere Distributed Switch version 5.5 or later.

For more information about LACP support on a vSphere Distributed Switch, see the vSphere 6 Networking documentation.

Note The number of LAGs you can use depends on the capabilities of the underlying physical environment and the topology of the virtual network.

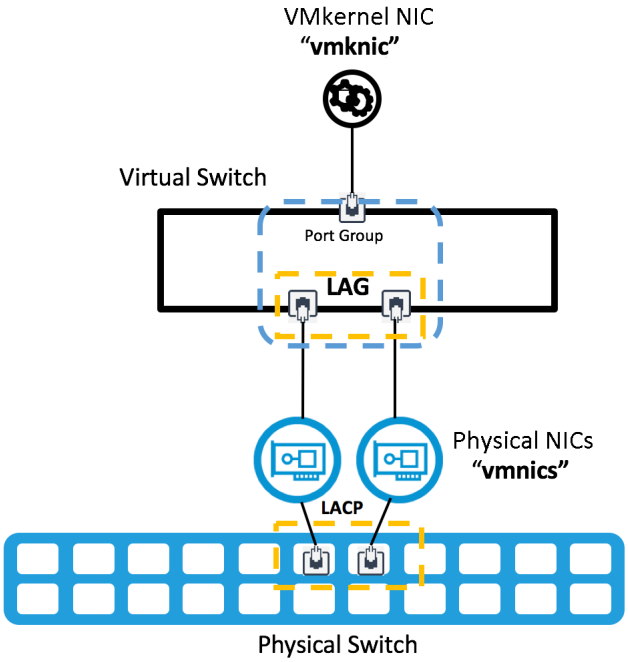
For more information about the different load-balancing options, see KB 2051826.

Static and Dynamic Link Aggregation

You can use LACP to combine and aggregate multiple network connections.

When LACP is in **active** or **dynamic** mode, a physical switch sends LACP messages to network devices, such as ESXi hosts, to negotiate the creation of a Link Aggregation Group (LAG).

To configure Link Aggregation on hosts using vSphere Standard Switches (and pre-5.5 vSphere Distributed Switches), configure a static channel-group on the physical switch. See your vendor documentation for more details.



Pros and Cons of Dynamic Link Aggregation

Consider the tradeoffs to using Dynamic Link Aggregation.

Pros

Improves performance and bandwidth. One vSAN host or VMkernel port can communicate with many other vSAN hosts using many different load-balancing options.

Provides network adapter redundancy. If a NIC fails and the link-state fails, the remaining NICs in the team continue to pass traffic.

Improves traffic balancing. Balancing of traffic after failures is automatic and fast.

Cons

Less flexible. Physical switch configuration requires that physical switch ports be configured in a port-channel configuration.

More complex. Use of multiple switches to produce full physical redundancy configuration is complex. Vendor-specific implementations add to the complexity.

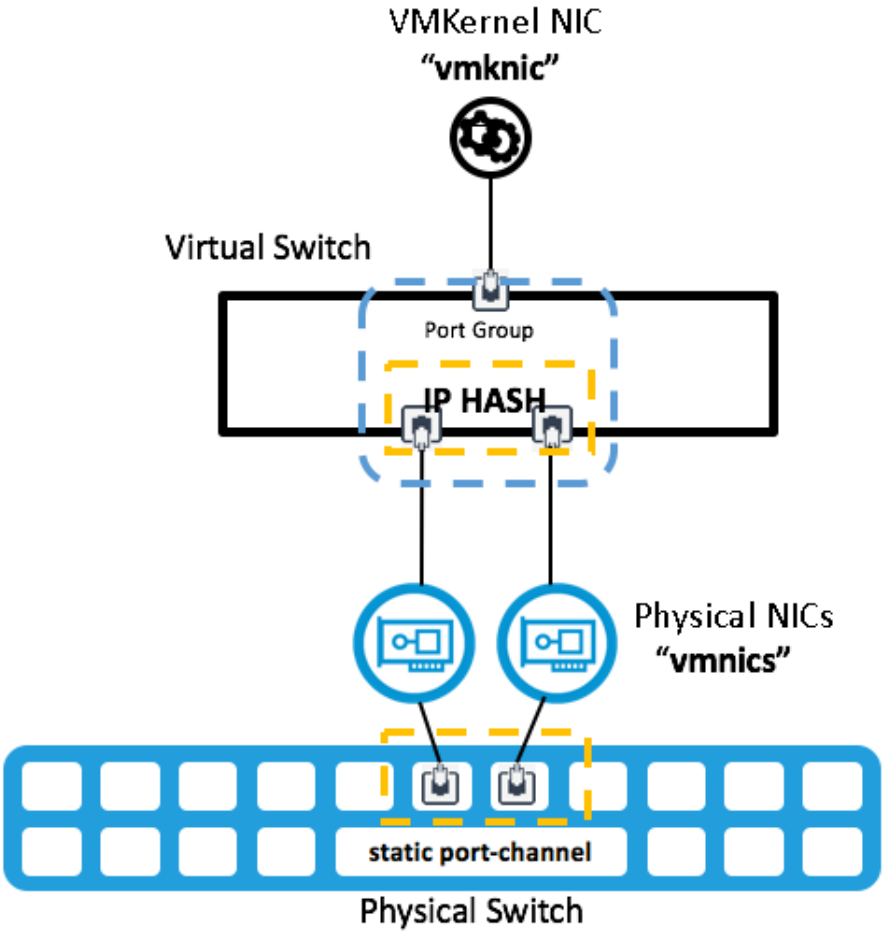
Static LACP with Route Based on IP Hash

You can create a vSAN 6.6 cluster using static LACP with an IP-hash policy. This section focuses on vSphere Standard Switches, but you also can use vSphere Distributed Switches.

You can use the Route based on IP Hash load balancing policy. For details about IP Hash, see the vSphere documentation.

Select **Route based on IP Hash** load balancing policy at a vSwitch or port-group level. Set all uplinks assigned to static channel group to the Active Uplink position on the Teaming and Failover Policies at the virtual switch or port-group level.

When IP Hash is configured on a vSphere port group, the port group uses the **Route based on IP Hash** policy. The number of ports in the port-channel must be same as the number of uplinks in the team.



Pros and Cons of Static LACP with IP Hash

Consider the tradeoffs to using Static LACP with IP Hash.

Pros

- **Improves performance and bandwidth**. One vSAN host or VMkernel port can communicate with many other vSAN hosts using the IP Hash algorithm.
- **Provides network adapter redundancy**. If a NIC fails and the link-state fails, the remaining NICs in the team continue to pass traffic.
- **Adds flexibility**. You can use IP Hash with both vSphere Standard Switches and vSphere Distributed Switches.

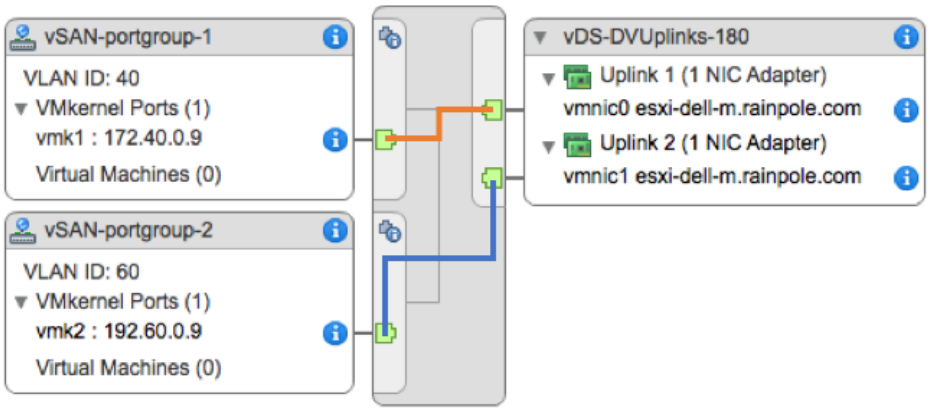
Cons

- **Physical switch configuration is less flexible**. Physical switch ports must be configured in a static port-channel configuration.
- **Increased chance of misconfiguration**. Static port-channels form without any verification on either end (unlike LACP dynamic port-channel).
- **More complex**. Introducing full physical redundancy configuration increases complexity when multiple switches are used. Implementations can become quite vendor specific.
- **Limited load balancing**. If your environment has only a few IP addresses, the virtual switch might consistently pass the traffic through one uplink in the team. This can be especially true for small vSAN clusters.

Understanding Network Air Gaps

You can use advanced NIC teaming methods to create an air-gap storage fabric. Two storage networks are used to create a redundant storage network topology, with each storage network physically and logically isolated from the other by an air gap.

You can configure a network air gap for vSAN in a vSphere environment. Configure multiple VMkernel ports per vSAN host. Associate each VMkernel port to dedicated physical uplinks, using either a single vSwitch or multiple virtual switches, such as vSphere Standard Switch or vSphere Distributed Switch.



Typically, each uplink must be connected to fully redundant physical infrastructure.

This topology is not ideal. The failure of components such as NICs on different hosts that reside on the same network can lead to interruption of storage I/O. To avoid this problem, implement physical NIC redundancy on all hosts and all network segments. Configuration example 2 addresses this topology in detail.

These configurations are applicable to both L2 and L3 topologies, with both unicast and multicast configurations.

Pros and Cons of Air Gap Network Configurations with vSAN

Network air gaps can be useful to separate and isolate vSAN traffic. Use caution when configuring this topology.

Pros

- Physical and logical separation of vSAN traffic.

Cons

- vSAN does not support multiple VMkernel adapters (vmknics) on the same subnet. For more information, see VMware KB 2010877.
- Setup is complex and error prone, so troubleshooting is more complex.
- Network availability is not guaranteed with multiple vmknics in some asymmetric failures, such as one NIC failure on one host and another NIC failure on another host.
- Load-balanced vSAN traffic across physical NICs is not guaranteed.
- Costs increase for vSAN hosts, as you might need multiple VMkernel adapters (vmknics) to protect multiple physical NICs (vmnics). For example, 2 x 2 vmnics might be required to provide redundancy for two vSAN vmknics.
- Required logical resources are doubled, such as VMkernel ports, IP addresses, and VLANs.
- vSAN does not implement port binding. This means that techniques such as multi-pathing are not available.
- Layer 3 topologies are not suitable for vSAN traffic with multiple vmknics. These topologies might not function as expected.
- Command-line host configuration might be required to change vSAN multicast addresses.

Dynamic LACP combines, or aggregates, multiple network connections in parallel to increase throughput and provide redundancy. When NIC teaming is configured with LACP, load balancing of the vSAN network across multiple uplinks occurs. This load balancing happens at the network layer, and is not done through vSAN.

Note Other terms sometimes used to describe link aggregation include port trunking, link bundling, Ethernet/network/NIC bonding, EtherChannel.

This section focuses on Link Aggregation Control Protocol (LACP). The IEEE standard is 802.3ad, but some vendors have proprietary LACP features, such as PAgP (Port Aggregation Protocol). Follow the best practices recommended by your vendor.

Note The LACP support introduced in vSphere Distributed Switch 5.1 only supports IP-hash load balancing. vSphere Distributed Switch 5.5 and later fully support LACP.

LACP is an industry standard that uses port-channels. Many hashing algorithms are available. The vSwitch port-group policy and the port-channel configuration must agree and match.

NIC Teaming Configuration Examples

The following NIC teaming configurations illustrate typical vSAN networking scenarios.

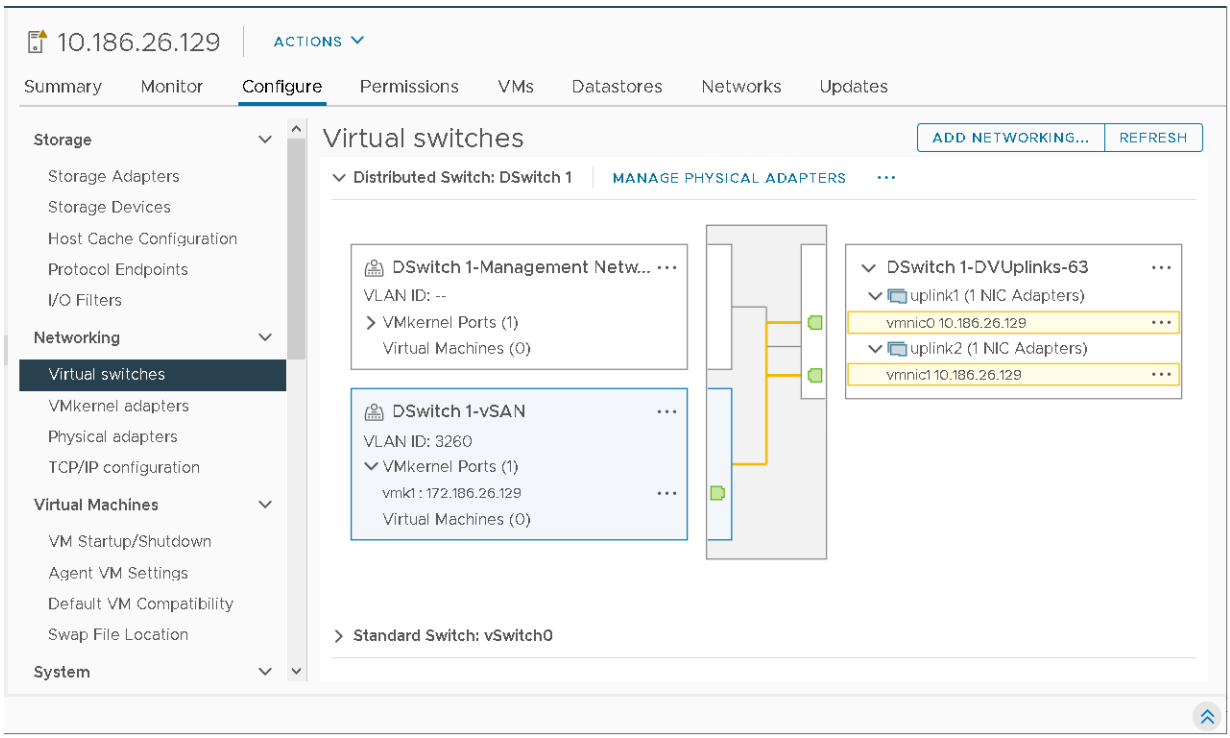
Configuration 1: Single vmknic, Route Based on Physical NIC Load

You can configure basic Active/Active NIC Teaming with the **Route based on Physical NIC Load** policy for vSAN hosts. Use a vSphere Distributed Switch (vDS).

For this example, the vDS must have two uplinks configured for each host. A distributed port group is designated for vSAN traffic and isolated to a specific VLAN. Jumbo frames are already enabled on the vDS with an MTU value of 9000.

Configure teaming and failover for the distributed port group for vSAN traffic as follows:

- Load balancing policy set to **Route Based on Physical Nic Load**.
- Network failure detection set to **Link status only**.
- Notify Switches set to **Yes**.
- Failback set to **No** . You can set Failback to **yes**, but not for this example.
- Ensure both uplinks are in the **Active uplinks** position.



Network Uplink Redundancy Lost

When the link down state is detected, the workload switches from one uplink to another. There is no noticeable impact to the vSAN cluster and VM workload.

Recovery and Failback

When you set **Failback** to **No**, traffic is not promoted back to the original vmnic. If **Failback** is set to **Yes**, traffic is promoted back to the original vmnic on recovery.

Load Balancing

Since this is a single VMkernel NIC, there is no performance benefit to using **Route based on physical load**.

Only one physical NIC is in use at any time. The other physical NIC is idle.

Configuration 2: Multiple vmknics, Route Based on Originating Port ID

You can use two non-routable VLANs that are logically and physically separated, to produce an air-gap topology.

This example provides configuration steps for a vSphere distributed switch, but you also can use vSphere standard switches. It uses two 10 Gb physical NICs and logically separates them on the vSphere networking layer.

Create two distributed port groups for each vSAN VMkernel vmknic. Each port group has a separate VLAN tag. For vSAN VMkernel configuration, two IP addresses on both VLANs are required for vSAN traffic.

Note Practical implementations typically use four physical uplinks for full redundancy.

For each port group, the teaming and failover policy use the default settings.

- Load balancing set to **Route based on originating port ID**
- Network failure detection set to **Link Status Only**
- Notify Switches set to the default value of **Yes**
- Failback set to the default value of **Yes**
- The uplink configuration has one uplink in the **Active** position and one uplink in the **Unused** position.

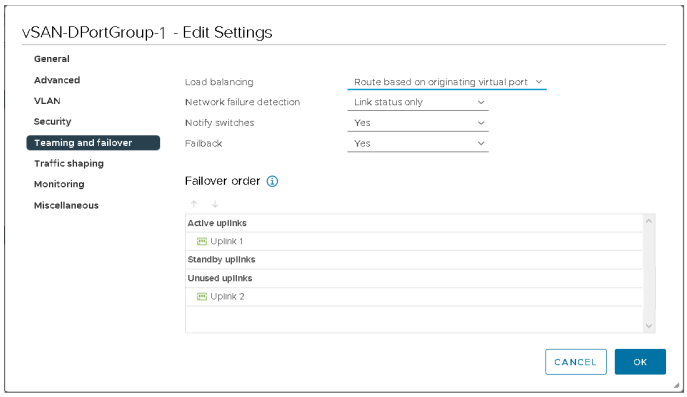
One network is completely isolated from the other network.

vSAN Port Group 1

This example uses a distributed port group called **vSAN-DPortGroup-1**. **VLAN 3266** is tagged for this port group with the following Teaming and Failover policy:

- Traffic on the port group tagged with VLAN 3266
- Load balancing set to **Route based on originating port ID**
- Network failure detection set to **Link Status Only**
- Notify Switches set to default value of **Yes**

- Failback set to default value of **Yes**
- The uplink configuration has **Uplink 1** in the **Active** position and **Uplink 2** in the **Unused** position.



vSAN Port Group 2

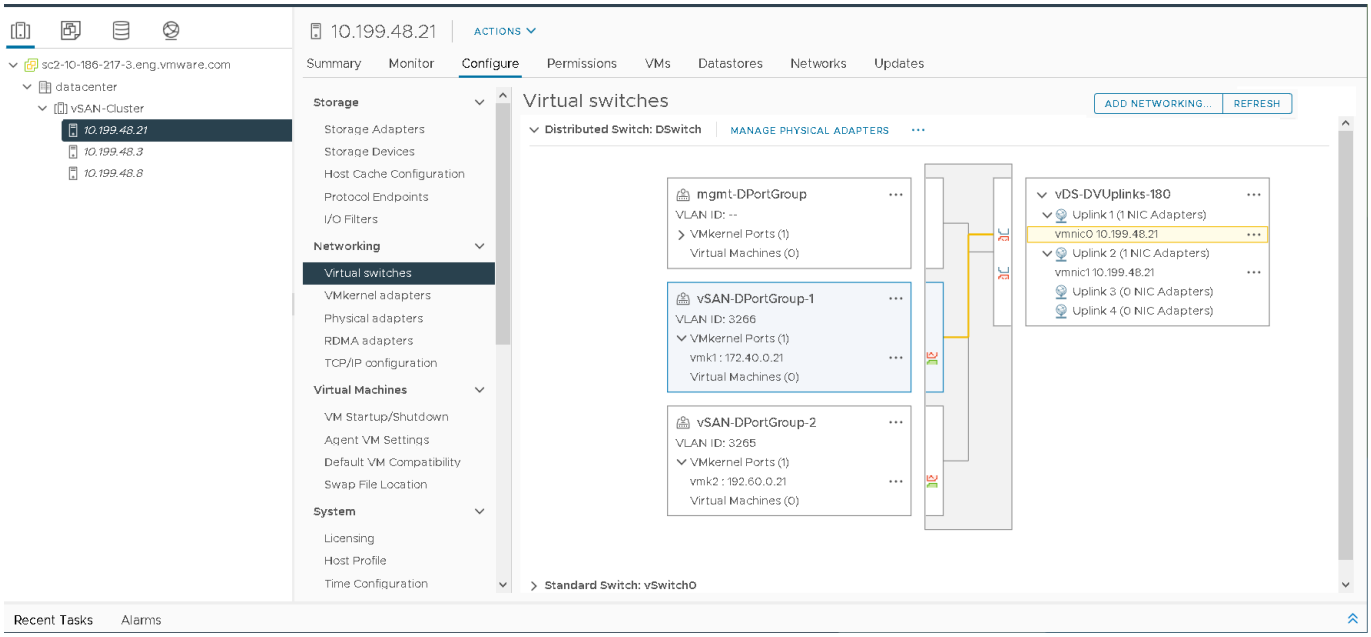
To complement vSAN port group 1, configure a second distributed port group called **vSAN-portgroup-2**, with the following differences:

- Traffic on the port group tagged with VLAN 3265
- The uplink configuration has **Uplink 2** in the **Active** position and **Uplink 1** in the **Unused** position.

vSAN VMkernel Port Configuration

Create two vSAN VMkernel interfaces and on both port groups. In this example, the port groups are named **vmk1** and **vmk2**.

- **vmk1** is associated with VLAN 3266 (172.40.0.xx), and as a result port group **vSAN-DPortGroup-1**.
- **vmk2** is associated with VLAN 3265 (192.60.0.xx), and as a result port group **vSAN-DPortGroup-2**.



Load Balancing

vSAN has no load balancing mechanism to differentiate between multiple vmknics, so the vSAN I/O path chosen is not deterministic across physical NICs. The vSphere performance charts show that one physical NIC is often more utilized than the other. A simple I/O test performed in our labs, using 120 VMs with a 70:30 read/write ratio with a 64K block size on a four-host all flash vSAN cluster, revealed an unbalanced load across NICs.

vSphere performance graphs show an unbalanced load across NICs.

Network Uplink Redundancy Lost

Consider a network failure introduced in this configuration. vmnic1 is not enabled on a given vSAN host. As a result, port **vmk2** is impacted. A failing NIC triggers both network connectivity alarms and redundancy alarms.

For vSAN, this failover process triggers approximately **10 seconds** after CMMDS (Cluster Monitoring, Membership, and Directory Services) detects a failure. During failover and recovery, vSAN stops any active connections on the failed network, and attempts to re-establish connections on the remaining functional network.

Since two separate vSAN VMkernel ports communicate on isolated VLANs, vSAN health check failures might be triggered. This is expected as **vmk2** can no longer communicate to its peers on VLAN 3265.

The performance charts show that the affected workload has restarted on vmnic0, since vmnic1 has a failure. This test illustrates an important distinction between vSphere NIC teaming and this topology. vSAN attempts to re-establish or restart connections on the remaining network.

However, in some failure scenarios, recovering the impacted connections might require up to **90 seconds** to complete, due to ESXi TCP connection timeout. Subsequent connection attempts might fail, but connection attempts time out at 5 seconds, and the attempts rotate through all possible IP addresses. This behavior might affect virtual machine guest I/O. As a result, application and virtual machine I/O might have to be retried.

For example, on Windows Server 2012 VMs, Event IDs 153 (device reset) and 129 (retry events) might be logged during the failover and recovery process. In the example, event ID 129 was logging for approximately 90 seconds until the I/O was recovered.

You might have to modify disk timeout settings of some guest OSes to ensure that they are not severely impacted. Disk timeout values might vary, depending on the presence of VMware Tools, and the specific guest OS type and version. For more information about changing guest OS disk timeout values, go to VMware KB 1009465.

Recovery and Failback

When the network is repaired, workloads are not automatically rebalanced unless another failure to force workload occurs, due to another failure. As soon as the impacted network is recovered, it becomes available for new TCP connections.

Configuration 3: Dynamic LACP

You can configure a two-port LACP port channel on a switch and a two-uplink Link Aggregation Group on a vSphere distributed switch.

In this example, use 10Gb networking with two physical uplinks per server.

Note vSAN over RDMA does not support this configuration.

Configure the Network Switch

Configure the vSphere distributed switch with the following settings.

- Identify the ports in question where the vSAN host will connect.
- Create a port channel.
- If using VLANs, then trunk the correct VLAN to the port channel.
- Configure the desired distribution or load-balancing options (hash).
- Setting LACP mode to active/dynamic.
- Verify MTU configuration.

Configure vSphere

Configure the vSphere network with the following settings.

- Configure vDS with the correct MTU.
- Add hosts to vDS.

- Create a LAG with the correct number of uplinks and matching attributes to port channel.
- Assign physical uplinks to the LAG.
- Create a distributed port group for vSAN traffic and assign correct VLAN.
- Configure VMkernel ports for vSAN with correct MTU.

Set Up the Physical Switch

Configure the physical switch with the following settings. For guidance about how to set up this configuration on Dell servers, refer to: http://www.dell.com/Support/Article/us/en/19/HOW10364.

Configure a two uplink LAG:

- Use switch ports 36 and 18.
- This configuration uses VLAN trunking, so port channel is in VLAN trunk mode, with the appropriate VLANs trunked.
- Use the following method for load-balancing or load distribution: **Source and destination IP addresses, TCP/UDP port and VLAN**
- Verify that the LACP mode is **Active** (Dynamic).

Use the following commands to configure an individual port channel on a Dell switch:

- Create a port-channel.

```
#interface port-channel 1
```

- Set port-channel to VLAN trunk mode.

```
#switchport mode trunk
```

- Allow VLAN access.

```
#switchport trunk allowed vlan 3262
```

- Configure the load balancing option.

```
#hashing-mode 6
```

- Assign the correct ports to the port-channel and set the mode to Active.
- Verify that the port channel is configured correctly.

```
#show interfaces port-channel 1


Channel Ports Ch-Type Hash Type Min-links Local Prf

------- ------------------------------ -------- --------- --------- ---------

Po1 Active: Te1/0/36, Te1/0/18 Dynamic 6 1 Disabled

Hash Algorithm Type

1 - Source MAC, VLAN, EtherType, source module and port Id
```

```
    2 - Destination MAC, VLAN, EtherType, source module and port Id

    3 - Source IP and source TCP/UDP port

    4 - Destination IP and destination TCP/UDP port

    5 - Source/Destination MAC, VLAN, EtherType, source MODID/port

    6 - Source/Destination IP and source/destination TCP/UDP port

    7 - Enhanced hashing mode

#interface range Te1/0/36, Te1/0/18

#channel-group 1 mode active
```

Full configuration:

```
#interface port-channel 1

#switchport mode trunk

#switchport trunk allowed vlan 3262

#hashing-mode 6

#exit

#interface range Te1/0/36,Te1/018

#channel-group 1 mode active

#show interfaces port-channel 1
```

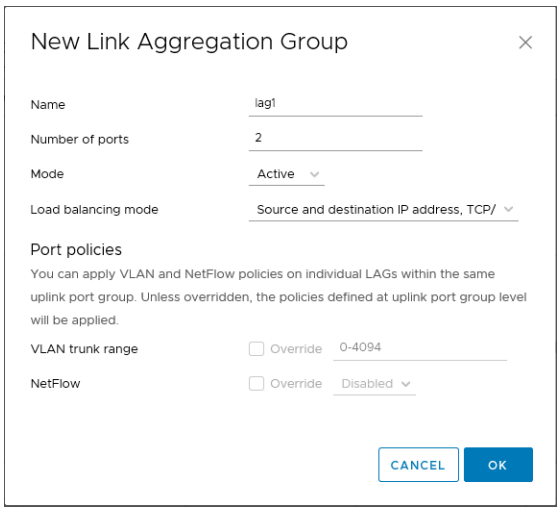
Note Repeat this procedure on all participating switch ports that are connected to vSAN hosts.

Set Up vSphere Distributed Switch

Before you begin, make sure that the vDS is upgraded to a version that supports LACP. To verify, right click the vDS, and check if the Upgrade option is available. You might have to upgrade the vDS to a version that supports LACP.

Create LAG on vDS

To create a LAG on a distributed switch, select the vDS, click the **Configure** tab, and select **LACP**. Add a new LAG.



Configure the LAG with the following properties:

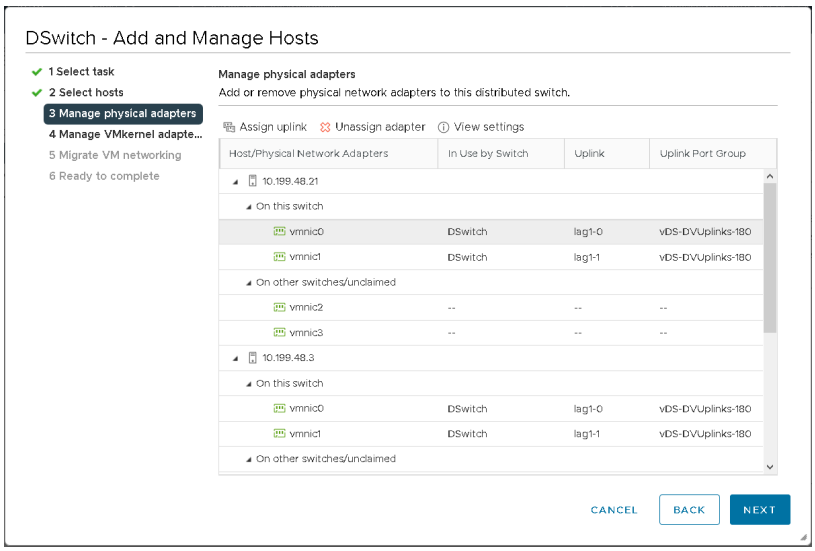
- LAG name: **lag1**
- Number of ports: **2** (to match port channel on switch)
- Mode: **Active**, to match the physical switch.
- Load balancing mode: **Source and destination IP addresses, TCP/UDP port and VLAN**

Add Physical Uplinks to LAG

vSAN hosts have been added to the vDS. Assign the individual vmnics to the appropriate LAG ports.

- Right click the vDS, and select **Add and Manage Hosts…**
- Select **Manage Host Networking**, and add your attached hosts.
- On **Manage Physical Adapters**, select the appropriate adapters and assign them to the LAG port.
- Migrate vmnic0 from Uplink 1 position to port 0 on LAG1.

Repeat the procedure for vmnic1 to the second LAG port position, lag1-1.



Configure Distributed Port Group Teaming and Failover Policy

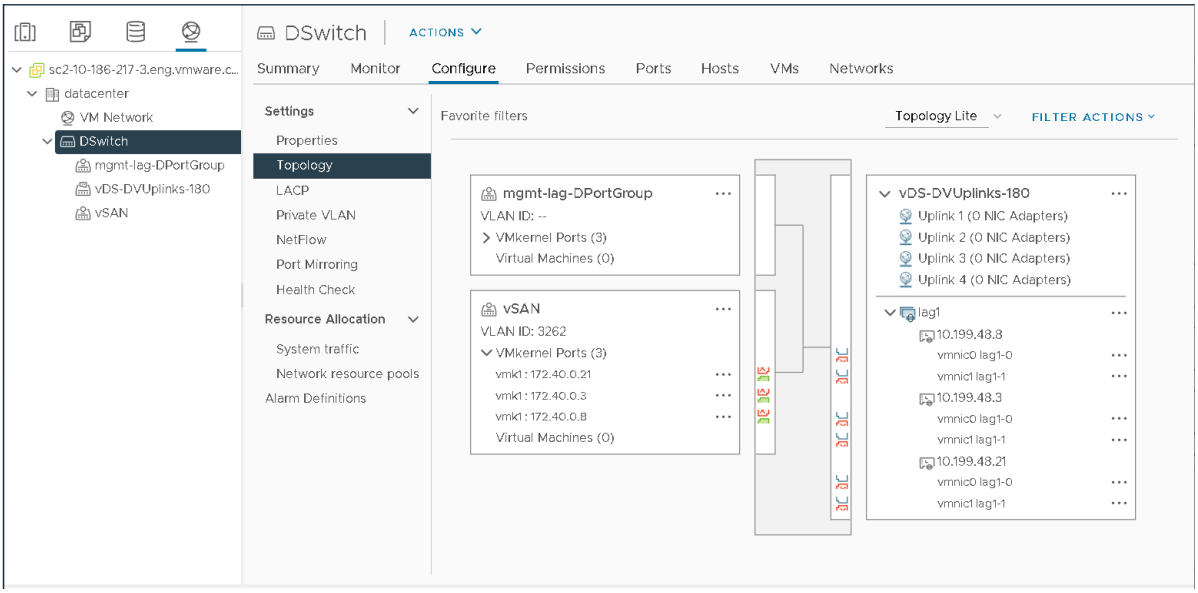
Assign the LAG group as an **Active uplink** on distributed port group teaming and failover policy. Select or create the designated distributed port group for vSAN traffic. This configuration uses a vSAN port group called **vSAN** with VLAN ID 3262 tagged. Edit the port group, and configure Teaming and Failover Policy to reflect the new LAG configuration.

Ensure the LAG group **lag1** is in the active uplinks position, and ensure the remaining uplinks are in the **Unused** position.

Note When a link aggregation group (LAG) is selected as the only active uplink, the load-balancing mode of the LAG overrides the load-balancing mode of the port group. Therefore, the following policy plays no role: **Route based on originating virtual port**.

Create the VMkernel Interfaces

The final step is to create the VMkernel interfaces to use the new distributed port group, ensuring that they are tagged for vSAN traffic. Observe that each vSAN vmknic can communicate over vmnic0 and vmnic1 on a LAG group to provide load balancing and failover.



Configure Load Balancing

From a load balancing perspective, there is not a consistent balance of traffic across all hosts on all vmnics in this LAG setup, but there is more consistency compared to **Route based on physical NIC load** used in Configuration 1 and the air-gapped/multiple vmknics method used in Configuration 2.

The individual hosts' vSphere performance graph shows improved load balancing.

Network Uplink Redundancy Lost

When vmnic1 is not enabled on a given vSAN host, a Network Redundancy alarm is triggered.

No vSAN health alarms are triggered, and the impact to Guest I/O is minimal compared to the air-gapped, multi-vmknics configuration. This configuration does not have to stop any TCP sessions with LACP configured.

Recovery and Failback

In a failback scenario, the behavior differs between Load Based Teaming, multiple vmknics, and LACP in a vSAN environment. After vmnic1 recovers, traffic is automatically balanced across both active uplinks. This behavior can be advantageous for vSAN traffic.

Failback Set to Yes or No?

A LAG load-balancing policy overrides the Teaming and Failover policy for vSphere distributed port groups. Also consider the guidance on Failback value. Lab tests show no discernable behavior differences between Failback set to **yes** or **no** with LACP. LAG settings takes priority over the port-group settings.

Note Network failure detection values remain as **link status only**, since beacon probing is not supported with LACP. See VMware KB Understanding IP Hash load balancing (2006129)

Configuration 4: Static LACP – Route Based on IP Hash

You can use a two-port LACP static port-channel on a switch, and two active uplinks on a vSphere Standard Switch.

In this configuration, use 10Gb networking with two physical uplinks per server. A single VMkernel interface (vmknic) for vSAN exists on each host.

For more information about host requirements and configuration examples, see the following VMware Knowledge Base articles:

- Host requirements for link aggregation for ESXi and ESX (1001938)
- Sample configuration of EtherChannel / Link Aggregation Control Protocol (LACP) with ESXi/ESX and Cisco/HP switches (KB 1004048)

Note vSAN over RDMA does not support this configuration.

Configure the Physical Switch

Configure a two-uplink static port-channel as follows:

- Switch ports 43 and 44
- VLAN trunking, so port-channel is in VLAN trunk mode, with the appropriate VLANs trunked.
- Do not specify the load-balancing policy on the port-channel group.

These steps can be used to configure an individual port-channel on the switch:

Step 1: Create a port-channel.

```
#interface port-channel 13
```

Step 2: Set port-channel to VLAN trunk mode.

```
#switchport mode trunk
```

Step 3: Allow appropriate VLANs.

```
#switchport trunk allowed vlan 3266
```

Step 4: Assign the correct ports to the port-channel and set mode to active.

```
#interface range Te1/0/43, Te1/0/44

#channel-group 1 mode on
```

Step 5: Verify that the port-channel is configured as a static port-channel.

```
#show interfaces port-channel 13


Channel Ports Ch-Type Hash Type Min-links Local Prf

------- ------------------------------ -------- --------- --------- --

Po13 Active: Te1/0/43, Te1/0/44 Static 7 1 Disabled
```

```
Hash Algorithm Type

1 - Source MAC, VLAN, EtherType, source module and port Id

2 - Destination MAC, VLAN, EtherType, source module and port Id

3 - Source IP and source TCP/UDP port

4 - Destination IP and destination TCP/UDP port

5 - Source/Destination MAC, VLAN, EtherType, source MODID/port

6 - Source/Destination IP and source/destination TCP/UDP port

7 - Enhanced hashing mode
```

Configure vSphere Standard Switch

This example assumes you understand the configuration and creation of vSphere Standard Switches.

This example uses the following configuration:

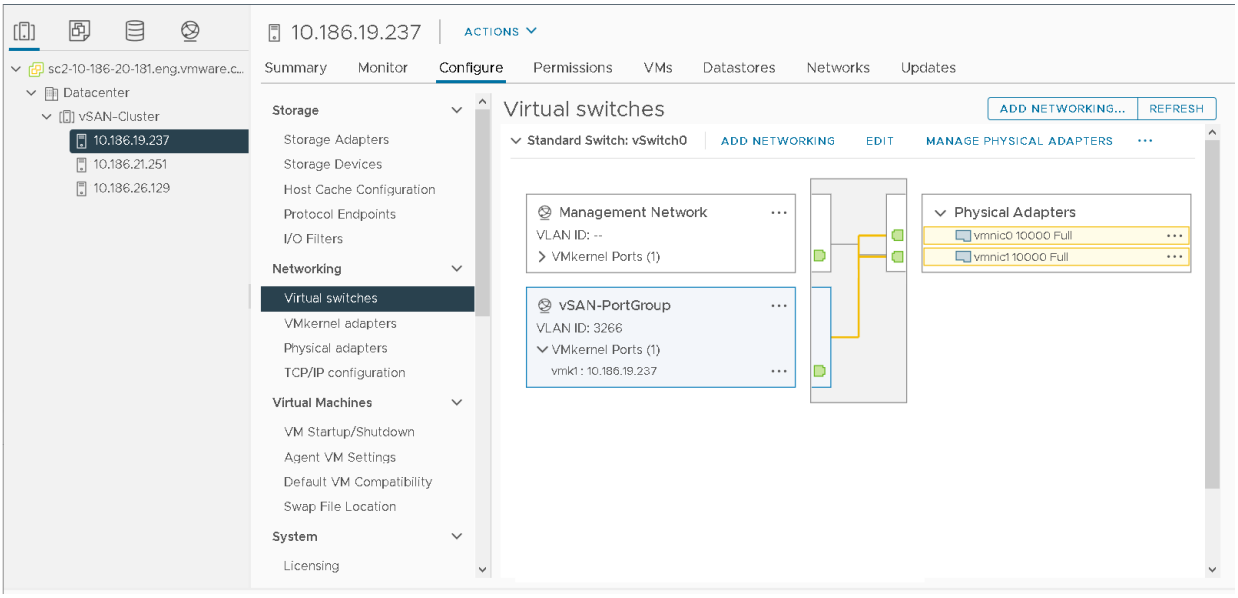
- Identical vSAN hosts
- Uplinks named vmnic0 and vmnic1
- VLAN 3266 trunked to the switch ports and port-channel
- Jumbo frames

On each host, create a **vSwitch1** with MTU set to 9000, and vmnic0 and vmnic1 added to the vSwitch. On the Teaming and Failover Policy, set both adapters to the **Active** position. Set the Load Balancing Policy to **Route Based on IP Hash**.

Configure teaming and failover for the distributed port group for vSAN traffic as follows:

- Load balancing policy set to **Route Based on IP hash**.
- Network failure detection set to **Link status only**.
- Notify Switches set to **Yes**.
- Failback set to **Yes**.
- Ensure both uplinks are in the **Active uplinks** position.

Use defaults for network detection, Notify Switches and Failback. All port groups inherit the Teaming and Failover Policy that was set at the vSwitch level. You can override individual port group teaming and failover polices to differ from the parent vSwitch, but make sure you use the same set of uplinks for IP hash load balancing for all port groups.



Configure Load Balancing

Although both physical uplinks are utilized, there is not a consistent balance of traffic across all physical vmnics. The figure shows that only active traffic is vSAN traffic, which was essentially four vmknics or IP addresses. The behavior might be caused by the low number of IP addresses and possible hashes. However, in some situations, the virtual switch might consistently pass the traffic through one uplink in the team. For further details on the IP Hash algorithm, see the official vSphere documentation about *Route Based on IP Hash.*

Network Redundancy

In this example, vmnic1 is connected to a port that has been disabled from the switch, to focus on failure and redundancy behavior. Note that a network uplink redundancy alarm has triggered.

No vSAN health alarms were triggered. Cluster and VM components are not affected and Guest Storage I/O is not interrupted by this failure.

Recovery and Failback

Once vmnic1 recovers, traffic is automatically balanced across both active uplinks.

Network I/O Control

11

Use vSphere Network I/O Control to set Quality of Service (QoS) levels on network traffic.

vSphere Network I/O Control is a feature available with vSphere Distributed Switches. Use it to implement Quality of Service (QoS) on network traffic. This can be useful for vSAN when vSAN traffic must share the physical NIC with other traffic types, such as vMotion, management, virtual machines.

Reservations, Shares, and Limits

You can set a **reservation** so that Network I/O Control guarantees minimum bandwidth is available on the physical adapter for vSAN.

Reservations can be useful when *bursty* traffic, such as vMotion or full host evacuation, might impact vSAN traffic. Reservations are only invoked if there is contention for network bandwidth. One disadvantage with reservations in Network I/O Control is that unused reservation bandwidth cannot be allocated to virtual machine traffic. The total bandwidth reserved among all system traffic types cannot exceed 75 percent of the bandwidth provided by the physical network adapter with the lowest capacity.

vSAN best practices for reservations. Traffic reserved for vSAN cannot be allocated to virtual machine traffic, so avoid using NIOC reservations in vSAN environments.

Setting **shares** makes a certain bandwidth available to vSAN when the physical adapter assigned for vSAN becomes saturated. This prevents 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 remaining adapters in the team. The **shares** option ensures that no other traffic impacts the vSAN network.

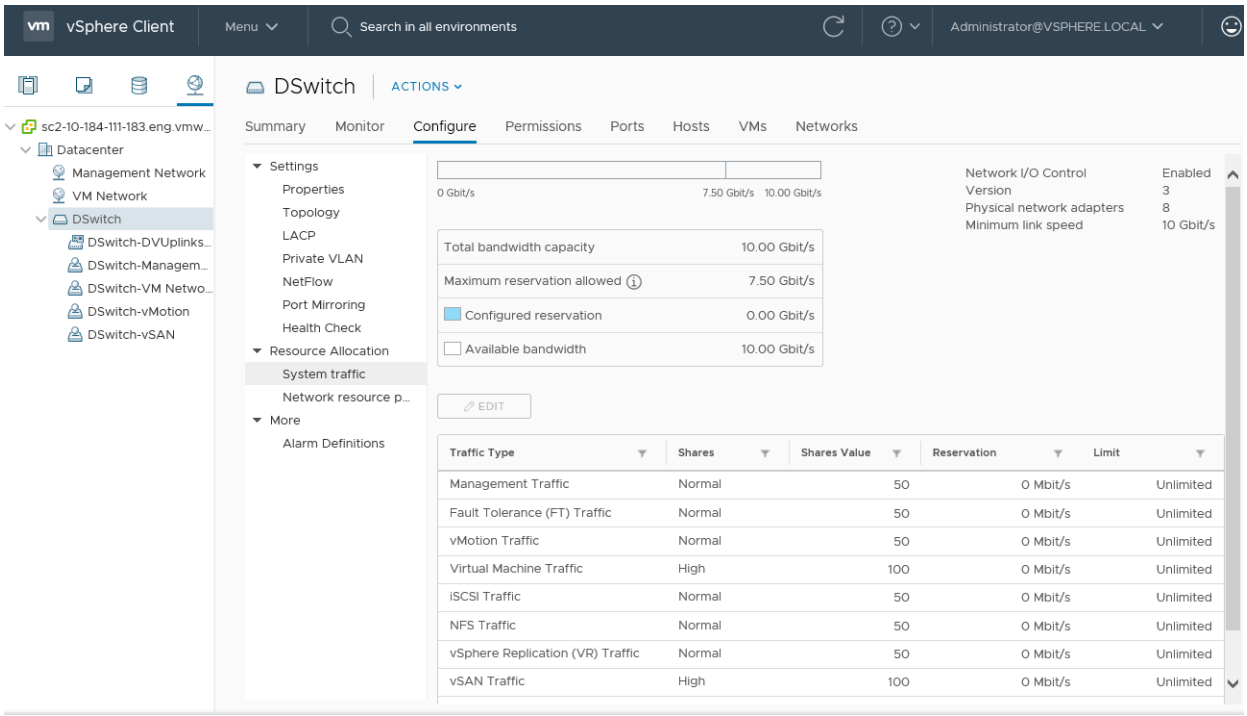
vSAN recommendation on shares. This is the fairest of the bandwidth allocation techniques in NIOC. This technique is preferred for use in vSAN environments.

Setting **limits** defines the maximum bandwidth that a certain traffic type can consume on an adapter. If no one else is using the additional bandwidth, the traffic type with the limit also cannot consume it.

vSAN recommendation on limits. As traffic types with limits cannot consume additional bandwidth, avoid using NIOC limits in vSAN environments.

Network Resource Pools

You can view all system traffic types that can be controlled with Network I/O Control. If you have multiple virtual machine networks, you can assign certain bandwidth to virtual machine traffic. Use network resource pools to consume parts of that bandwidth based on the virtual machine port group.



Enabling Network I/O Control

You can enable Network I/O Control in the configuration properties of the vDS. Right-click the vDS in the vSphere Client, and choose menu **Settings > Edit Settings**.

Note Network I/O Control is only available on vSphere distributed switches, not on standard vSwitches.

You can use Network I/O Control to reserve bandwidth for network traffic based on the capacity of the physical adapters on a host. For example, if vSAN traffic uses 10 GbE physical network adapters, and those adapters are shared with other system traffic types, you can use vSphere Network I/O Control to guarantee a certain amount of bandwidth for vSAN. This can be useful when traffic such as vSphere vMotion, vSphere HA, and virtual machine traffic share the same physical NIC as the vSAN network.

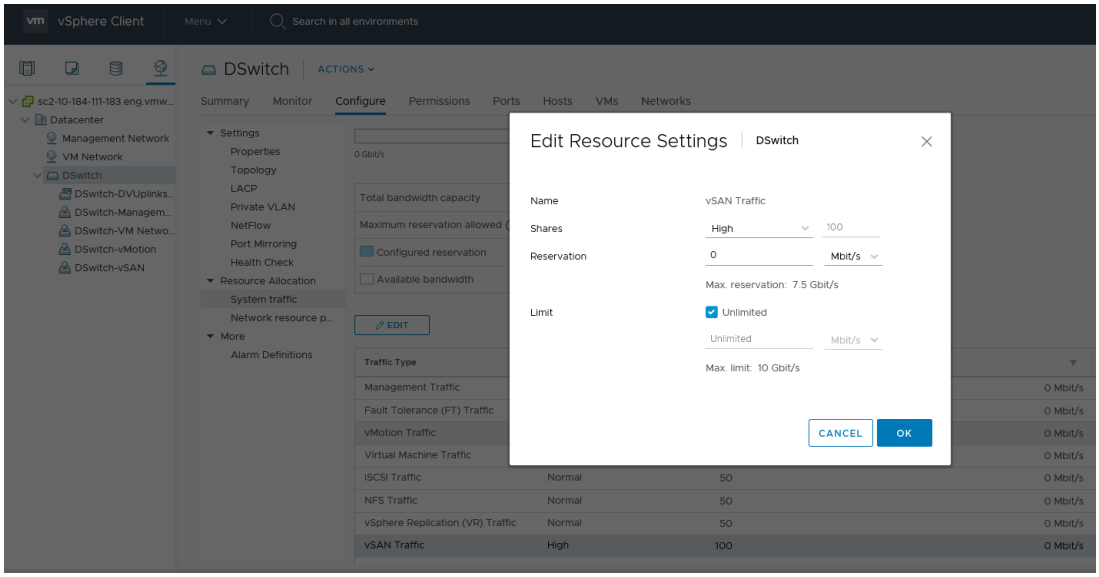
This chapter includes the following topics:

- Network I/O Control Configuration Example

Network I/O Control Configuration Example

You can configure Network I/O Control for a vSAN cluster.

Consider a vSAN cluster with a single 10 GbE physical adapter. This NIC handles traffic for vSAN, vSphere vMotion, and virtual machines. To change the shares value for a traffic type, select that traffic type from the System Traffic view (**VDS > Configure > Resource Allocation > System Traffic**), and click **Edit**. The shares value for vSAN traffic has been changed from the default of Normal/50 to High/100.



Edit the other traffic types to match the share values shown in the table.

Table 11-1. Sample NIOC Settings

Traffic Type	Shares	Value
vSAN	High	100
vSphere vMotion	Low	25
Virtual machine	Normal	50
iSCSI/NFS	Low	25

If the 10 GbE adapter becomes saturated, Network I/O Control allocates 5 Gbps to vSAN on the physical adapter, 3.5 Gbps to virtual machine traffic, and 1.5 Gbps to vMotion. Use these values as a starting point to configure NIOC configuration on your vSAN network. Ensure that vSAN has the highest priority of any protocol.

For more details about the various parameters for bandwidth allocation, see *vSphere Networking* documentation.

With each of the vSphere editions for vSAN, VMware provides a vSphere Distributed Switch as part of the edition. Network I/O Control can be configured with any vSAN edition.

Understanding vSAN Network Topologies

12

vSAN architecture supports different network topologies. These topologies impact on the overall deployment and management of vSAN.

The introduction of unicast support in vSAN 6.6 simplifies the network design.

This chapter includes the following topics:

Standard Deployments

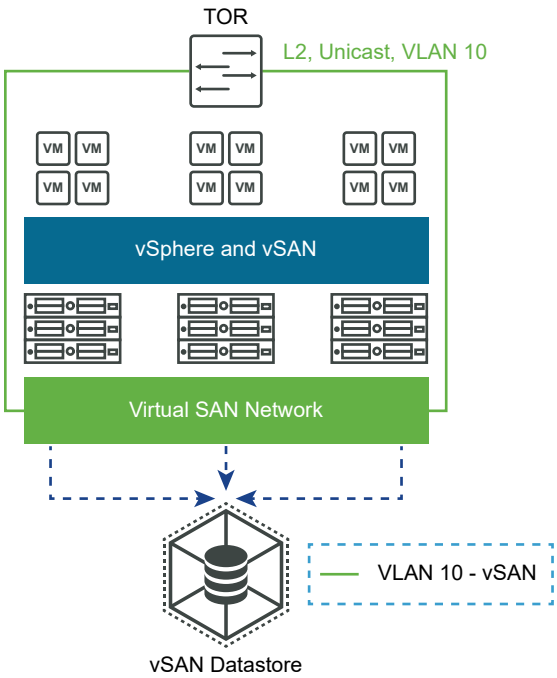
vSAN supports several single-site deployment types.

Layer-2, Single Site, Single Rack

This network topology is responsible for forwarding packets through intermediate Layer 2 devices such as hosts, bridges, or switches.

The Layer 2 network topology offers the simplest implementation and management of vSAN. VMware recommends the use and configuration of IGMP Snooping to avoid sending unnecessary multicast traffic on the network. In this first example, we are looking at a single site, and perhaps even a single rack of servers using vSAN 6.5 or earlier. This version uses multicast, so enable IGMP Snooping. Since everything is on the same L2, you need not configure routing for multicast traffic.

Layer 2 implementations are simplified even further with vSAN 6.6 and later, which introduces unicast support. IGMP Snooping is not required.



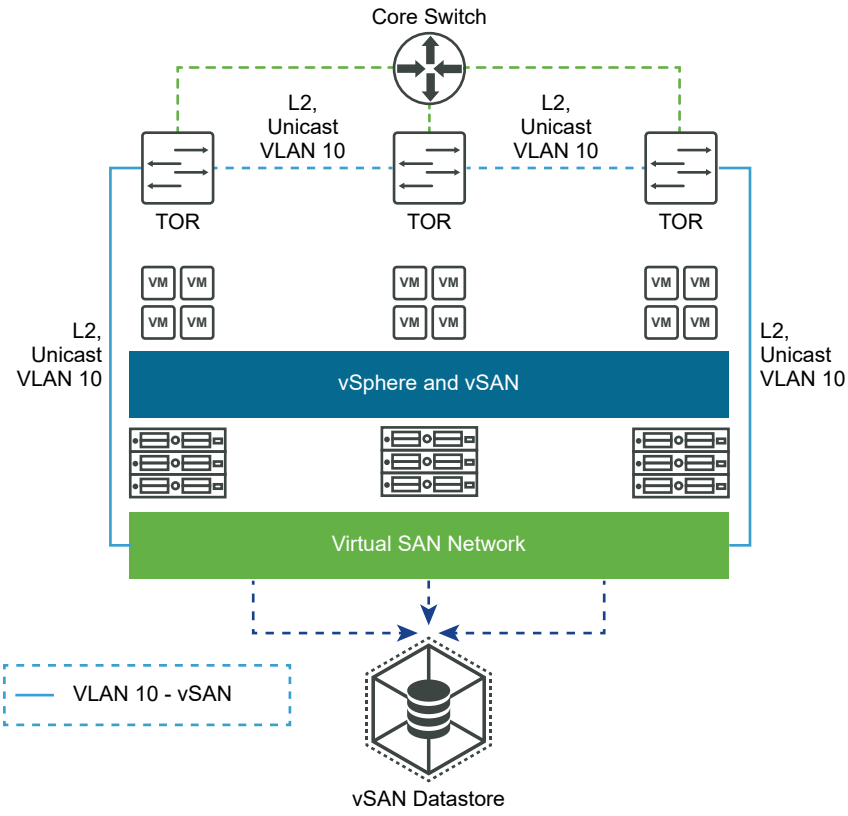
Layer 2, Single Site, Multiple Racks

This network topology works with the Layer 2 implementation where there are multiple racks, and multiple top-of-rack switches, or TORs, connected to a core switch.

In the following figures, the blue dotted line between the TORs shows that the vSAN network is available and accessible to all the hosts in the vSAN cluster. However, the hosts in the different racks communicate to each other over Layer 3, which implies using PIM to route multicast traffic between the hosts. The TORs are not physically connected to each other.

VMware recommends that all TORs are configured for IGMP Snooping, to prevent unnecessary multicast traffic on the network. As there is no routing of the traffic, there is no need to configure PIM to route the multicast traffic.

This implementation is easier in vSAN 6.6 and later, because vSAN traffic is unicast. With unicast traffic, there is no need to configure IGMP Snooping on the switches.



Layer 3, Single Site, Multiple Racks

This network topology works for vSAN deployments where Layer 3 is used to route vSAN traffic.

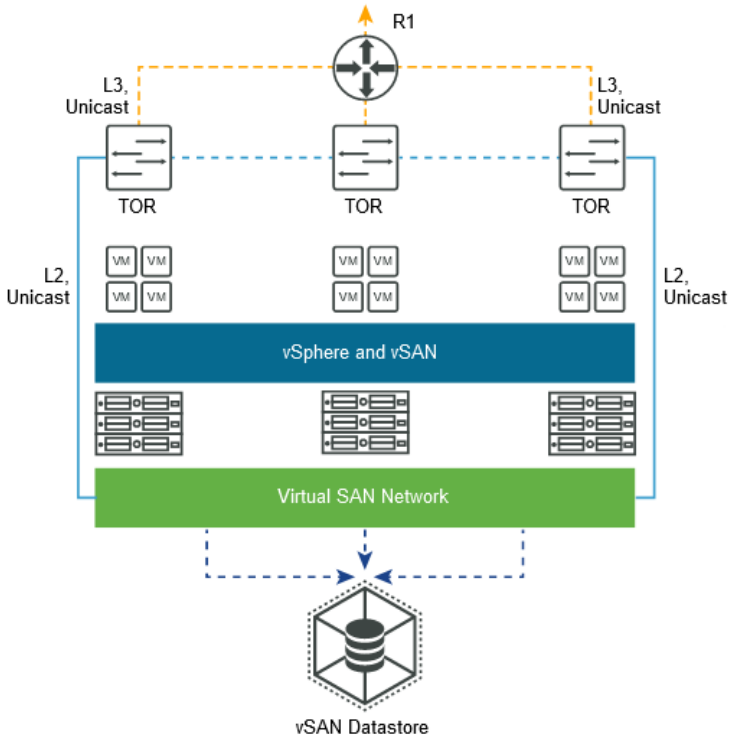
This simple Layer 3 network topology uses multiple racks in the same data center, each with its own TOR switch. Route the vSAN network between the different racks over L3, to allow all the hosts in the vSAN cluster to communicate. Place the vSAN VMkernel ports on different subnets or VLANs, and use a separate subnet or VLAN for each rack.

This network topology routes packets through intermediate Layer 3 capable devices, such as routers and Layer 3 capable switches. Whenever hosts are deployed across different Layer 3 network segments, the result is a routed network topology.

With vSAN 6.5 and earlier, VMware recommends the use and configuration of IGMP Snooping, because these deployments require multicast. Configure PIM on the physical switches to facilitate the routing of the multicast traffic.

vSAN 6.6 and later simplifies this topology. As there is no multicast traffic, there is no need to configure IGMP Snooping. You do not need to configure PIM to route multicast traffic.

Here is an overview of an example vSAN 6.6 deployment over L3. There is no requirement for IGMP Snooping or PIM, because there is no multicast traffic.



Stretched Cluster Deployments

vSAN supports stretched cluster deployments that span two locations.

In vSAN 6.5 and earlier, vSAN traffic between data sites is **multicast** for metadata and **unicast** for I/O.

In vSAN 6.6 and later, all traffic is **unicast**. In all versions of vSAN, the witness traffic between a data site and the witness host is unicast.

Layer 2 Everywhere

You can configure a vSAN stretched cluster in a Layer 2 network, but this configuration is not recommended.

Consider a design where the vSAN stretched cluster is configured in one large Layer 2 design. Data Site 1 and Site 2 are where the virtual machines are deployed. Site 3 contains the witness host.

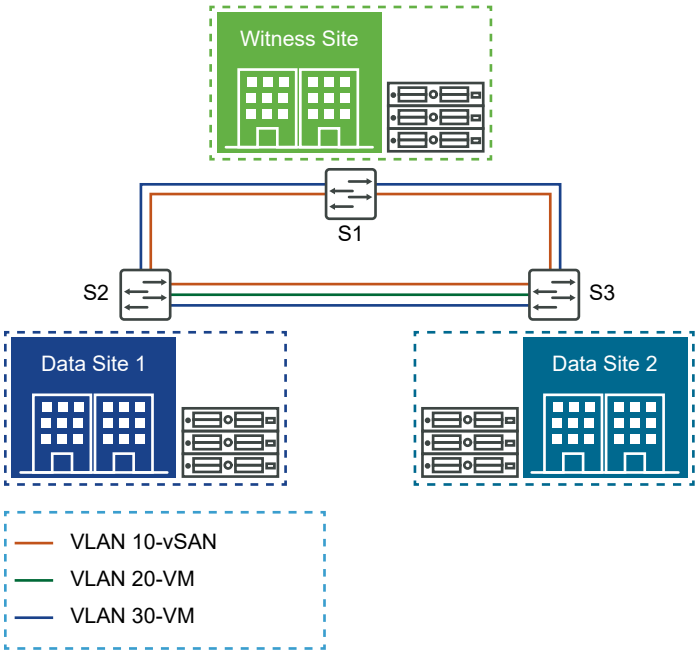
Note For best results, do not use a stretched Layer 2 network across all sites.

To demonstrate Layer 2 everywhere as simply as possible, we use switches (and not routers) in the topologies.

Layer 2 networks cannot have any loops (multiple paths), so features such as Spanning Tree Protocol (STP) are needed to block one of the connections between Site 1 and Site 2. Now consider a situation where the link between Site 2 and Site 3 is broken (the link between Site 1 and Site 2). Network traffic can be switched from Site 1 to Site 2 through the witness host at Site 3. As VMware supports a much lower bandwidth and higher latency for the witness host, you see a significant decrease in performance if data network traffic passes through a lower specification witness site.

If switching traffic between data sites through the witness site does not impact latency of applications, and bandwidth is acceptable, a stretched L2 configuration between sites is possible. In most cases, such a configuration is not feasible, and adds complexity to the networking requirements.

With vSAN 6.5 or earlier, which uses multicast traffic, you must configure IGMP snooping on the switches. This is not necessary with vSAN 6.6 and later. PIM is not necessary because there is no routing of multicast traffic.



Supported Stretched Cluster Configurations

vSAN supports stretched cluster configurations.

The following configuration prevent traffic from Site 1 being routed to Site 2 through the witness host, in the event of a failure on either of the data sites' network. This configuration avoids performance degradation. To ensure that data traffic is not switched through the witness host, use the following network topology.

Between Site 1 and Site 2, implement a stretched Layer 2 switched configuration or a Layer 3 routed configuration. Both configurations are supported.

Between Site 1 and the witness host, implement a Layer 3 routed configuration.

Between Site 2 and the witness host, implement a Layer 3 routed configuration.

These configurations (L2+L3, and L3 everywhere) are described with considerations given to multicast in vSAN 6.5 and earlier, and unicast only, which is available in vSAN 6.6. Multicast traffic introduces additional configuration steps for IGMP snooping, and PIM for routing multicast traffic.

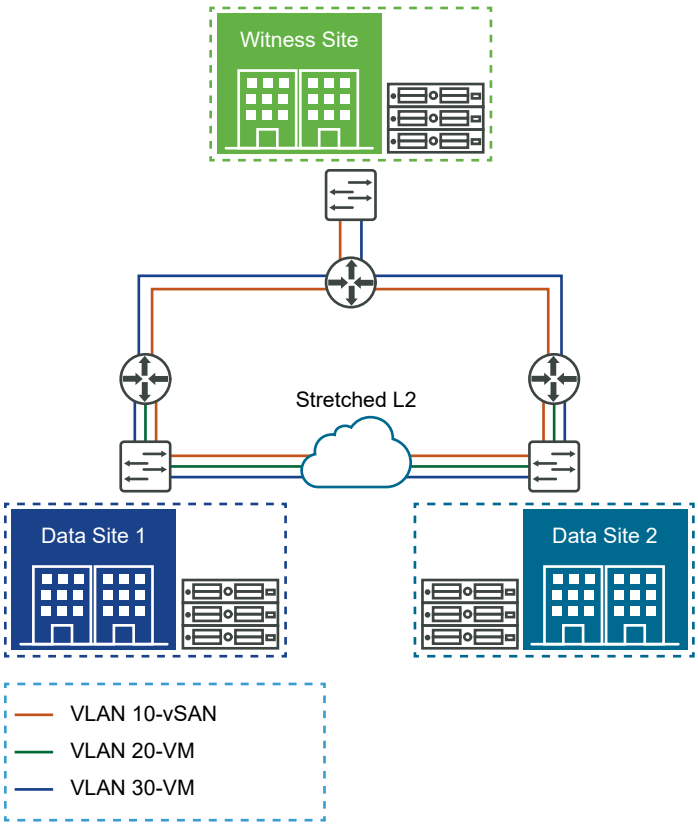
We shall examine a stretched Layer 2 network between the data sites and a Layer 3 routed network to the witness site. To demonstrate a combination of Layer 2 and Layer 3 as simply as possible, use a combination of switches and routers in the topologies.

Stretched Layer 2 Between Data Sites, Layer 3 to Witness Host

vSAN supports stretched Layer 2 configurations between data sites.

The only traffic that is routed in this case is the witness traffic. With vSAN 6.5 and earlier, which uses multicast, use IGMP snooping for the multicast traffic on the stretched L2 vSAN between data sites. However, since the witness traffic is unicast, there is no need to implement PIM on the Layer 3 segments.

With vSAN 6.6, which uses unicast, there is no requirement to consider IGMP snooping or PIM.



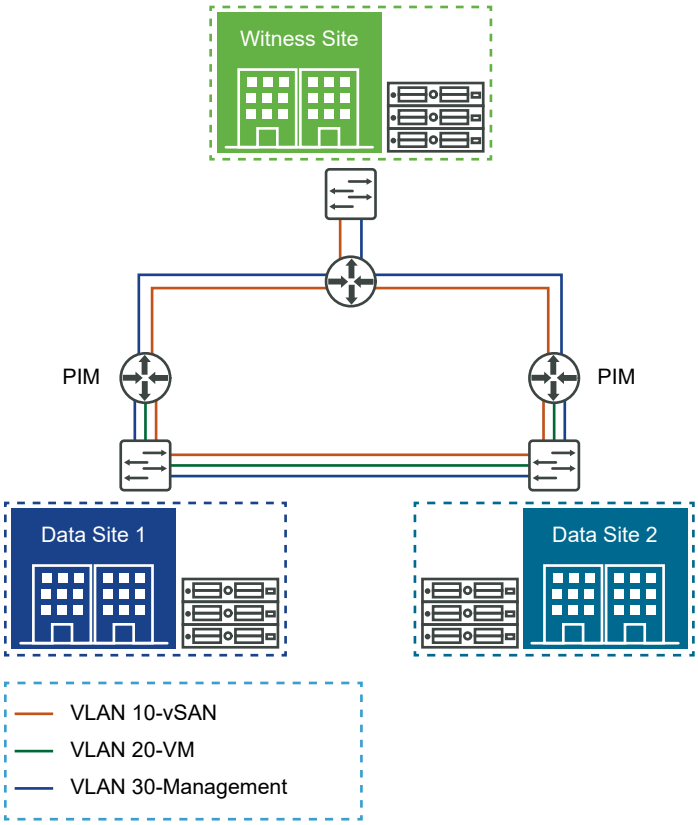
Layer 3 Everywhere

In this vSAN stretched cluster configuration, the data traffic is routed between the data sites and the witness host.

To implement Layer 3 everywhere as simply as possible, use routers or routing switches in the topologies.

For example, consider an environment with vSAN 6.5 or earlier, which uses multicast traffic. In this case, configure IGMP snooping on the data site switches to manage the amount of multicast traffic on the network. This is unnecessary at the witness host since witness traffic is unicast. The multicast traffic is routed between the data sites, so configure PIM to allow multicast routing.

With vSAN 6.6 and later, neither IGMP snooping nor PIM are needed because all the routed traffic is unicast.



Separating Witness Traffic on vSAN Stretched Clusters

vSAN supports separating witness traffic on stretched clusters.

Inn vSAN 6.5 and later releases, you can separate witness traffic from vSAN traffic in two-node configurations. This means that the two vSAN hosts can be directly connected without a 10 Gb switch.

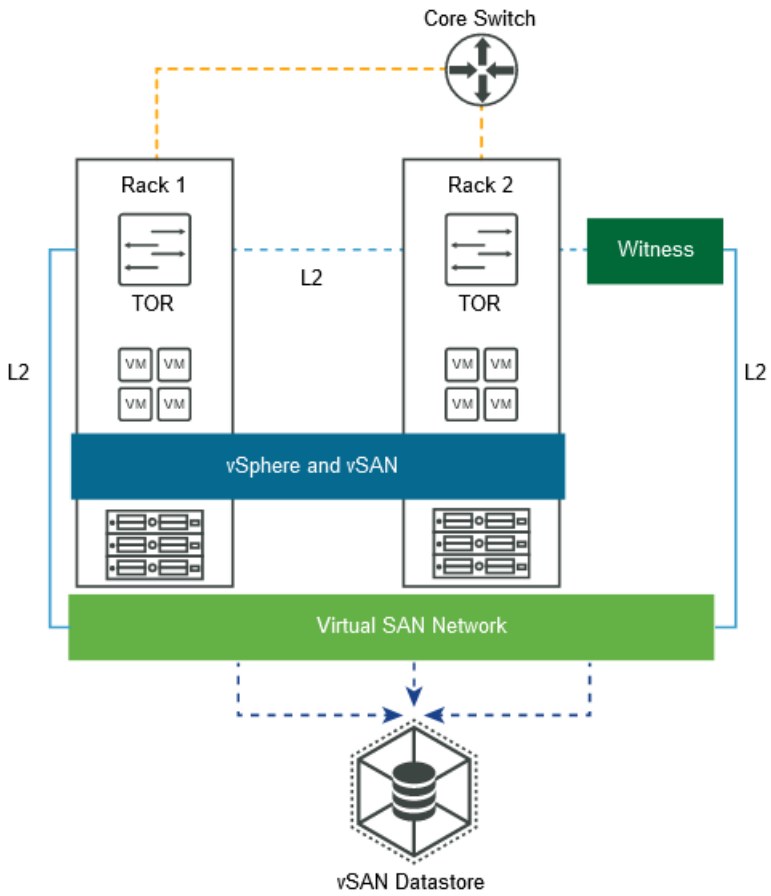
This witness traffic separation is only supported on two-node deployments in vSAN 6.6. Separating the witness traffic on vSAN stretched clusters is supported in vSAN 6.7 and later.

Using Stretched Cluster to Achieve Rack Awareness

With stretched clusters, vSAN provides rack awareness in a single site.

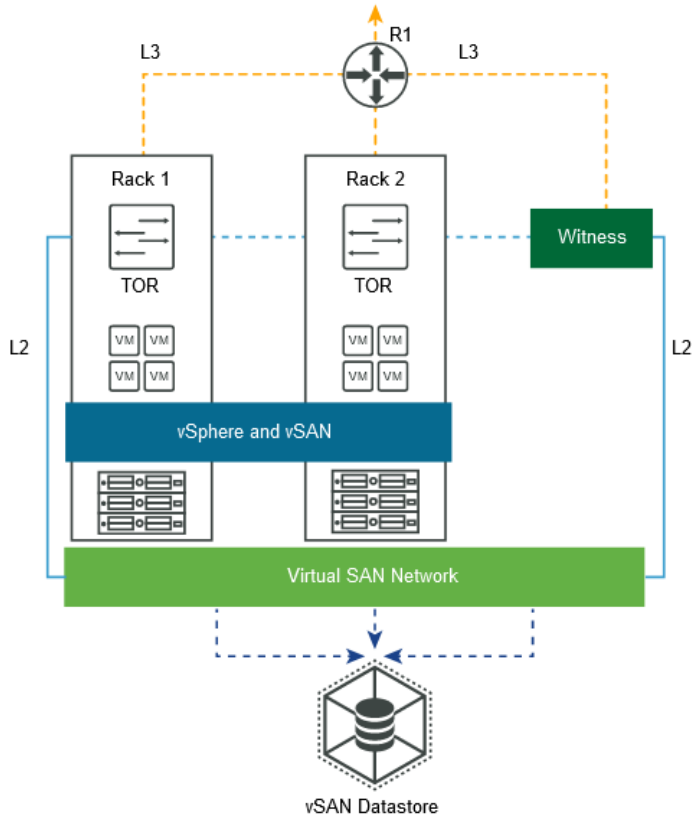
If you have two racks of vSAN hosts, you can continue to run your vSAN cluster after a complete rack failure. In this case, availability of the VM workloads is provided by the remaining rack and a remote witness host.

Note For this configuration to be supported, do not place the witness host within the two racks of vSAN hosts.



In this example, if rack 1 fails, rack 2 and the witness host provide VM availability. This configuration is a pre-vSAN 6.6 environment, and needs multicast configured on the network. The witness host must be on the vSAN network. Witness traffic is unicast. In vSAN 6.6 and later, all traffic is unicast.

This topology is also supported over L3. Place the vSAN VMkernel ports on different subnets or VLANs, and use a separate subnet or VLAN for each rack.



This topology supports deployments with two racks to achieve rack awareness (fault domains) with a vSAN stretched cluster. This solution uses a witness host that is external to the cluster.

Two-Node vSAN Deployments

vSAN supports two-node deployments. Two-node vSAN deployments are used for remote offices/branch offices (ROBO) that have a small number of workloads, but require high availability.

vSAN two-node deployments use a third witness host, which can be located remotely from the branch office. Often the witness is maintained in the branch office, along with the management components, such as the vCenter Server.

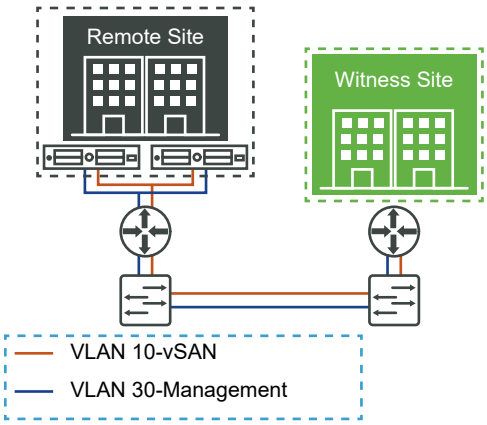
Two-Node vSAN Deployments Earlier than vSAN 6.5

vSAN releases earlier than 6.5 that support two-node deployments require a physical switch at the remote site.

Early two-node vSAN have a requirement to include a physical 10 Gb switch at the remote site. If the only servers at this remote site were the vSAN hosts, this could be an inefficient solution.

With this deployment, if there are no other devices using the 10 Gb switch, then no consideration needs to be given to IGMP snooping. If other devices at the remote site share the 10 Gb switch, use IGMP snooping to prevent excessive and unnecessary multicast traffic.

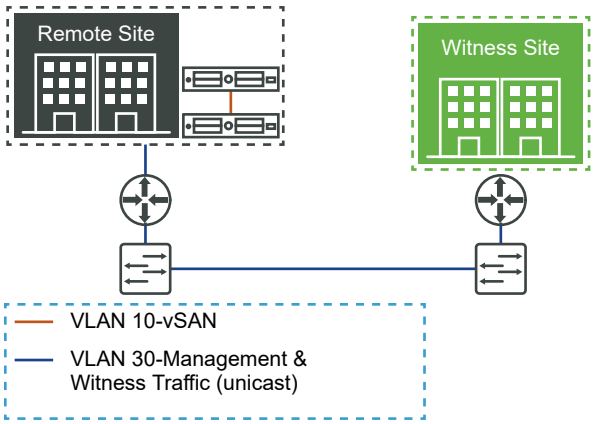
PIM is not required because the only routed traffic is witness traffic, which is unicast.



Two-Node Deployments for vSAN 6.5 and Later

vSAN 6.5 and later supports two-node deployments.

With vSAN version 6.5 and later, this two-node vSAN implementation is much simpler. vSAN 6.5 and later allows the two hosts at the data site to be directly connected.

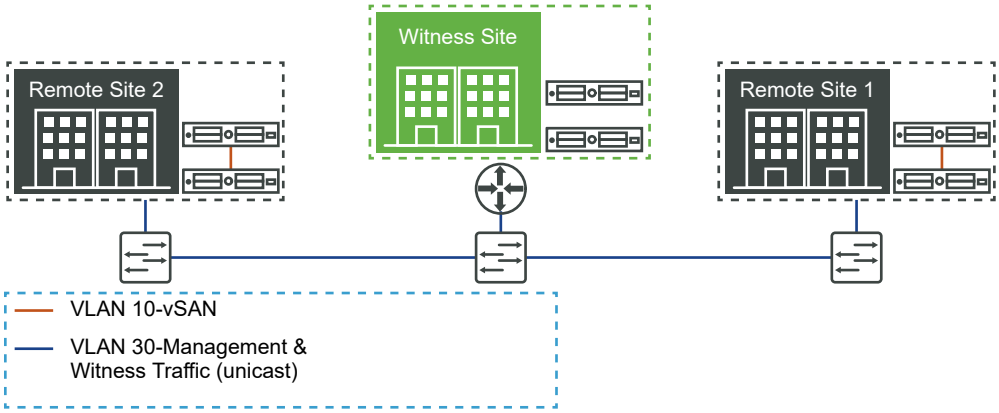


To enable this functionality, the witness traffic is separated completely from the vSAN data traffic. The vSAN data traffic can flow between the two nodes on the direct connect, while the witness traffic can be routed to the witness site over the management network.

The witness appliance can be located remotely from the branch office. For example, the witness might be running back in the main data center, alongside the management infrastructure (vCenter Server, vROps, Log Insight, and so on). Another supported place where the witness can reside remotely from the branch office is in vCloud Air.

In this configuration, there is no switch at the remote site. As a result, there is no need configure support for multicast traffic on the vSAN back-to-back network. You do not need to consider multicast on the management network because the witness traffic is unicast.

vSAN 6.6 and later uses all uncast, so there are no multicast considerations. Multiple remote office/branch office two-node deployment are also supported, so long as each has their own unique witness.



Common Considerations for Two-Node vSAN Deployments

Two-node vSAN deployments provide support to other topologies. This section describes common configurations.

For more information about two-node configurations and detailed deployment considerations outside of network, see the vSAN core documentation.

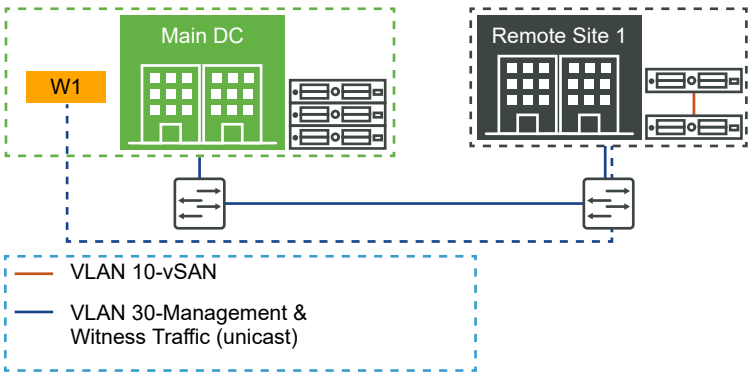
Running the Witness on Another Two-Node Cluster

vSAN does not support running the witness on another two-node cluster.

Witness Running on Another Standard vSAN Deployment

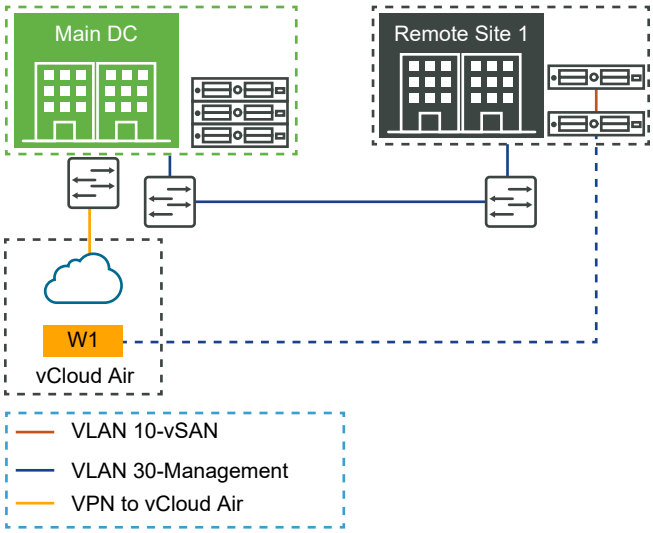
vSAN supports witness running on another standard vSAN deployment.

This configuration is supported. Any failure on the two-node vSAN at the remote site does not impact the availability of the standard vSAN environment at the main data center.



Witness Running in vCloud Air

vSAN allows you to run a witness in vCloud Air.



Configuration of Network from Data Sites to Witness Host

The host interfaces in the data sites communicate to the witness host over the vSAN network. There are different configuration options available.

This topic discusses how to implement these configurations. It addresses how the interfaces on the hosts in the data sites, which communicate to each other over the vSAN network, communicate with the witness host.

Option 1: Physical ESXi Witness Connected over L3 with Static Routes

The data sites can be connected over a stretched L2 network. Use this also for the data sites' management network, vSAN network, vMotion network and virtual machine network.

The physical network router in this network infrastructure does not automatically transfer traffic from the hosts in the data sites (site 1 and site 2) to the host in the witness site (site 3). In order to configure the vSAN stretched cluster successfully, all hosts in the cluster must communicate. It is possible to deploy a stretched cluster in this environment.

The solution is to use *static routes* configured on the ESXi hosts, so that the vSAN traffic from site1 and site 2 can reach the witness host in site 3. In the case of the ESXi hosts on the data sites, add a static route to the vSAN interface, which redirects traffic to the witness host on site 3 over a specified gateway for that network. In the case of the witness host, the vSAN interface must have a static route added, which redirects vSAN traffic destined for the hosts in the data sites. Use the following command to add a static route on each ESXi host in the stretched cluster:

`esxcli network ip route ipv4 add -g` *<gateway>* **`-n`** *<network>*

Note The vCenter Server must be able to manage the ESXi hosts at both the data sites and the witness site. As long as there is direct connectivity from the witness host to vCenter Server, there are no additional concerns regarding the management network.

There is no need to configure a vMotion network or a VM network, or add any static routes for these networks in the context of a vSAN stretched cluster. Virtual machines are never migrated or deployed to the vSAN witness host. Its purpose is to maintain witness objects only, and does not require either of these networks for this task.

Option 2: Virtual ESXi Witness Appliance Connected over L3 with Static Routes

Since the witness host is a virtual machine that gets deployed on a physical ESXi host, which is not part of the vSAN cluster, that physical ESXi host must have a minimum of one VM network pre-configured. This VM network must reach both the management network and the vSAN network shared by the ESXi hosts on the data sites.

Note The witness host does not need to be a dedicated host. It can be used for many other VM workloads, while simultaneously hosting the witness.

An alternative option is to have two preconfigured VM networks on the underlying physical ESXi host, one for the management network and one for the vSAN network. When the virtual ESXi witness is deployed on this physical ESXi host, the network needs to be attached and configured accordingly.

Once you have deployed the virtual ESXi witness host, configure the static route. Assume that the data sites are connected over a stretched L2 network. Use this also for the data sites' management network, vSAN network, vMotion network, and virtual machine network. vSAN traffic is not routed from the hosts in the data sites (site 1 and site 2) to the host in the witness site (site 3) over the default gateway. In order to configure the vSAN stretched cluster successfully, all hosts in the cluster require static routes, so that the vSAN traffic from site 1 and site 2 can reach the witness host in site 3. Use the **`esxcli network ip route`** command to add a static route on each ESXi host.

Corner Case Deployments

It is possible to deploy vSAN in unusual, or corner-case configurations.

These unusual topologies require special considerations.

Three Locations, No Stretched Cluster, Distributed Witness Hosts

You can deploy vSAN across multiple rooms, buildings or sites, rather than deploy a stretched cluster configuration.

This configuration is supported. The one requirement is that the latency between the sites must be at the same level as the latency expected for a normal vSAN deployment in the same data center. The latency must be **<1ms** between all hosts. If latency is greater than this value, consider a stretched cluster which tolerates latency of 5ms. With vSAN 6.5 or earlier, additional considerations for multicast must be addressed.

For best results, maintain a uniform configuration across all sites in such a topology. To maintain availability of VMs, configure fault domains, where the hosts in each room, building, or site are placed in the same fault domain. Avoid asymmetric partitioning of the cluster, where host A cannot communicate to host B, but host B can communicate to host A.

Two-Node Deployed as 1+1+W Stretched Cluster

You can deploy a two-node configuration as a stretched cluster configuration, placing each host in different rooms, buildings, or sites.

Attempt to increase the number of hosts at each site fail with an error related to licensing. For any cluster that is larger than two hosts and that uses the dedicated witness appliance/host feature (N+N+Witness, where N>1), the configuration is considered a vSAN stretched cluster.

Troubleshooting the vSAN Network

13

vSAN allows you to examine and troubleshoot the different types of issues that arise from a misconfigured vSAN network.

vSAN operations depend on the network configuration, reliability, and performance. Many support requests stem from an incorrect network configuration, or the network not performing as expected.

Use the vSAN health service to resolve network issues. Network health checks can direct you to an appropriate Knowledge Base article, depending on the results of the health check. The Knowledge Base article provides instructions to solve the network problem.

Network Health Checks

The health service includes a category for networking health checks.

Each health check has an **Ask VMware** link. If a health check fails, click **Ask VMware** and read the associated VMware Knowledge Base article for further details, and guidance on how to address the issue at hand.

The following networking health checks provide useful information about your vSAN environment.

- vSAN**: Basic (unicast) connectivity check**. This check verifies that IP connectivity exists among all ESXi hosts in the vSAN cluster, by pinging each ESXi host on the vSAN network from each other ESXi host.
- **vMotion: Basic (unicast) connectivity check**. This check verifies that IP connectivity exists among all ESXi hosts in the vSAN cluster that have vMotion configured. Each ESXi host on the vMotion network pings all other ESXi hosts.
- **All hosts have a vSAN vmknic configured**. This check ensures each ESXi host in the vSAN cluster has a VMkernel NIC configured for vSAN traffic.
- All hosts have matching multicast settings. This check ensures that each hosts have a properly configured multicast address.
- **All hosts have matching subnets**. This check tests that all ESXi hosts in a vSAN cluster have been configured so that all vSAN VMkernel NICs are on the same IP subnet.

- **Hosts disconnected from VC**. This check verifies that the vCenter Server has an active connection to all ESXi hosts in the vSAN cluster.
- **Hosts with connectivity issues**. This check refers to situations where vCenter Server lists the host as connected, but API calls from vCenter to the host are failing. It can highlight connectivity issues between a host and the vCenter Server.
- **Network latency**. This check performs a network latency check of vSAN hosts. If the threshold exceeds 100 ms, a warning is displayed. If the latency threshold exceeds 200 ms, and error is raised.
- **vMotion: MTU checks (ping with large packet size)**. This check complements the basic vMotion ping connectivity check. Maximum Transmission Unit size is increased to improve network performance. Incorrectly configured MTUs might not appear as a network configuration issue, but can cause performance issues.
- **vSAN cluster partition**. This health check examines the cluster to see how many partitions exist. It displays an error if there is more than a single partition in the vSAN cluster.
- **Multicast assessment based on other checks**. This health check aggregates data from all network health checks. If this check fails, it indicates that multicast is likely the root cause of a network partition.

Commands to Check the Network

When the vSAN network has been configured, use these commands to check its state. You can check which VMkernel Adapter (vmknic) is used for vSAN, and what attributes it contains.

Use ESXCLI and RVC commands to verify that the network is fully functional, and to troubleshoot any network issues with vSAN.

You can verify that the vmknic used for the vSAN network is uniformly configured correctly across all hosts, check that multicast is functional, and verify that hosts participating in the vSAN cluster can successfully communicate with one another.

esxcli vsan network list

This command enables you to identify the VMkernel interface used by the vSAN network.

The output below shows that the vSAN network is using vmk2. This command continues to work even if vSAN has been turned off and the hosts no longer participate in vSAN.

The Agent Group Multicast and Master Group Multicast are also important to check.

```
[root@esxi-dell-m:~] esxcli vsan network list
Interface
   VmkNic Name: vmk1
   IP Protocol: IP
   Interface UUID: 32efc758-9ca0-57b9-c7e3-246e962c24d0
   Agent Group Multicast Address: 224.2.3.4
   Agent Group IPv6 Multicast Address: ff19::2:3:4
```

```
   Agent Group Multicast Port: 23451
   Master Group Multicast Address: 224.1.2.3
   Master Group IPv6 Multicast Address: ff19::1:2:3
   Master Group Multicast Port: 12345
   Host Unicast Channel Bound Port: 12321
   Multicast TTL: 5
   Traffic Type: vsan
```

This provides useful information, such as which VMkernel interface is being used for vSAN traffic. In this case, it is **vmk1**. However, also shown are the multicast addresses. This information might be displayed even when the cluster us running in unicast mode. There is the group multicast address and port. Port 23451 is used for the heartbeat, sent every second by the primary, and is visible on every other host in the cluster. Port 12345 is used for the CMMDS updates between the primary and backup.

esxcli network ip interface list

This command enables you to verify items such as vSwitch or distributed switch.

Use this command to check which vSwitch or distributed switch that it is attached to, and the MTU size, which can be useful if jumbo frames have been configured in the environment. In this case, MTU is at the default of 1500.

```
[root@esxi-dell-m:~] esxcli network ip interface list
vmk0
   Name: vmk0
   <<truncated>>
vmk1
   Name: vmk1
   MAC Address: 00:50:56:69:96:f0
   Enabled: true
   Portset: DvsPortset-0
   Portgroup: N/A
   Netstack Instance: defaultTcpipStack
   VDS Name: vDS
   VDS UUID: 50 1e 5b ad e3 b4 af 25-18 f3 1c 4c fa 98 3d bb
   VDS Port: 16
   VDS Connection: 1123658315
   Opaque Network ID: N/A
   Opaque Network Type: N/A
   External ID: N/A
   MTU: 9000
   TSO MSS: 65535
   Port ID: 50331814
```

The Maximum Transmission Unit size is shown as 9000, so this VMkernel port is configured for jumbo frames, which require an MTU of about 9,000. VMware does not make any recommendation around the use of jumbo frames. However, jumbo frames are supported for use with vSAN.

esxcli network ip interface ipv4 get –i vmk2

This command displays information such as IP address and netmask of the vSAN VMkernal interface.

With this information, an administrator can now begin to use other commands available at the command line to check that the vSAN network is working correctly.

```
[root@esxi-dell-m:~] esxcli network ip interface ipv4 get -i vmk1
Name  IPv4 Address  IPv4 Netmask   IPv4 Broadcast  Address Type  Gateway  DHCP DNS
----  ------------  -------------  --------------  ------------  -------  --------
vmk1  172.40.0.9   255.255.255.0   172.40.0.255   STATIC         0.0.0.0 false
```

vmkping

The `vmkping` command verifies whether all the other ESXi hosts on the network are responding to your ping requests.

```
~ # vmkping -I vmk2 172.32.0.3 -s 1472 -d
 PING 172.32.0.3 (172.32.0.3): 56 data bytes
 64 bytes from 172.32.0.3: icmp_seq=0 ttl=64 time=0.186 ms
 64 bytes from 172.32.0.3: icmp_seq=1 ttl=64 time=2.690 ms
 64 bytes from 172.32.0.3: icmp_seq=2 ttl=64 time=0.139 ms

 --- 172.32.0.3 ping statistics ---
 3 packets transmitted, 3 packets received, 0% packet loss
 round-trip min/avg/max = 0.139/1.005/2.690 ms
```

While it does not verify multicast functionality, it can help identify a rogue ESXi host that has network issues. You can also examine the response times to see if there is any abnormal latency on the vSAN network.

If jumbo frames are configured, this command does not report any issues if the jumbo frame MTU size is incorrect. By default, this command uses an MTU size of 1500. If there is a need to verify if jumbo frames are successfully working end-to-end, use vmkping with a larger packet size (-s) option as follows:

```
~ # vmkping -I vmk2 172.32.0.3 -s 8972 -d
PING 172.32.0.3 (172.32.0.3): 8972 data bytes
9008 bytes from 172.32.0.3: icmp_seq=0 ttl=64 time=0.554 ms
9008 bytes from 172.32.0.3: icmp_seq=1 ttl=64 time=0.638 ms
9008 bytes from 172.32.0.3: icmp_seq=2 ttl=64 time=0.533 ms

--- 172.32.0.3 ping statistics ---
3 packets transmitted, 3 packets received, 0% packet loss
round-trip min/avg/max = 0.533/0.575/0.638 ms
~ #
```

Consider adding -d to the vmkping command to test if packets can be sent without fragmentation.

esxcli network ip neighbor list

This command helps to verify if all vSAN hosts are on the same network segment.

In this configuration, we have a four-host cluster, and this command returns the ARP (Address Resolution Protocol) entries of the other three hosts, including their IP addresses and their vmknic (vSAN is configured to use vmk1 on all hosts in this cluster).

```
[root@esxi-dell-m:~] esxcli network ip neighbor list -i vmk1
Neighbor     Mac Address        Vmknic   Expiry  State  Type
-----------  -----------------  ------  -------  -----  -------
172.40.0.12  00:50:56:61:ce:22  vmk1    164 sec         Unknown
172.40.0.10  00:50:56:67:1d:b2  vmk1    338 sec         Unknown
172.40.0.11  00:50:56:6c:fe:c5  vmk1    162 sec         Unknown
[root@esxi-dell-m:~]
```

esxcli network diag ping

This command checks for duplicates on the network, and round-trip times.

To get even more detail regarding the vSAN network connectivity between the various hosts, ESXCLI provides a powerful network diagnostic command. Here is an example of one such output, where the VMkernel interface is on vmk1 and the remote vSAN network IP of another host on the network is 172.40.0.10

```
[root@esxi-dell-m:~] esxcli network diag ping -I vmk1 -H 172.40.0.10
   Trace:
         Received Bytes: 64
         Host: 172.40.0.10
         ICMP Seq: 0
         TTL: 64
         Round-trip Time: 1864 us
         Dup: false
         Detail:

         Received Bytes: 64
         Host: 172.40.0.10
         ICMP Seq: 1
         TTL: 64
         Round-trip Time: 1834 us
         Dup: false
         Detail:

         Received Bytes: 64
         Host: 172.40.0.10
         ICMP Seq: 2
         TTL: 64
         Round-trip Time: 1824 us
         Dup: false
         Detail:
   Summary:
         Host Addr: 172.40.0.10
```

```
          Transmitted: 3
          Recieved: 3
          Duplicated: 0
          Packet Lost: 0
          Round-trip Min: 1824 us
          Round-trip Avg: 1840 us
          Round-trip Max: 1864 us
[root@esxi-dell-m:~]
```

vsan.lldpnetmap

This RVC command displays uplink port information.

If there are non-Cisco switches with Link Layer Discovery Protocol (LLDP) enabled in the environment, there is an RVC command to display uplink <-> switch <-> switch port information. For more information on RVC, refer to the RVC Command Guide.

This helps you determine which hosts are attached to which switches when the vSAN cluster is spanning multiple switches. It can help isolate a problem to a particular switch when only a subset of the hosts in the cluster is impacted.

```
> vsan.lldpnetmap 02013-08-15 19:34:18 -0700: This operation will take
30-60 seconds ...+---------------+---------------------------+| Host          | LLDP
info                 |+---------------+---------------------------+| 10.143.188.54 | w2r13-
vsan-x650-2: vmnic7 ||               | w2r13-vsan-x650-1: vmnic5 |+---------------
+---------------------------+
```

This is only available with switches that support LLDP. To configure it, log in to the switch and run the following:

```
switch# config t
Switch(Config)# feature lldp
```

To verify that LLDP is enabled:

```
switch(config)#do show running-config lldp
```

Note LLDP operates in both send and receive mode, by default. Check the settings of your vDS properties if the physical switch information is not being discovered. By default, vDS is created with discovery protocol set to CDP, Cisco Discovery Protocol. To resolve this, set the discovery protocol to LLDP, and set operation to **both** on the vDS.

Checking Multicast Communications

Multicast configurations can cause issues for initial vSAN deployment.

One of the simplest ways to verify if multicast is working correctly in your vSAN environment is by using the `tcpdump-uw` command. This command is available from the command line of the ESXi hosts.

This `tcpdump-uw` command shows if the primary is correctly sending multicast packets (port and IP info) and if all other hosts in the cluster are receiving them.

On the primary, this command shows the packets being sent out to the multicast address. On all other hosts, the same packets are visible (from the primary to the multicast address). If they are not visible, multicast is not working correctly. Run the `tcpdump-uw` command shown here on any host in the cluster, and the heartbeats from the primary are visible. In this case, the primary is at IP address 172.32.0.2. The **-v** for verbosity is optional.

```
[root@esxi-hp-02:~] tcpdump-uw -i vmk2 multicast -v
tcpdump-uw: listening on vmk2, link-type EN10MB (Ethernet), capture size 96 bytes
11:04:21.800575 IP truncated-ip - 146 bytes missing! (tos 0x0, ttl 5, id 34917, offset 0,
flags [none], proto UDP (17), length 228)
    172.32.0.4.44824 > 224.1.2.3.12345: UDP, length 200
11:04:22.252369 IP truncated-ip - 234 bytes missing! (tos 0x0, ttl 5, id 15011, offset 0,
flags [none], proto UDP (17), length 316)
    172.32.0.2.38170 > 224.2.3.4.23451: UDP, length 288
11:04:22.262099 IP truncated-ip - 146 bytes missing! (tos 0x0, ttl 5, id 3359, offset 0,
flags [none], proto UDP (17), length 228)
    172.32.0.3.41220 > 224.2.3.4.23451: UDP, length 200
11:04:22.324496 IP truncated-ip - 146 bytes missing! (tos 0x0, ttl 5, id 20914, offset 0,
flags [none], proto UDP (17), length 228)
    172.32.0.5.60460 > 224.1.2.3.12345: UDP, length 200
11:04:22.800782 IP truncated-ip - 146 bytes missing! (tos 0x0, ttl 5, id 35010, offset 0,
flags [none], proto UDP (17), length 228)
    172.32.0.4.44824 > 224.1.2.3.12345: UDP, length 200
11:04:23.252390 IP truncated-ip - 234 bytes missing! (tos 0x0, ttl 5, id 15083, offset 0,
flags [none], proto UDP (17), length 316)
    172.32.0.2.38170 > 224.2.3.4.23451: UDP, length 288
11:04:23.262141 IP truncated-ip - 146 bytes missing! (tos 0x0, ttl 5, id 3442, offset 0,
flags [none], proto UDP (17), length 228)
    172.32.0.3.41220 > 224.2.3.4.23451: UDP, length 200
```

While this output might seem a little confusing, suffice to say that the output shown here indicates that the four hosts in the cluster are getting a heartbeat from the primary. This **`tcpdump-uw`** command must be run on every host to verify that they are all receiving the heartbeat. This verifies that the primary is sending the heartbeats, and every other host in the cluster is receiving them, which indicates that multicast is working.

If some of the vSAN hosts are not able to pick up the one-second heartbeats from the primary, the network administrator needs to check the multicast configuration of their switches.

To avoid the annoying **`truncated-ip - 146 bytes missing!`** message, use the **-s0** option to the same command to stop trunacating of packets:

```
[root@esxi-hp-02:~] tcpdump-uw -i vmk2 multicast -v -s0
tcpdump-uw: listening on vmk2, link-type EN10MB (Ethernet), capture size 65535 bytes
11:18:29.823622 IP (tos 0x0, ttl 5, id 56621, offset 0, flags [none], proto UDP (17), length
228)
    172.32.0.4.44824 > 224.1.2.3.12345: UDP, length 200
11:18:30.251078 IP (tos 0x0, ttl 5, id 52095, offset 0, flags [none], proto UDP (17), length
228)
```

```
    172.32.0.3.41220 > 224.2.3.4.23451: UDP, length 200
11:18:30.267177 IP (tos 0x0, ttl 5, id 8228, offset 0, flags [none], proto UDP (17), length
316)
    172.32.0.2.38170 > 224.2.3.4.23451: UDP, length 288
11:18:30.336480 IP (tos 0x0, ttl 5, id 28606, offset 0, flags [none], proto UDP (17), length
228)
    172.32.0.5.60460 > 224.1.2.3.12345: UDP, length 200
11:18:30.823669 IP (tos 0x0, ttl 5, id 56679, offset 0, flags [none], proto UDP (17), length
228)
    172.32.0.4.44824 > 224.1.2.3.12345: UDP, length 200
```

The **tcpdump** command is related to IGMP (Internet Group Management Protocol) membership. Hosts (and network devices) use IGMP to establish multicast group membership.

Each ESXi host in the vSAN cluster sends out regular IGMP membership reports (Join).

The **tcpdump** command shows IGMP member reports from a host:

```
[root@esxi-dell-m:~] tcpdump-uw -i vmk1 igmp
tcpdump-uw: verbose output suppressed, use -v or -vv for full protocol decode
listening on vmk1, link-type EN10MB (Ethernet), capture size 262144 bytes
15:49:23.134458 IP 172.40.0.9 > igmp.mcast.net: igmp v3 report, 1 group record(s)
15:50:22.994461 IP 172.40.0.9 > igmp.mcast.net: igmp v3 report, 1 group record(s)
```

The output shows IGMP v3 reports are taking place, indicating that the ESXi host is regularly updating its membership. If a network administrator has any doubts whether or not vSAN ESXi hosts are doing IGMP correctly, running this command on each ESXi host in the cluster and showing this trace can be used to verify.

If you have multicast communications, use IGMP v3.

In fact, the following command can be used to look at multicast and IGMP traffic at the same time:

```
[root@esxi-hp-02:~] tcpdump-uw -i vmk2 multicast or igmp -v -s0
```

A common issue is that the vSAN cluster is configured across multiple physical switches, and while multicast has been enabled on one switch, it has not been enabled across switches. In this case, the cluster forms with two ESXi hosts in one partition, and another ESXi host (connected to the other switch) is unable to join this cluster. Instead it forms its own vSAN cluster in another partition. The `vsan.lldpnetmap` command seen earlier can help you determine network configuration, and which hosts are attached to which switch.

While a vSAN cluster forms, there are indicators that show multicast might be an issue.

Assume that the checklist for subnet, VLAN, MTU has been followed, and each host in the cluster can `vmkping` every other host in the cluster.

If there is a multicast issue when the cluster is created, a common symptom is that each ESXi host forms its own vSAN cluster, with itself as the primary. If each host has a unique network partition ID, this symptom suggests that there is no multicast between any of the hosts.

However, if there is a situation where a subset of the ESXi hosts form a cluster, and another subset form another cluster, and each have unique partitions with their own primary, backup and perhaps even agent hosts, multicast is enabled in the switch, but not across switches. vSAN shows hosts on the first physical switch forming their own cluster partition, and hosts on the second physical switch forming their own cluster partition, each with its own primary. If you can verify which switches the hosts in the cluster connect to, and hosts in a cluster are connected to the same switch, then this probably is the issue.

Checking vSAN Network Performance

Make that there is sufficient bandwidth between your ESXi hosts. This tool can assist you in testing whether your vSAN network is performing optimally.

To check the performance of the vSAN network, you can use `iperf` tool to measure maximum TCP bandwidth and latency. It is located in `/usr/lib/vmware/vsan/bin/iperf.copy`. Run it with `--help` to see the various options. Use this tool to check network bandwidth and latency between ESXi hosts participating in a vSAN cluster.

VMware KB 2001003 can assist with setup and testing.

This is most useful when a vSAN cluster is being commissioned. Running **iperf** tests on the vSAN network when the cluster is already in production can impact the performance of the virtual machines running on the cluster.

Checking vSAN Network Limits

The **`vsan.check.limits`** command verifies that none of the vSAN thresholds are being breached.

```
> ls
0 /
1 vcsa-04.rainpole.com/
> cd 1
/vcsa-04.rainpole.com> ls
0 Datacenter (datacenter)
/vcsa-04.rainpole.com> cd 0
/vcsa-04.rainpole.com/Datacenter> ls
0 storage/
1 computers [host]/
2 networks [network]/
3 datastores [datastore]/
4 vms [vm]/
/vcsa-04.rainpole.com/Datacenter> cd 1
/vcsa-04.rainpole.com/Datacenter/computers> ls
0 Cluster (cluster): cpu 155 GHz, memory 400 GB
1 esxi-dell-e.rainpole.com (standalone): cpu 38 GHz, memory 123 GB
2 esxi-dell-f.rainpole.com (standalone): cpu 38 GHz, memory 123 GB
3 esxi-dell-g.rainpole.com (standalone): cpu 38 GHz, memory 123 GB
4 esxi-dell-h.rainpole.com (standalone): cpu 38 GHz, memory 123 GB
/vcsa-04.rainpole.com/Datacenter/computers> vsan.check_limits 0
```

```
2017-03-14 16:09:32 +0000: Querying limit stats from all hosts ...
2017-03-14 16:09:34 +0000: Fetching vSAN disk info from esxi-dell-m.rainpole.com (may take a
moment) ...
2017-03-14 16:09:34 +0000: Fetching vSAN disk info from esxi-dell-n.rainpole.com (may take a
moment) ...
2017-03-14 16:09:34 +0000: Fetching vSAN disk info from esxi-dell-o.rainpole.com (may take a
moment) ...
2017-03-14 16:09:34 +0000: Fetching vSAN disk info from esxi-dell-p.rainpole.com (may take a
moment) ...
2017-03-14 16:09:39 +0000: Done fetching vSAN disk infos
+--------------------------+--------------------
+-----------------------------------------------------------------+
| Host                     | RDT
| Disks                                                           |
+--------------------------+--------------------
+-----------------------------------------------------------------+
| esxi-dell-m.rainpole.com |
Assocs: 1309/45000 | Components: 485/9000                                          |
|                          | Sockets:
89/10000  | naa.500a075113019b33: 0% Components: 0/0                        |
|                          | Clients:
136       | naa.500a075113019b37: 40% Components: 81/47661                  |
|                          | Owners:
138        | t10.ATA_____Micron_P420m2DMTFDGAR1T4MAX_____ 0% Components: 0/0 |
|                          |                      |
naa.500a075113019b41: 37% Components: 80/47661                  |
|                          |                      |
naa.500a07511301a1eb: 38% Components: 81/47661                  |
|                          |                      |
naa.500a075113019b39: 39% Components: 79/47661                  |
|                          |                      |
naa.500a07511301a1ec: 41% Components: 79/47661                  |
<<truncated>>
```

From a network perspective, it is the RDT associations (Assocs) and sockets count that are important. There are 45,000 associations per host in vSAN 6.0 and later. An RDT association is used to track peer-to-peer network state within vSAN. vSAN is sized so that it never runs out of RDT associations. vSAN also limits how many TCP sockets it is allowed to use, and vSAN is sized so that it never runs out of its allocation of TCP sockets. There is a limit of 10,000 sockets per host.

A vSAN **client** represents object's access in the vSAN cluster. The client typically represents a virtual machine running on a host. The client and the object might not be on the same host. There is no hard defined limit, but this metric is shown to help understand how clients balance across hosts.

There is only one vSAN **owner** for a given vSAN object, typically co-located with the vSAN client accessing this object. vSAN owners coordinate all access to the vSAN object and implement functionality, such as mirroring and striping. There is no hard defined limit, but this metric is once again shown to help understand how owners balance across hosts.

Using Multicast in vSAN Network

14

Multicast is a network communication technique that sends information packets to a group of destinations over an IP network.

Releases earlier than vSAN version 6.6 support IP multicast and used IP multicast communication as a discovery protocol to identify the nodes trying to join a vSAN cluster. Releases earlier than vSAN version 6.6 depend on IP multicast communication while joining and leaving the cluster groups and during other intra-cluster communication operations. Ensure that you enable and configure the IP multicast in the IP network segments to carry the vSAN traffic service.

An IP multicast address is called a Multicast Group (MG). IP multicast sends source packets to multiple receivers as a group transmission. IP multicast relies on communication protocols that hosts, clients, and network devices use to participate in multicast-based communications. Communication protocols such as Internet Group Management Protocol (IGMP) and Protocol Independent Multicast (PIM) are the main components and dependencies for the use of IP multicast communications.

While creating a vSAN cluster, a default multicast address is assigned to each vSAN cluster. The vSAN traffic service automatically assigns the default multicast address settings to each host. This multicast address sends frames to a default multicast group and multicast group agent.

When multiple vSAN clusters reside on the same Layer 2 network, VMware recommends changing the default multicast address within the additional vSAN clusters. This prevents multiple clusters from receiving all multicast streams. See VMware KB 2075451 for more information about changing the default vSAN multicast address.

This chapter includes the following topics:

- Internet Group Management Protocol
- Protocol Independent Multicast

Internet Group Management Protocol

You can use Internet Group Management Protocol (IGMP) to add receivers to the IP Multicast group membership within the Layer 2 domains.

IGMP allows receivers to send requests to the multicast groups they want to join. Becoming a member of a multicast group allows routers to forward traffic for the multicast groups on the Layer 3 segment where the receiver is connected to switch port.

You can use IGMP snooping to limit the physical switch ports participating in the multicast group to only vSAN VMkernel port uplinks. IGMP snooping is configured with an IGMP snooping querier. The need to configure an IGMP snooping querier to support IGMP snooping varies by switch vendor. Consult your specific switch vendor for IGMP snooping configuration.

vSAN supports both IGMP version 2 and IGMP version 3. When you perform the vSAN deployment across Layer 3 network segments, you can configure a Layer 3 capable device such as a router or a switch with a connection and access to the same Layer 3 network segments.

All VMkernel ports on the vSAN network subscribe to a multicast group using IGMP to avoid multicast flooding all network ports.

Note You can deactivate IGMP snooping if vSAN is on a non-routed or trunked VLAN that you can extend to the vSAN ports of all the hosts in the cluster.

Protocol Independent Multicast

Protocol Independent Multicast (PIM) consists of Layer 3 multicast routing protocols.

It provides different communication techniques for IP multicast traffic to reach receivers that are in different Layer 3 segments from the multicast groups sources. For earlier vSAN version 6.6 cluster, you must use PIM to enable the multicast traffic to flow across different subnets. Consult your network vendor for the implementation of PIM.

Networking Considerations for vSAN File Service

15

vSAN File Service is a layer that sits on top of vSAN to provide file shares. It currently supports SMB, NFSv3, and NFSv4.1 file shares.

Following are the network considerations for vSAN File Service:

- You must allocate static IP addresses as file server IPs from vSAN File Service network, each IP is the access point to vSAN file shares.
  - For best performance, the number of IP addresses must be equal to the number of hosts in the vSAN cluster.
  - All the static IP addresses should be from the same subnet.
  - Every static IP address has a corresponding FQDN, which should be part of the Forward lookup and Reverse lookup zones in the DNS server.
- You must ensure to prepare the network as vSAN File Service network:
  - If using standard switch based network, the Promiscuous Mode and Forged Transmits are enabled as part of the vSAN File Services enablement process.
  - If using DVS based network, vSAN File Services are supported on DVS version 6.6.0 or later. Create a dedicated port group for vSAN File Services in the DVS. MacLearning and Forged Transmits are enabled as part of the vSAN File Services enablement process for a provided DVS port group.

    Note If using NSX-based network, ensure that MacLearning is enabled for the provided network entity from the NSX admin console, and all the hosts and File Services nodes are connected to the desired NSX-T network.
- For SMB share and NFS share with Kerberos security, you must provide information about your AD domain and organizational unit (optional). In addition, a user account with sufficient privileges to create and delete objects is required.
- Ensure that the file server can access AD server and DNS server. The file server must be able to access all the ports required by AD service.

Following are the ports that vSAN File Service uses for network connectivity. Ensure that these ports are not blocked by the firewall.

Service	Port Number	Entity	Connectivity Requirements
Server Message Block (SMB)	TCP port 445	File Servers	External network to file servers
Quotas for a user of a local filesystem (RQUOTA)	TCP port 875	File Servers	External network to file servers
Network File System (NFS)	TCP and UDP port 2049	File Servers	External network to file servers. NFSv3 can use both TCP and UDP ports but NFSv4.1 uses only TCP.
NFS Mount	TCP and UDP port 20048	File Servers	External network to file servers
Network Status Monitor (NSM) server daemon	TCP and UDP port 27689	File Servers	External network to file servers. Both inward and outward communication must be permitted.
Network Lock Manager (NLM)	TCP and UDP port 32803	File Servers	External network to file servers. Allows the connection initiated from File Server to client. Inbound and outbound connections must be allowed on firewall. The default port is UDP.
LDAP	TCP port 389	Active Directory (AD) servers (if AD domain is configured)	File servers to AD servers
LDAP to Global Catalog	TCP port 3268	AD servers (if AD domain is configured)	File servers to AD servers
Kerberos	TCP port 88	AD servers (if AD domain is configured)	File servers to AD servers
Kerberos password change	TCP port 464	AD servers (if AD domain is configured)	File servers to AD servers
Domain Name Server (DNS)	TCP and UDP port 53	DNS servers	File servers to DNS servers
vSAN Distributed File System (VDFS) Server	TCP port 1564	ESXi hosts	Inside vSAN network

Networking Considerations for iSCSI on vSAN

16

vSAN iSCSI target service allows hosts and physical workloads that reside outside the vSAN cluster to access the vSAN datastore. This feature enables an iSCSI initiator on a remote host to transport block-level data to an iSCSI target on a storage device within the vSAN cluster.

The iSCSI targets on vSAN are managed using Storage Policy Based Management (SPBM) similar to other vSAN objects. For the iSCSI LUNs, this space savings the space through deduplication and compression, and provides security through encryption. For enhanced security, vSAN iSCSI target service uses Challenge Handshake Authentication Protocol (CHAP) and Mutual CHAP authentication.

vSAN identifies each iSCSI target by a unique iSCSI qualified Name (IQN). The iSCSI target is presented to a remote iSCSI initiator using the IQN, so that the initiator can access the LUN of the target. vSAN iSCSI target service allows creating iSCSI initiator groups. The iSCSI initiator group restricts access to only those initiators that are members of the group.

This chapter includes the following topics:

Characteristics of vSAN iSCSI Network

Following are the characteristics of a vSAN iSCSI network:

- iSCSI Routing - iSCSI initiators can make routed connections to vSAN iSCSI targets over an L3 network.
- IPv4 and IPv6 - vSAN iSCSI network supports both IPv4 and IPv6.
- IP Security - IPSec on the vSAN iSCSI network provides increased security.

  Note ESXi hosts support IPsec using IPv6 only.

- Jumbo Frames - Jumbo Frames are supported on the vSAN iSCSI network.
- NIC Teaming - All NIC teaming configurations are supported on the vSAN iSCSI network.
- Multiple Connections per Session (MCS) - vSAN iSCSI implementation does not support MCS.

Migrating from Standard to Distributed vSwitch

17

You can migrate from a vSphere Standard Switch to a vSphere Distributed Switch, and use Network I/O Control. This enables you to prioritize the QoS (Quality of Service) on vSAN traffic.

Warning It is best to have access to the ESXi hosts, although you might not need it. If something goes wrong, you can access the console of the ESXi hosts.

Make a note of the original vSwitch setup. In particular, note the load-balancing and NIC teaming settings on the source. Make sure the destination configuration matches the source.

Create a Distributed Switch

Create the distributed vSwitch and give it a name.

1 In the vSphere Client Host and Clusters view, right-click a data center and select menu **New Distributed Switch**.

2 Enter a name.

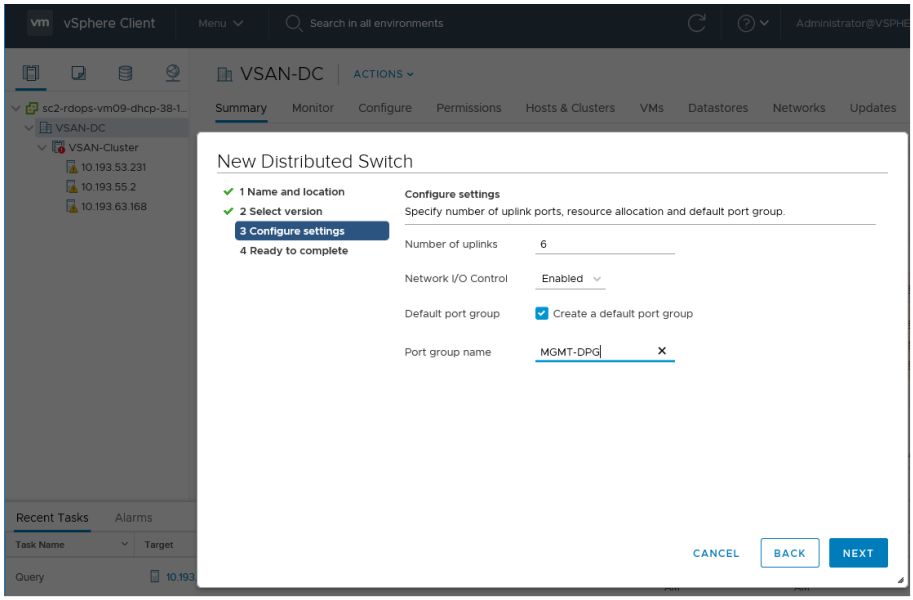
3 Select the version of the vSphere Distributed Switch. In this example, version 6.6.0 is used for the migration.

4 Add the settings. Determine how many uplinks you are currently using for networking. This example has six: management, vMotion, virtual machines, and three for vSAN (a LAG configuration). Enter 6 for the number of uplinks. Your environment might be different, but you can edit it later.

   You can create a default port group at this point, but additional port groups are needed.

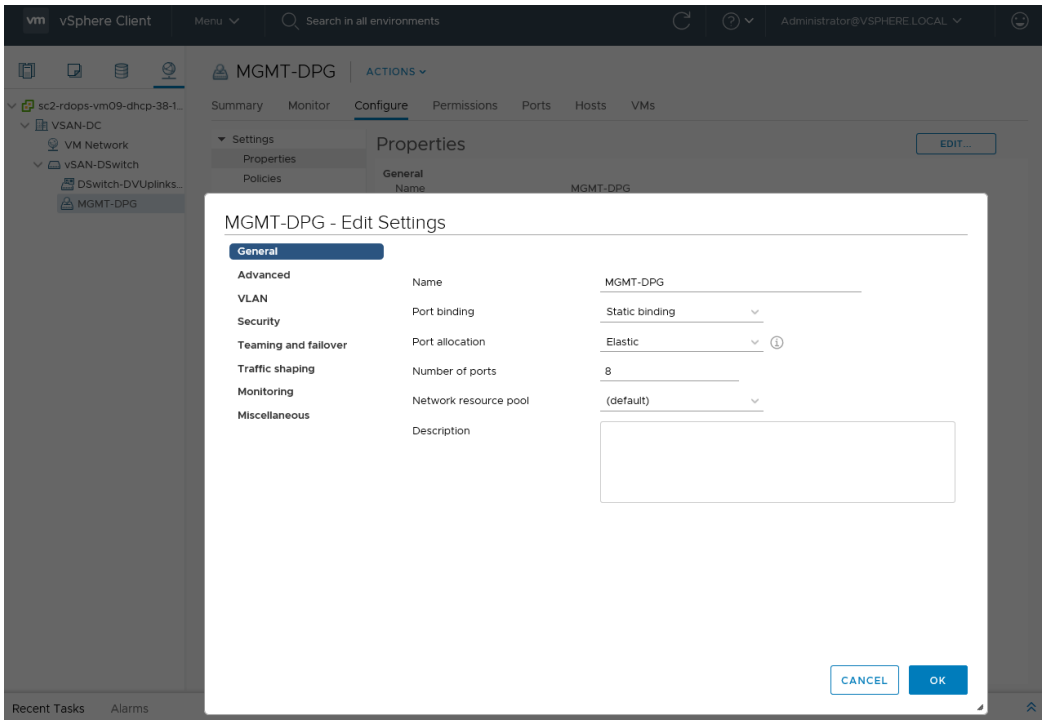
5 Finish the configuration of the distributed vSwitch.

The next step is to configure and create the additional port groups.



Create Port Groups

A single default port group was created for the management network. Edit this port group to make sure it has all the characteristics of the management port group on the standard vSwitch, such as VLAN and NIC teaming, and failover settings.



Configure the management port group.

1 In the vSphere Client Networking view, select the distributed port group, and click **Edit**.

2 For some port groups, you must change the VLAN. Since VLAN 51 is the management VLAN, tag the distributed port group accordingly.

3 Click **OK**.

Create distributed port groups for vMotion, virtual machine networking, and vSAN networking.

1 Right-click the vSphere Distributed Switch and select menu **Distributed Port Group > New Distributed Port Group**.

2 For this example, create a port group for the vMotion network.

Create all the distributed port groups on the distributed vSwitch. Then migrate the uplinks, VMkernel networking, and virtual machine networking to the distributed vSwitch and associated distributed port groups.

Warning Migrate the uplinks and networks in step-by-step fashion to proceed smoothly and with caution.

Migrate Management Network

Migrate the management network (vmk0) and its associated uplink (vmnic0) from the standard vSwitch to the distributed vSwitch (vDS).

1 Add hosts to the vDS.

   a Right-click the vDS and select menu **Add and Manage Hosts**.

   b Add hosts to the vDS. Click the green Add icon (+), and add all hosts from the cluster.

2 Configure the physical adapters and VMkernel adapters.

   a Click **Manage physical adapters** to migrate the physical adapters and VMkernel adapters, vmnic0 and vmk0 to the vDS.

   b Select an appropriate uplink on the vDS for physical adapter vmnic0. For this example, use Uplink1. The physical adapter is selected and an uplink is chosen.

3 Migrate the management network on vmk0 from the standard vSwitch to the distributed vSwitch. Perform these steps on each host.

   a Select vmk0, and click **Assign port group**.

   b Assign the distributed port group created for the management network earlier.

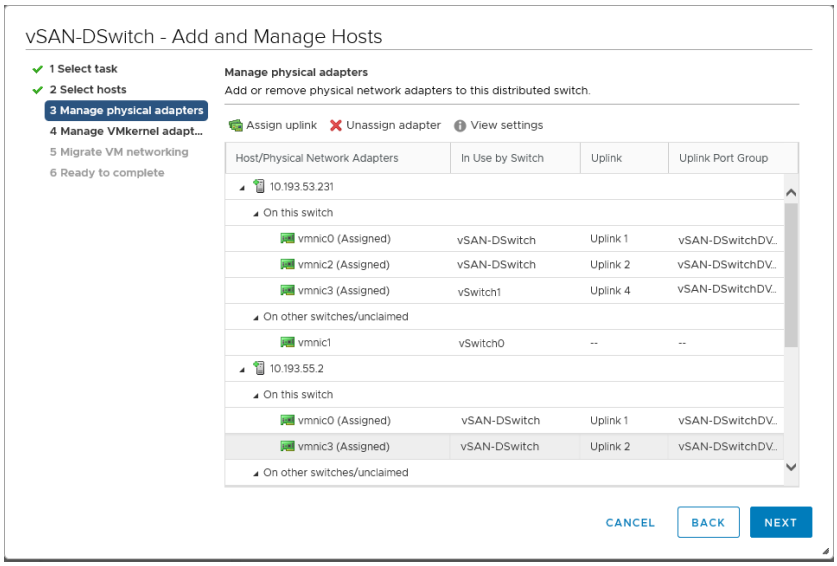
4 Finish the configuration.

   a Review the changes to ensure that you are adding four hosts, four uplinks (vmnic0 from each host), and four VMkernel adapters (vmk0 from each host).

   b Click **Finish**.

When you examine the networking configuration of each host, review the switch settings, with one uplink (vmnic0) and the vmk0 management port on each host.

Repeat this process for the other networks.



Migrate vMotion

To migrate the vMotion network, use the same steps used for the management network.

Before you begin, ensure that the distributed port group for the vMotion network has the same attributes as the port group on the standard vSwitch. Then migrate the uplink used for vMotion (vmnic1), with the VMkernel adapter (vmk1).

Migrate vSAN Network

If you have a single uplink for the vSAN network, then use the same process as before. However, if you are using more than one uplink, there are additional steps.

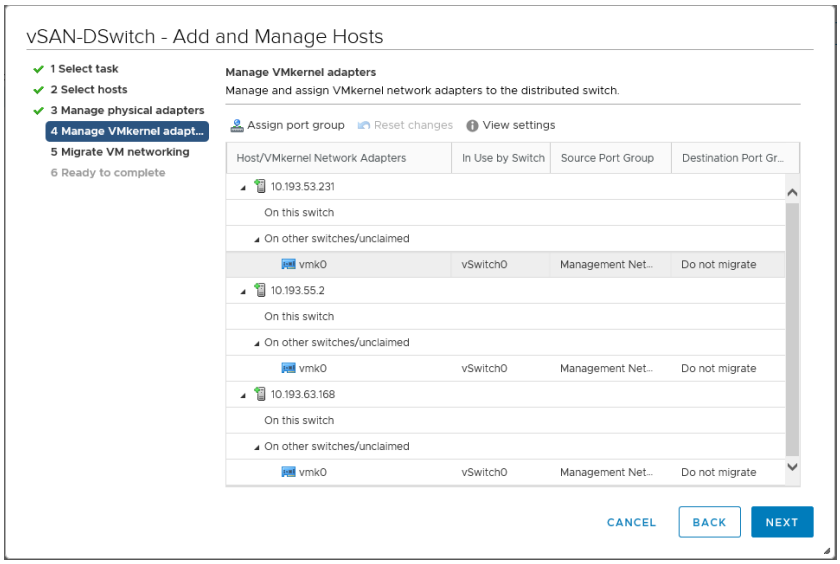
If the vSAN network is using Link Aggregation (LACP), or it is on a different VLAN to the other VMkernel networks, place some of the uplinks into an unused state for certain VMkernel adapters.

For example, VMkernel adapter vmk2 is used for vSAN. However, uplinks vmnic3, 4 and 5 are used for vSAN and they are in a LACP configuration. Therefore, for vmk2, all other vmnics (0, 1 and 2) must be placed in an unused state. Similarly, for the management adapter (vmk0) and vMotion adapter (vmk0), place the vSAN uplinks/vmnics in an unused state.

Modify the settings of the distributed port group and change the path policy and failover settings. On the **Manage physical network adapter** page, perform the steps for multiple adapters.

Assign the vSAN VMkernel adapter (vmk2) to the distributed port group for vSAN.

Note If you are only now migrating the uplinks for the vSAN network, you might not be able to change the distributed port group settings until after the migration. During this time, vSAN might have communication issues. After the migration, move to the distributed port group settings and make any policy changes and mark any uplinks to be unused. vSAN networking then returns to normal when this task is finished. Use the vSAN health service to verify that everything is functional.



Migrate VM Network

The final task needed to migrate the network from a standard vSwitch to a distributed vSwitch is to migrate the VM network.

Manage host networking.

1 Right-click the vDS and choose menu **Add and Manage Hosts**.

2 Select all the hosts in the cluster, to migrate virtual machine networking for all hosts to the distributed vSwitch.

   Do not move any uplinks. However, if the VM networking on your hosts used a different uplink, then migrate the uplink from the standard vSwitch.

3 Select the VMs to migrate from a virtual machine network on the standard vSwitch to the virtual machine distributed port group on the distributed vSwitch. Click **Assign port group**, and select the distributed port group.

4 Review the changes and click **Finish**. In this example, you are moving to VMs. Any templates using the original standard vSwitch virtual machine network must be converted to virtual machines, and edited. The new distributed port group for virtual machines must be selected as the network. This step cannot be achieved through the migration wizard.

Since the standard vSwitch no longer has any uplinks or port groups, it can be safely removed.

This completes the migration from a vSphere Standard Switch to a vSphere Distributed Switch.

Checklist Summary for vSAN Network

18

Use the checklist summary to verify your vSAN network requirements.

- Check if you use shared 10Gb NIC or dedicated 1Gb NIC. All-flash clusters require 10Gb NICs.
- Verify that redundant NIC teaming connections are configured.
- Verify that flow control is enabled on the ESXi host NICs.
- Verify that VMkernel port for vSAN network traffic is configured on each host.
- Verify that you have identical VLAN, MTU and subnet across all interfaces.
- Verify that you can run **`vmkping`** successfully between all hosts. Use the health service to verify.
- If you use jumbo frames, verify that you can run **`vmkping`** successfully with 9000 packet size between all hosts. Use the health service to verify.
- If your vSAN version is earlier than v6.6, verify that multicast is enabled on the network.
- If your vSAN version is earlier than v6.6 and multiple vSAN clusters are on the same network, configure multicast to use unique multicast addresses.
- If your vSAN version is earlier than v6.6 and spans multiple switches, verify that multicast is configured across switches.
- If your vSAN version is earlier than v6.6 and is routed, verify that PIM is configured to allow multicast routing.
- Ensure that the physical switch can meet vSAN requirements (multicast, flow control, feature interoperability).
- Verify that the network does not have performance issues, such as excessive dropped packets or pause frames.
- Verify that network limits are within acceptable margins.
- Test vSAN network performance with **`iperf`**, and verify that it meets expectations.