Architecture and Design

VMware Validated Design 4.0
VMware Validated Design for Micro-Segmentation 4.0

vmware®

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

https://docs.vmware.com/

If you have comments about this documentation, submit your feedback to

docfeedback@vmware.com

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

Contents

About Architecture and Design for VMware Validated Design for Micro-Segmentation

The *Architecture and Design* document for the VMware Validated Design for Micro-Segmentation use case contains a validated model of the use case and provides a detailed design of each component.

The document discusses the building blocks and the main principles of each layer and provides the available design options according to the design objective. A set of design decisions clearly lays out the decisions that were made and includes justification and potential implementation of each decision.

See the *Planning and Preparation* document for the VMware Validated Design for Micro-Segmentation for more information about supported product versions.

Note Design decisions in this document are based on design decisions in the *Architecture and Design* document for the VMware Validated Design for the Software-Defined Data Center, but some decision have been removed or changed. As a result, the decisions are not always numbered consecutively.

Intended Audience

The *Architecture and Design* document is intended for cloud architects, infrastructure administrators and cloud administrators who are familiar with and want to use VMware software to deploy in a short time and manage an SDDC that meets the requirements for capacity, scalability, backup and restore, and extensibility for disaster recovery support.

VMware Validated Design for the SDDC and this Use Case Documentation

Some of the information in this guide, in particular illustrations, show a dual-region design or include VMware vSAN. The Validated Design for Micro-Segmentation is a single-region design that does not include VMware vSAN. This design can be expanded to use vSAN or two regions.

Updated Information

This *VMware Validated Design for Micro-Segmentation Architecture and Design* is updated with each release of the product or when necessary.

This table provides the update history of the *VMware Validated Design for Micro-Segmentation Architecture and Design*.

Revision	Description
002236-01	■ Changed iBGP to BGP in Physical Network Design Decisions and Network Transport. ■ Removed design decision SDDC-VI-VC-024 in Compute Cluster Design. ■ Change the share value for NFS traffic from Normal to Low in design decision SDDC-VI-NET-010 in Network I/O Control
002236-00	Initial release.

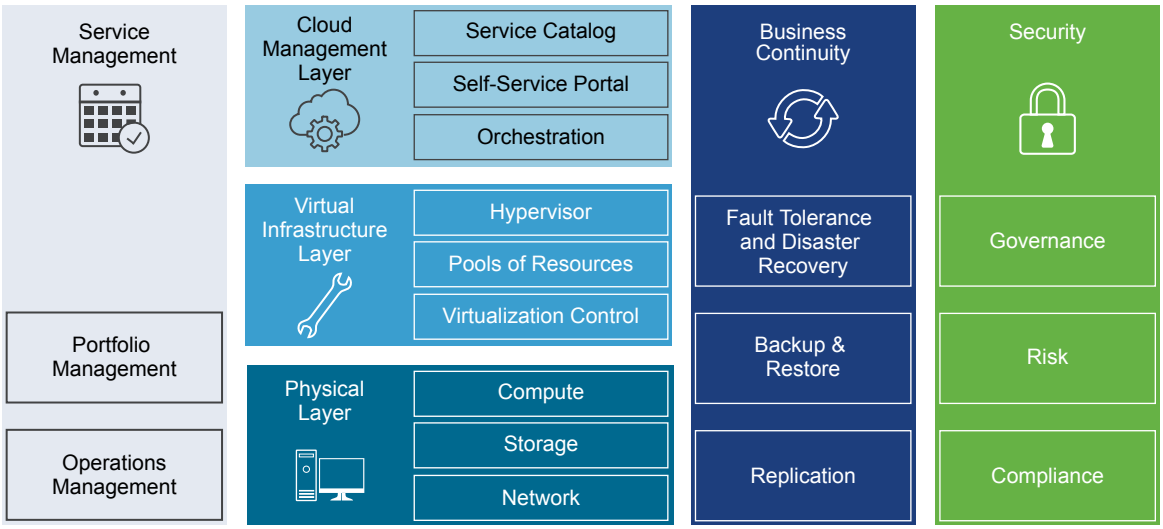
Architecture Overview

1

The VMware Validated Design for Software-Defined Data Center (SDDC) enables an IT organization to automate the provisioning of common repeatable requests and to respond to business needs with more agility and predictability. Traditionally this has been referred to as IaaS, or Infrastructure as a Service, however the VMware Validated Design for Software-Defined Data Center extends the typical IaaS solution to include a broader and more complete IT solution.

The VMware Validated Design architecture is based on a number of layers and modules, which allows interchangeable components be part of the end solution or outcome such as the SDDC. If a particular component design does not fit a business or technical requirement for whatever reason, it should be able to be swapped out for another similar component. The VMware Validated Designs are one way of putting an architecture together. They are rigorously tested to ensure stability, scalability and compatibility. Ultimately, the system is designed in such a way as to ensure the desired IT outcome will be achieved.

Figure 1-1. Architecture Overview



Physical Layer

The lowest layer of the solution is the Physical Layer, sometimes referred to as the 'core', which consists of three main components, Compute, Network and Storage. Inside the compute component sit the x86 based servers that run the management, edge and tenant compute workloads. There is some guidance around the physical capabilities required to run this architecture, however no recommendations on the type or brand of hardware is given. All components must be supported on the *VMware Hardware Compatibility* guide.

Virtual Infrastructure Layer

Sitting on the Physical Layer components is the Virtual Infrastructure Layer. Within the Virtual Infrastructure Layer, access to the physical underlying infrastructure is controlled and allocated to the management and tenant workloads. The Virtual Infrastructure Layer consists primarily of the physical host's hypervisor and the control of these hypervisors. The management workloads consist of elements in the virtual management layer itself, along with elements in the Cloud Management Layer, Service Management, Business Continuity and Security areas.

Cloud Management Layer

The Cloud Management Layer is the "top" layer of the stack and is where the service consumption occurs. Typically through a UI or API, this layer calls for resources and then orchestrates the actions of the lower layers to achieve the request. While the SDDC can stand on its own without any other ancillary services, for a complete SDDC experience other supporting components are needed. The Service Management, Business Continuity and Security areas complete the architecture by providing this support.

Service Management

When building any type of IT infrastructure, portfolio and operations management play key roles in continued day-to-day service delivery. The Service Management area of this architecture mainly focuses on operations management in particular monitoring, alerting and log management.

Business Continuity

To ensure a system is enterprise ready, it must contain elements to support business continuity in the area of data backup, restoration and disaster recovery. This area ensures that when data loss occurs, the right elements are in place to prevent permanent loss to the business. The design provides comprehensive guidance on how to operate backup and restore functions, along with run books detailing how to fail over components in the event of a disaster.

Security

All systems need to be inherently secure by design. This is to reduce risk and increase compliance while still providing a governance structure. The security area outlines what is needed to ensure the entire SDDC is resilient to both internal and external threats.

Physical Infrastructure Architecture 2

The architecture of the data center physical layer is based on logical hardware pods, a leaf-and-spine network topology, and zones and regions for high availability.

This chapter includes the following topics:

Pod Architecture

The VMware Validated Design for SDDC uses a small set of common building blocks called pods.

Pod Architecture Characteristics

Pods can include different combinations of servers, storage equipment, and network equipment, and can be set up with varying levels of hardware redundancy and varying quality of components. Pods are connected to a network core that distributes data between them. The pod is not defined by any hard physical properties, as it is a standard unit of connected elements within the SDDC network fabric.

A pod is a logical boundary of functionality for the SDDC platform. While each pod usually spans one rack, it is possible to aggregate multiple pods into a single rack in smaller setups. For both small and large setups, homogeneity and easy replication are important.

Different pods of the same type can provide different characteristics for varying requirements. For example, one compute pod could use full hardware redundancy for each component (power supply through memory chips) for increased availability. At the same time, another compute pod in the same setup could use low-cost hardware without any hardware redundancy. With these variations, the architecture can cater to the different workload requirements in the SDDC.

One of the guiding principles for such deployments is that VLANs are not spanned beyond a single pod by the network virtualization layer. Although this VLAN restriction appears to be a simple requirement, it has widespread impact on how a physical switching infrastructure can be built and on how it scales.

Pod to Rack Mapping

Pods are not mapped one-to-one to 19" data center racks. While a pod is an atomic unit of a repeatable building block, a rack is merely a unit of size. Because pods can have different sizes, how pods are mapped to 19" data center racks depends on the use case.

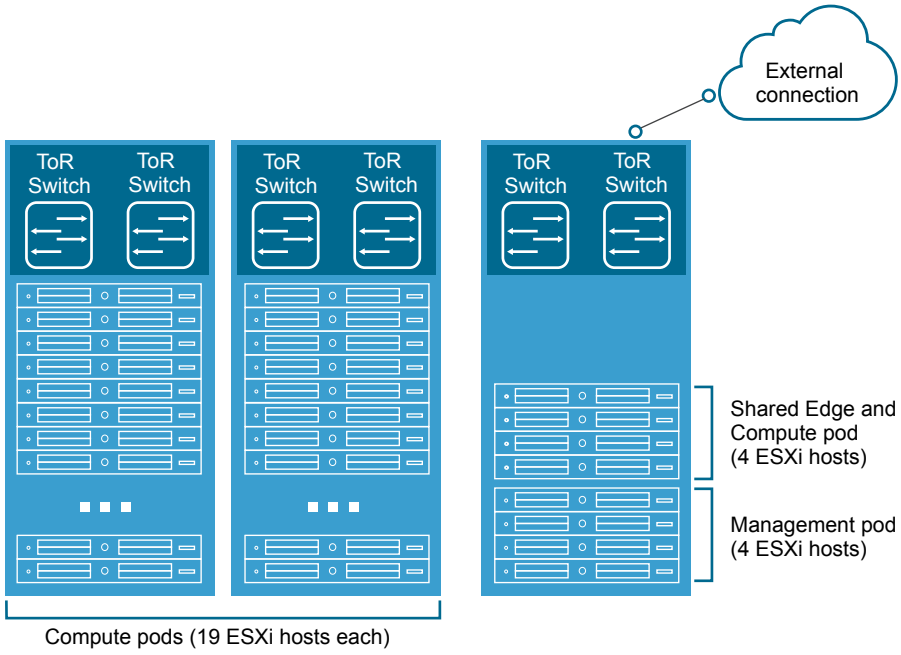
- One Pod in One Rack. One pod can occupy exactly one rack.
- Multiple Pods in One Rack. Two or more pods can occupy a single rack, for example, one management pod and one shared edge and compute pod can be deployed to a single rack.
- Single Pod Across Multiple Racks. A single pod can stretch across multiple adjacent racks. For example, a storage pod with filer heads and disk shelves can span more than one rack or a compute pod that has more host then a single rack can support.

Note The mangement and the shared edge and compute pods can not span racks. This is due to NSX controllers and other virtual machines on a VLAN backed network migrating to a different rack where that IP subnet is not available due to layer 2 termination at the Top of Rack switch.

Pod Types

The SDDC differentiates between different types of pods including management pod, compute pod, shared edge and compute pod, and storage pod. Each design includes several pods.

Figure 2-1. Pods in the SDDC



Management Pod

The management pod runs the virtual machines that manage the SDDC. These virtual machines host vCenter Server, NSX Manager, NSX Controller, and vRealize Log Insight. Different types of management pods can support different SLAs. Because the management pod hosts critical infrastructure, consider implementing a basic level of hardware redundancy for this pod.

Management pod components must not have tenant-specific addressing.

Shared Edge and Compute Pod

The shared edge and compute pod runs the required NSX services to enable north-south routing between the SDDC and the external network, and east-west routing inside the SDDC. This shared pod also hosts the SDDC tenant virtual machines (sometimes referred to as workloads or payloads). As the SDDC grows, additional compute-only pods can be added to support a mix of different types of workloads for different types of Service Level Agreements (SLAs).

Compute Pod

Compute pods host the SDDC tenant virtual machines (sometimes referred to as workloads or payloads). An SDDC can mix different types of compute pods and provide separate compute pools for different types of SLAs.

Storage Pod

Storage pods provide network-accessible storage using NFS or iSCSI. Different types of storage pods can provide different levels of SLA, ranging from just a bunch of disks (JBODs) using IDE drives with minimal to no redundancy, to fully redundant enterprise-class storage arrays. For bandwidth-intense IP-based storage, the bandwidth of these pods can scale dynamically.

Physical Network Architecture

The physical network architecture is tightly coupled with the pod-and-core architecture, and uses a Layer 3 leaf-and-spine network instead of the more traditional 3-tier data center design.

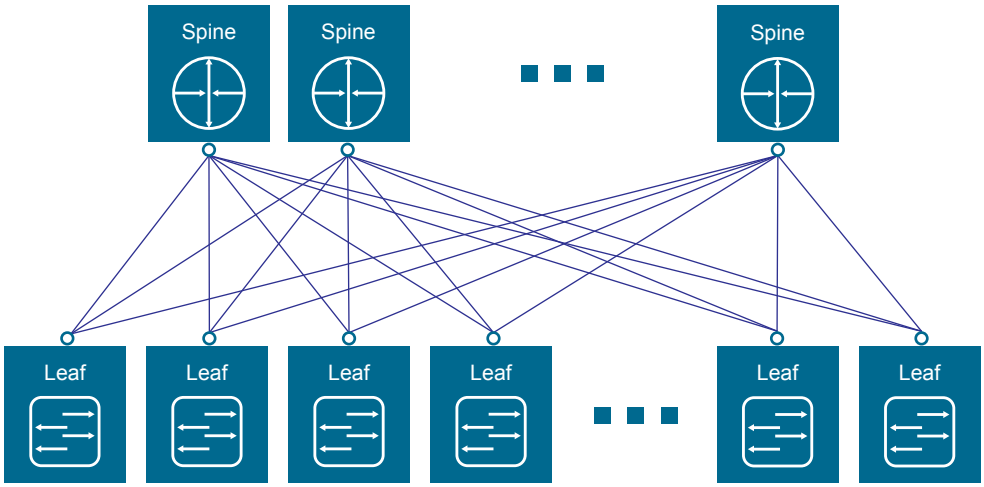
Leaf-and-Spine Network Architecture

A leaf-and-spine network is the core building block for the physical network in the SDDC.

- A leaf switch is typically located inside a rack and provides network access to the servers inside that rack, it is also referred to as a Top of Rack (ToR) switch.

- A spine switch is in the spine layer and provides connectivity between racks. Links between spine switches are typically not required. If a link failure occurs between a spine switch and a leaf switch, the routing protocol ensures that no traffic for the affected rack is sent to the spine switch that has lost connectivity to that rack.

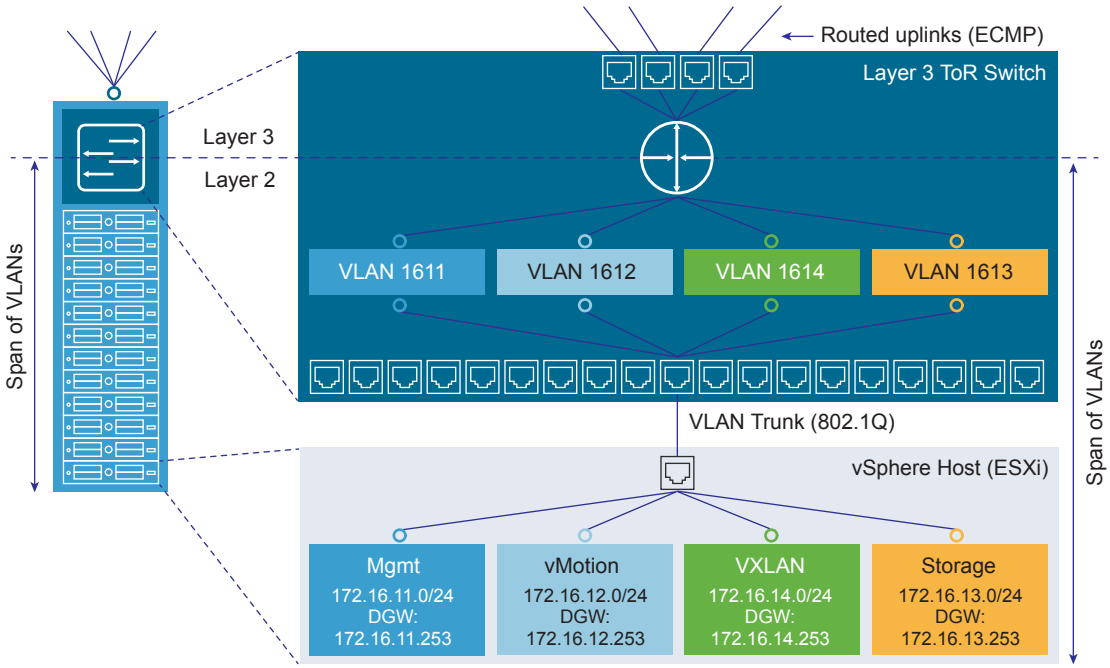
Figure 2-2. Leaf-and-Spine Physical Network Design



Ports that face the servers inside a rack should have a minimal configuration, shown in the following high-level physical and logical representation of the leaf node.

Note Each leaf node has identical VLAN configuration with unique /24 subnets assigned to each VLAN.

Figure 2-3. High-Level Physical and Logical Representation of a Leaf Node



Network Transport

You can implement the physical layer switch fabric for a SDDC by offering Layer 2 transport services or Layer 3 transport services to all components. For a scalable and vendor-neutral data center network, use Layer 3 transport.

Benefits and Drawbacks for Layer 2 Transport

In a design that uses Layer 2 transport, leaf switches and spine switches form a switched fabric, effectively acting like one large switch. Using modern data center switching fabric products such as Cisco FabricPath, you can build highly scalable Layer 2 multipath networks without the Spanning Tree Protocol (STP). Such networks are particularly suitable for large virtualization deployments, private clouds, and high-performance computing (HPC) environments.

Using Layer 2 routing has the following benefits and drawbacks:

- The benefit of this approach is more design freedom. You can span VLANs, which is useful for vSphere vMotion or vSphere Fault Tolerance (FT).
- The drawback is that the size of such a deployment is limited because the fabric elements have to share a limited number of VLANs. In addition, you have to rely on a specialized data center switching fabric product from a single vendor because these products are not designed for interoperability between vendors.

Benefits and Drawbacks for Layer 3 Transport

A design using Layer 3 transport requires these considerations:

- Layer 2 connectivity is limited within the data center rack up to the leaf switch.
- The leaf switch terminates each VLAN and provides default gateway functionality. That is, it has a switch virtual interface (SVI) for each VLAN.
- Uplinks from the leaf switch to the spine layer are routed point-to-point links. VLAN trunking on the uplinks is not allowed.
- A dynamic routing protocol, such as OSPF, ISIS, or BGP, connects the leaf switches and spine switches. Each leaf switch in the rack advertises a small set of prefixes, typically one per VLAN or subnet. In turn, the leaf switch calculates equal cost paths to the prefixes it received from other leaf switches.

Using Layer 3 routing has the following benefits and drawbacks:

- The benefit is that you can chose from a wide array of Layer 3 capable switch products for the physical switching fabric. You can mix switches from different vendors due to general interoperability between implementation of OSPF, ISIS or BGP. This approach is typically more cost effective because it makes use of only the basic functionality of the physical switches.
- A design restriction, and thereby a drawback of using Layer 3 routing, is that VLANs are restricted to a single rack. This affects vSphere vMotion, vSphere Fault Tolerance, and storage networks.

Infrastructure Network Architecture

A key goal of network virtualization is to provide a virtual-to-physical network abstraction.

To achieve this, the physical fabric must provide a robust IP transport with the following characteristics:

- Simplicity
- Scalability
- High bandwidth
- Fault-tolerant transport
- Support for different levels of quality of service (QoS)

Simplicity and Scalability

Simplicity and scalability are the first and most critical requirements for networking.

Simplicity

Configuration of the switches inside a data center must be simple. General or global configuration such as AAA, SNMP, syslog, NTP, and others should be replicated line by line, independent of the position of the switches. A central management capability to configure all switches at once is an alternative.

Scalability

Scalability factors include, but are not limited to, the following:

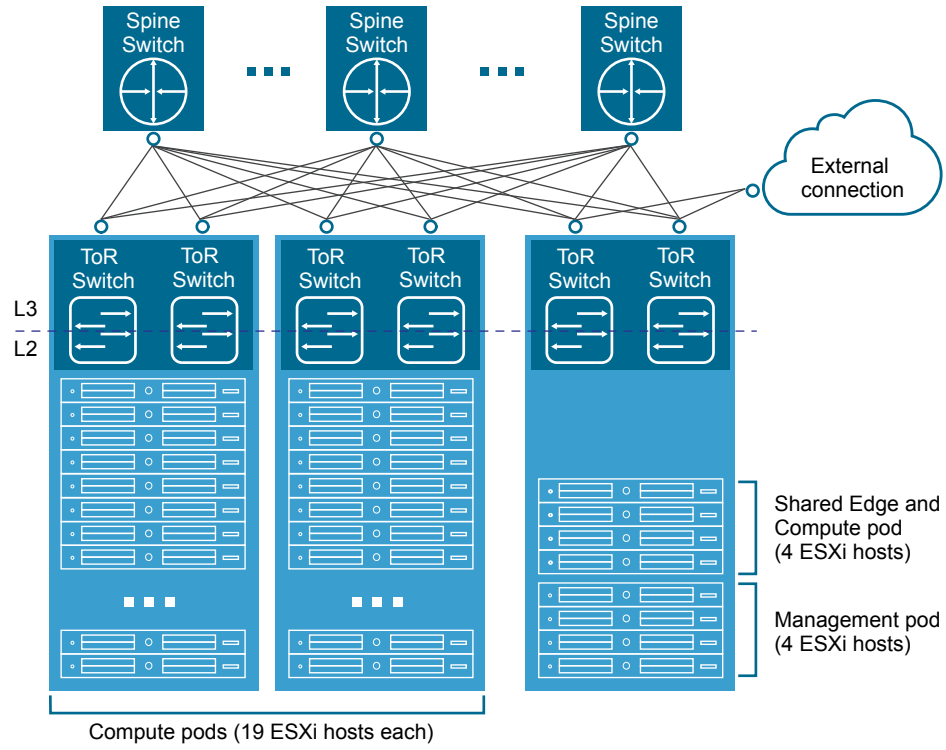
- Number of racks supported in a fabric.
- Amount of bandwidth between any two racks in a data center.
- Number of paths from which a leaf switch can select when communicating with another rack.

The total number of ports available across all spine switches and the oversubscription that is acceptable determine the number of racks supported in a fabric. Different racks may host different types of infrastructure, which can result in different bandwidth requirements.

- Racks with storage systems might attract or source more traffic than other racks.
- Compute racks, such as racks hosting hypervisors with workloads or virtual machines, might have different bandwidth requirements than shared edge and compute racks, which provide connectivity to the outside world.

Link speed and the number of links vary to satisfy different bandwidth demands. You can vary them for each rack without sacrificing other aspects of the leaf-and-spine architecture.

Figure 2-4. Pod Network Design



The number of links to the spine switches dictates how many paths for traffic from this rack to another rack are available. Because the number of hops between any two racks is consistent, the architecture can utilize equal-cost multipathing (ECMP). Assuming traffic sourced by the servers carries a TCP or UDP header, traffic distribution can occur on a per-flow basis.

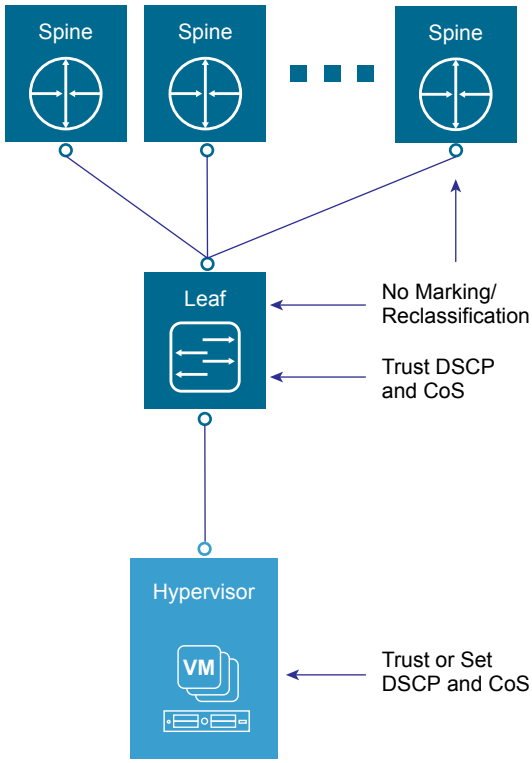
Quality of Service Differentiation

Virtualized environments carry different types of traffic, including tenant, storage and management traffic, across the switching infrastructure. Each traffic type has different characteristics and makes different demands on the physical switching infrastructure.

- Management traffic, although typically low in volume, is critical for controlling physical and virtual network state.
- IP storage traffic is typically high in volume and generally stays within a data center.

For virtualized environments, the hypervisor sets the QoS values for the different traffic types. The physical switching infrastructure has to trust the values set by the hypervisor. No reclassification is necessary at the server-facing port of a leaf switch. If there is a congestion point in the physical switching infrastructure, the QoS values determine how the physical network sequences, prioritizes, or potentially drops traffic.

Figure 2-5. Quality of Service (Differentiated Services) Trust Point



Two types of QoS configuration are supported in the physical switching infrastructure.

- Layer 2 QoS, also called class of service.
- Layer 3 QoS, also called DSCP marking.

A vSphere Distributed Switch supports both class of service and DSCP marking. Users can mark the traffic based on the traffic type or packet classification. When the virtual machines are connected to the VXLAN-based logical switches or networks, the QoS values from the internal packet headers are copied to the VXLAN-encapsulated header. This enables the external physical network to prioritize the traffic based on the tags on the external header.

Server Interfaces (NICs)

If the server has more than one server interface (NIC) of the same speed, use two as uplinks with VLANs trunked to the interfaces.

The vSphere Distributed Switch supports many different NIC Teaming options. Load-based NIC teaming supports optimal use of available bandwidth and supports redundancy in case of a link failure. Use two 10 GbE connections for each server in combination with a pair of leaf switches. 802.1Q network trunks can support a small number of VLANs. For example, management, storage, VXLAN, and VMware vSphere vMotion traffic.

Availability Zones and Regions

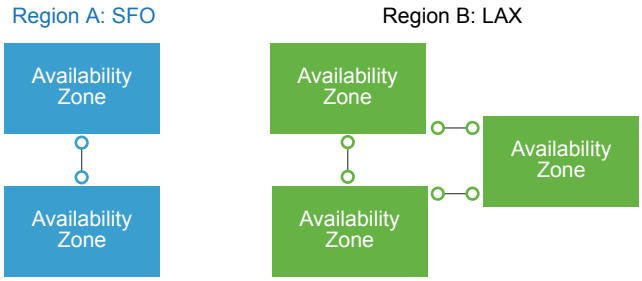
In an SDDC, availability zones are collections of infrastructure components. Regions support disaster recovery solutions and allow you to place workloads closer to your customers. Typically multiple availability zones form a single region.

This VMware Validated Design uses two regions, but uses only one availability zone in each region.

Note Some of the use cases, such as the VMware Validated Design for Micro-Segmentation, are validated for one regions. However, they can be expanded to use two regions, and from there to use two availability zones.

The following diagram shows how the design could be expanded to include multiple availability zones.

Figure 2-6. Availability Zones and Regions



Availability Zones

Each availability zone is isolated from other availability zones to stop the propagation of failure or outage across zone boundaries.

Together, multiple availability zones provide continuous availability through redundancy, helping to avoid outages and improve SLAs. An outage that is caused by external factors (such as power, cooling, and physical integrity) affects only one zone. Those factors most likely do not lead to an outage in other zones except in the case of major disasters.

Each availability zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. Each zone should have independent power supply, cooling system, network, and security. Common points of failures within a physical data center, like generators and cooling equipment, should not be shared across availability zones. Additionally, these zones should be physically separate so that even uncommon disasters affect only a single availability zone. Availability zones are usually either two distinct data centers within metro distance (latency in the single digit range) or two safety/fire sectors (data halls) within the same large scale data center.

Multiple availability zones (usually two) belong to a single region. The physical distance between availability zones can be up to approximately 50 kilometers (30 miles), which offers low, single-digit latency and large bandwidth by using dark fiber between the zones. This architecture allows the SDDC equipment across the availability zone to operate in an active/active manner as a single virtual data center or region.

You can operate workloads across multiple availability zones within the same region as if they were part of a single virtual data center. This supports an architecture with very high availability that is suitable for mission critical applications. When the distance between two locations of equipment becomes too large, these locations can no longer function as two availability zones within the same region and need to be treated as separate regions.

Regions

Multiple regions support placing workloads closer to your customers, for example, by operating one region on the US east coast and one region on the US west coast, or operating a region in Europe and another region in the US.

Regions are helpful in several ways.

- Regions can support disaster recovery solutions: One region can be the primary site and another region can be the recovery site.
- You can use multiple regions to address data privacy laws and restrictions in certain countries by keeping tenant data within a region in the same country.

The distance between regions can be rather large. This design uses two example regions, one region is San Francisco (SFO), the other region is Los Angeles (LAX).

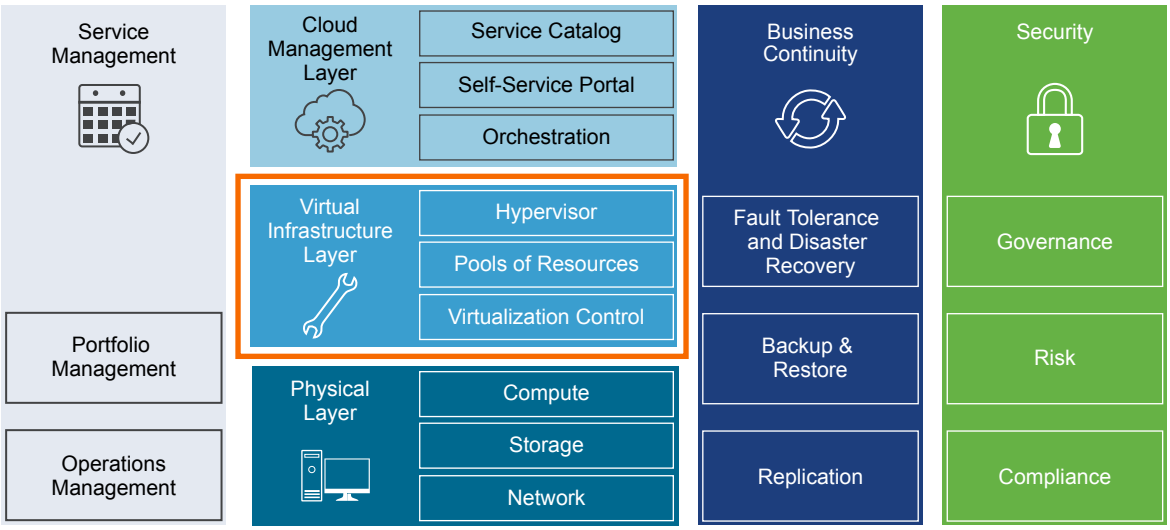
Virtual Infrastructure Architecture

3

The virtual infrastructure is the foundation of an operational SDDC.

Within the virtual infrastructure layer, access to the physical underlying infrastructure is controlled and allocated to the management and tenant workloads. The virtual infrastructure layer consists primarily of the physical hosts' hypervisors and the control of these hypervisors. The management workloads consist of elements in the virtual management layer itself, along with elements in the cloud management layer and in the service management, business continuity, and security areas.

Figure 3-1. Virtual Infrastructure Layer in the SDDC



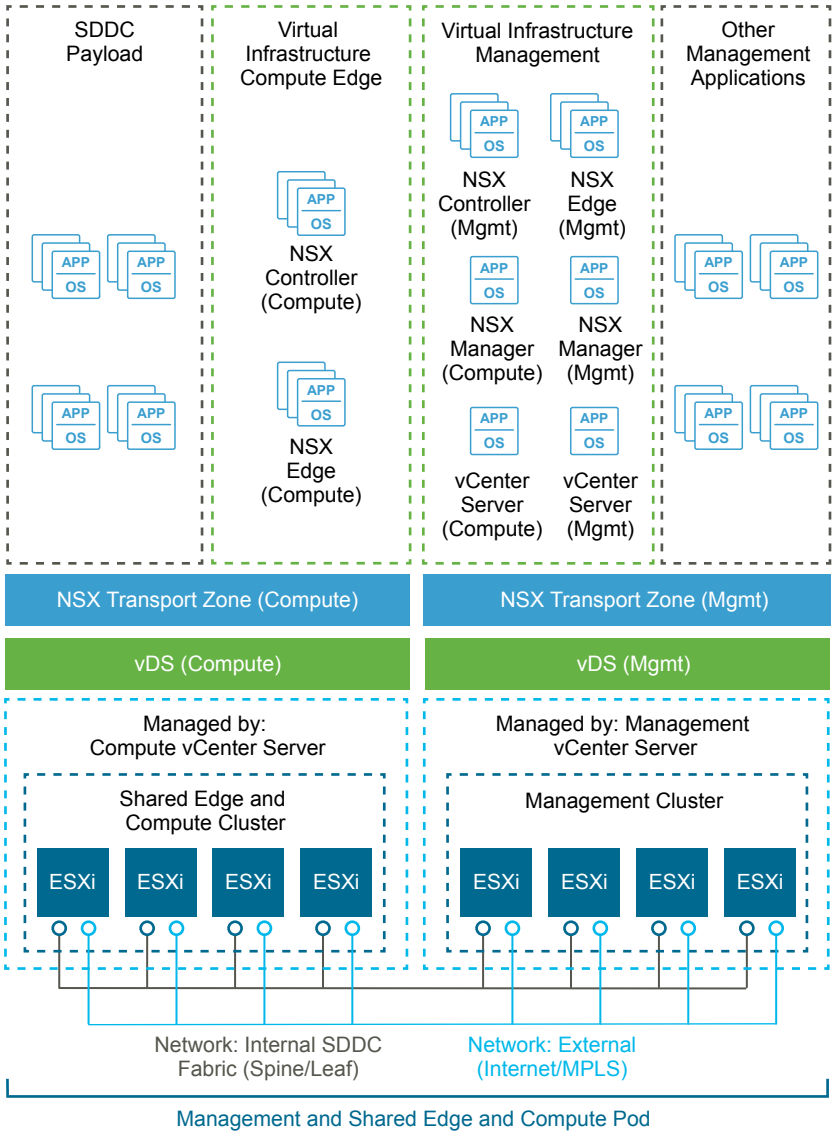
This chapter includes the following topics:

- Virtual Infrastructure Overview
- Network Virtualization Components
- Network Virtualization Services

Virtual Infrastructure Overview

The SDDC virtual infrastructure consists of two regions. Each region includes a management pod and a shared edge and compute pod.

Figure 3-2. SDDC Logical Design



Management Pod

Management pods run the virtual machines that manage the SDDC. These virtual machines host vCenter Server, vSphere Update Manager, NSX Manager, NSX Controller, vRealize Operations, vRealize Log Insight, vRealize Automation, Site Recovery Manager and other shared management components. All management, monitoring, and infrastructure services are provisioned to a vSphere cluster which provides high availability for these critical services. Permissions on the management cluster limit access to only administrators. This limitation protects the virtual machines that are running the management, monitoring, and infrastructure services.

Shared Edge and Compute Pod

The shared edge and compute pod runs the required NSX services to enable north-south routing between the SDDC and the external network and east-west routing inside the SDDC. This pod also hosts the SDDC tenant virtual machines (sometimes referred to as workloads or payloads). As the SDDC grows additional compute-only pods can be added to support a mix of different types of workloads for different types of SLAs.

Network Virtualization Components

VMware NSX for vSphere, the network virtualization platform, is a key solution in the SDDC architecture. The NSX for vSphere platform consists of several components that are relevant to the network virtualization design.

NSX for vSphere Platform

NSX for vSphere creates a network virtualization layer. All virtual networks are created on top of this layer, which is an abstraction between the physical and virtual networks. Several components are required to create this network virtualization layer:

- vCenter Server
- NSX Manager
- NSX Controller
- NSX Virtual Switch

These components are separated into different planes to create communications boundaries and provide isolation of workload data from system control messages.

Data plane	Workload data is contained wholly within the data plane. NSX logical switches segregate unrelated workload data. The data is carried over designated transport networks in the physical network. The NSX Virtual Switch, distributed routing, and the distributed firewall are also implemented in the data plane.
Control plane	Network virtualization control messages are located in the control plane. Control plane communication should be carried on secure physical networks (VLANs) that are isolated from the transport networks that are used for the data plane. Control messages are used to set up networking attributes on NSX Virtual Switch instances, as well as to configure and manage disaster recovery and distributed firewall components on each ESXi host.
Management plane	The network virtualization orchestration happens in the management plane. In this layer, cloud management platforms such as VMware vRealize® Automation™ can request, consume, and destroy networking resources for virtual workloads. Communication is directed from the cloud management platform to vCenter Server to create and manage virtual machines, and to NSX Manager to consume networking resources.

Network Virtualization Services

Network virtualization services include logical switches, logical routers, logical firewalls, and other components of NSX for vSphere.

Logical Switches

NSX for vSphere logical switches create logically abstracted segments to which tenant virtual machines can connect. A single logical switch is mapped to a unique VXLAN segment ID and is distributed across the ESXi hypervisors within a transport zone. This allows line-rate switching in the hypervisor without creating constraints of VLAN sprawl or spanning tree issues.

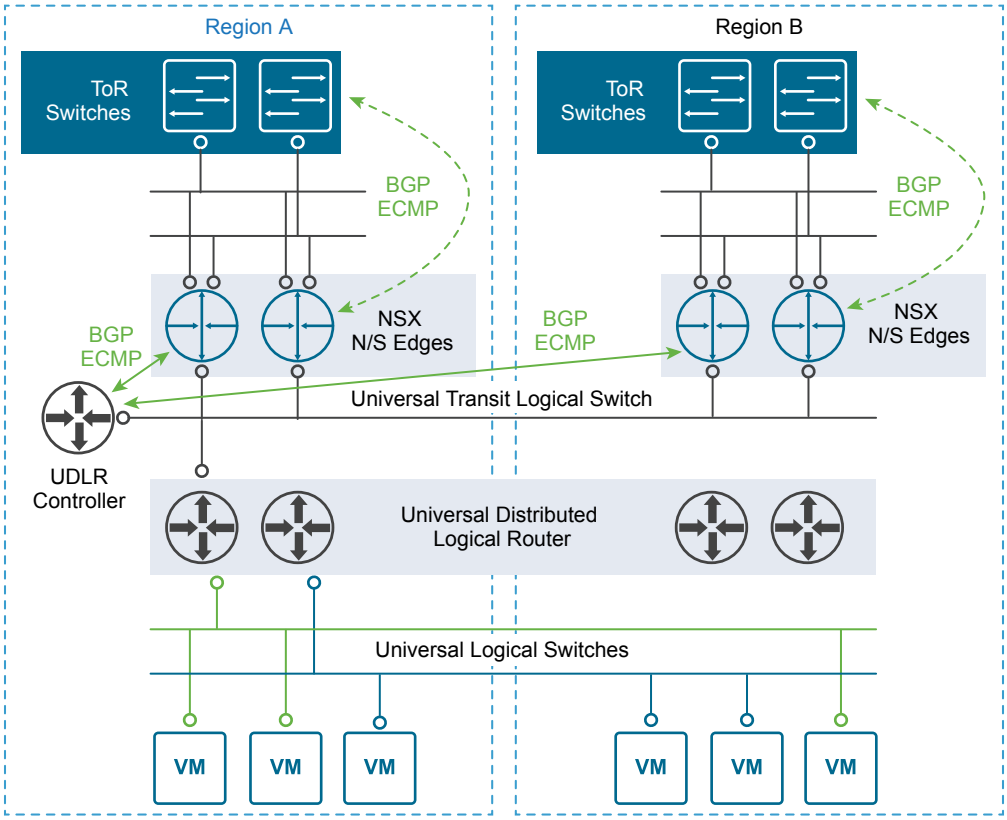
Universal Distributed Logical Router

The NSX for vSphere Universal Distributed Logical Router is optimized for forwarding in the virtualized space (between VMs, on VXLAN- or VLAN-backed port groups). Features include:

- High performance, low overhead first hop routing.
- Scaling the number of hosts.
- Support for up to 1,000 logical interfaces (LIFs) on each distributed logical router.

The Universal Distributed Logical Router is installed in the kernel of every ESXi host, as such it requires a VM to provide the control plane. The universal distributed logical router Control VM is the control plane component of the routing process, providing communication between NSX Manager and NSX Controller cluster through the User World Agent. NSX Manager sends logical interface information to the Control VM and NSX Controller cluster, and the Control VM sends routing updates to the NSX Controller cluster.

Figure 3-3. NSX for vSphere Universal Distributed Logical Router



Designated Instance

The designated instance is responsible for resolving ARP on a VLAN LIF. There is one designated instance per VLAN LIF. The selection of an ESXi host as a designated instance is performed automatically by the NSX Controller cluster and that information is pushed to all other hosts. Any ARP requests sent by the distributed logical router on the same subnet are handled by the same host. In case of host failure, the controller selects a new host as the designated instance and makes that information available to other hosts.

User World Agent

User World Agent (UWA) is a TCP and SSL client that enables communication between the ESXi hosts and NSX Controller nodes, and the retrieval of information from NSX Manager through interaction with the message bus agent.

Edge Services Gateway

While the Universal Logical Router provides VM to VM or east-west routing, the NSX Edge services gateway provides north-south connectivity, by peering with upstream Top of Rack switches, thereby enabling tenants to access public networks.

Logical Firewall

NSX for vSphere Logical Firewall provides security mechanisms for dynamic virtual data centers.

- The Distributed Firewall allows you to segment virtual data center entities like virtual machines. Segmentation can be based on VM names and attributes, user identity, vCenter objects like data centers, and hosts, or can be based on traditional networking attributes like IP addresses, port groups, and so on.
- The Edge Firewall component helps you meet key perimeter security requirements, such as building DMZs based on IP/VLAN constructs, tenant-to-tenant isolation in multi-tenant virtual data centers, Network Address Translation (NAT), partner (extranet) VPNs, and user-based SSL VPNs.

The Flow Monitoring feature displays network activity between virtual machines at the application protocol level. You can use this information to audit network traffic, define and refine firewall policies, and identify threats to your network.

Logical Virtual Private Networks (VPNs)

SSL VPN-Plus allows remote users to access private corporate applications. IPSec VPN offers site-to-site connectivity between an NSX Edge instance and remote sites. L2 VPN allows you to extend your datacenter by allowing virtual machines to retain network connectivity across geographical boundaries.

Logical Load Balancer

The NSX Edge load balancer enables network traffic to follow multiple paths to a specific destination. It distributes incoming service requests evenly among multiple servers in such a way that the load distribution is transparent to users. Load balancing thus helps in achieving optimal resource utilization, maximizing throughput, minimizing response time, and avoiding overload. NSX Edge provides load balancing up to Layer 7.

Service Composer

Service Composer helps you provision and assign network and security services to applications in a virtual infrastructure. You map these services to a security group, and the services are applied to the virtual machines in the security group.

Data Security provides visibility into sensitive data that are stored within your organization's virtualized and cloud environments. Based on the violations that are reported by the NSX for vSphere Data Security component, NSX security or enterprise administrators can ensure that sensitive data is adequately protected and assess compliance with regulations around the world.

NSX for vSphere Extensibility

VMware partners integrate their solutions with the NSX for vSphere platform to enable an integrated experience across the entire SDDC. Data center operators can provision complex, multi-tier virtual networks in seconds, independent of the underlying network topology or components.

Operations Architecture

4

The architecture of the operations management layer includes management components that provide support for the main types of operations in an SDDC. For the micro-segmentation use case, you can perform monitoring, logging with vRealize Log Insight.

Logging Architecture

vRealize Log Insight provides real-time log management and log analysis with machine learning-based intelligent grouping, high-performance searching, and troubleshooting across physical, virtual, and cloud environments.

Overview

vRealize Log Insight collects data from ESXi hosts using the syslog protocol. It can connect to other VMware products, like vCenter Server, to collect events, tasks, and alarms data, and can integrate with vRealize Operations Manager to send notification events and enable launch in context. vRealize Log Insight also functions as a collection and analysis point for any system capable of sending syslog data. In addition to syslog data an ingestion agent can be installed on Linux or Windows servers or may come pre-installed on certain VMware products to collect logs. This agent approach is especially useful for custom application logs and operating systems that don't natively support the syslog protocol, such as Windows.

Installation Models

You can deploy vRealize Log Insight as a virtual appliance in one of the following configurations:

- Standalone node
- Highly available cluster of one master and at least two worker nodes using an integrated load balancer (ILB)

The compute and storage resources of the vRealize Log Insight instances can scale-up as growth demands.

Cluster Nodes

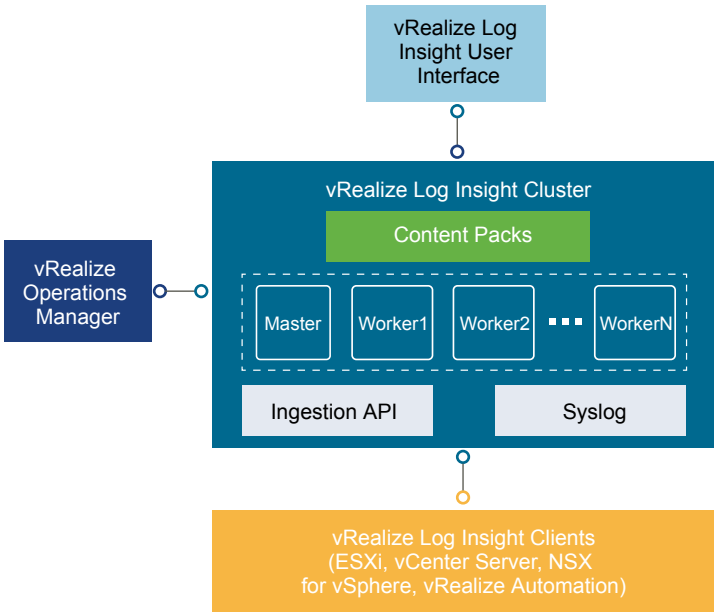
For high availability and scalability, you can deploy several vRealize Log Insight instances in a cluster. Each instance can have one of the following roles.

Master Node	Required initial node in the cluster. The master node is responsible for queries and log ingestion. The Web user interface of the master node serves as the single pane of glass for the cluster. All queries against data are directed to the master, which in turn queries the workers as appropriate.
Worker Node	Enables scale-out in larger environments. A worker node is responsible for ingestion of logs. A worker node stores logs locally. If a worker node is down, the logs on that worker becomes unavailable. You need at least two worker nodes to form a cluster with the master node.
Integrated Load Balancer (ILB)	Provides high availability (HA). The ILB runs on one of the cluster nodes. If the node that hosts the ILB Virtual IP (VIP) address stops responding, the VIP address is failed over to another node in the cluster.

Architecture of a Cluster

The architecture of vRealize Log Insight enables several channels for HA collection of log messages.

Figure 4-1. Cluster Architecture of vRealize Log Insight



vRealize Log Insight clients connect to ILB VIP address and use the Web user interface and ingestion (via Syslog or the Ingestion API) to send logs to vRealize Log Insight.

By default, the vRealize Log Insight Solution collects data from vCenter Server systems and ESXi hosts. For forwarding logs from NSX for vSphere, use content packs which contain extensions or provide integration with other systems in the SDDC.

Authentication Models

You can configure vRealize Log Insight for integration with Active Directory for user authentication in one or both of the following configurations:

- Embedded Active Directory Integration
- VMware Identity Manager

Archiving

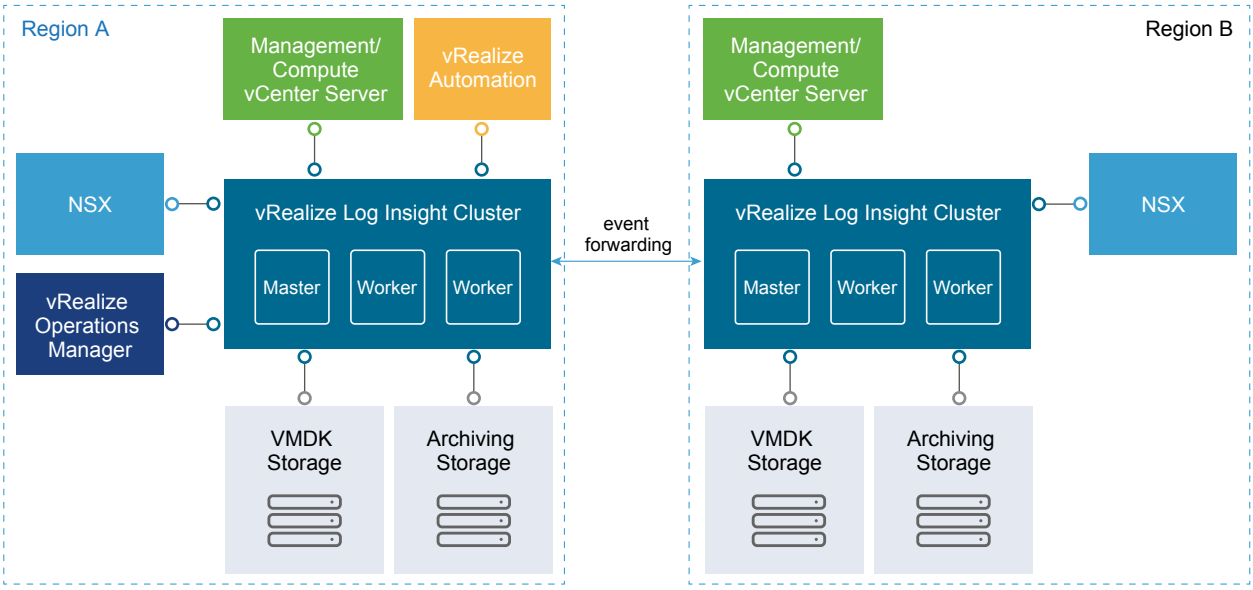
vRealize Log Insight supports data archiving on NFS shared storage that each vRealize Log Insight node can access.

Multi-Region vRealize Log Insight Deployment

The scope of the SDDC design covers multiple regions. Using vRealize Log Insight in a multi-region design can provide a syslog infrastructure in all regions of the SDDC. Using vRealize Log Insight across multiple regions requires deploying a cluster in each region.

vRealize Log Insight supports event forwarding to other vRealize Log Insight deployments across regions in the SDDC. Implementing failover by using vSphere Replication or disaster recovery by using Site Recovery Manager is not necessary. The event forwarding feature adds tags to log message that identify the source region and event filtering prevents looping messages between the regions.

Figure 4-2. Event Forwarding in vRealize Log Insight



Detailed Design

5

The Software-Defined Data Center (SDDC) detailed design considers both physical and virtual infrastructure design. It includes numbered design decisions and the justification and implications of each decision.

Each section also includes detailed discussion and diagrams.

Physical Infrastructure Design	Focuses on the three main pillars of any data center, compute, storage and network. In this section you find information about availability zones and regions. The section also provides details on the rack and pod configuration, and on physical hosts and the associated storage and network configurations.
Virtual Infrastructure Design	Provides details on the core virtualization software configuration. This section has information on the ESXi hypervisor, vCenter Server, the virtual network design including VMware NSX, and on software-defined storage for VMware vSAN. This section also includes details on business continuity (backup and restore) and on disaster recovery.
Operations Infrastructure Design	Explains how to architect, install, and configure vRealize Log Insight. You learn how to ensure that service management within the SDDC is comprehensive.

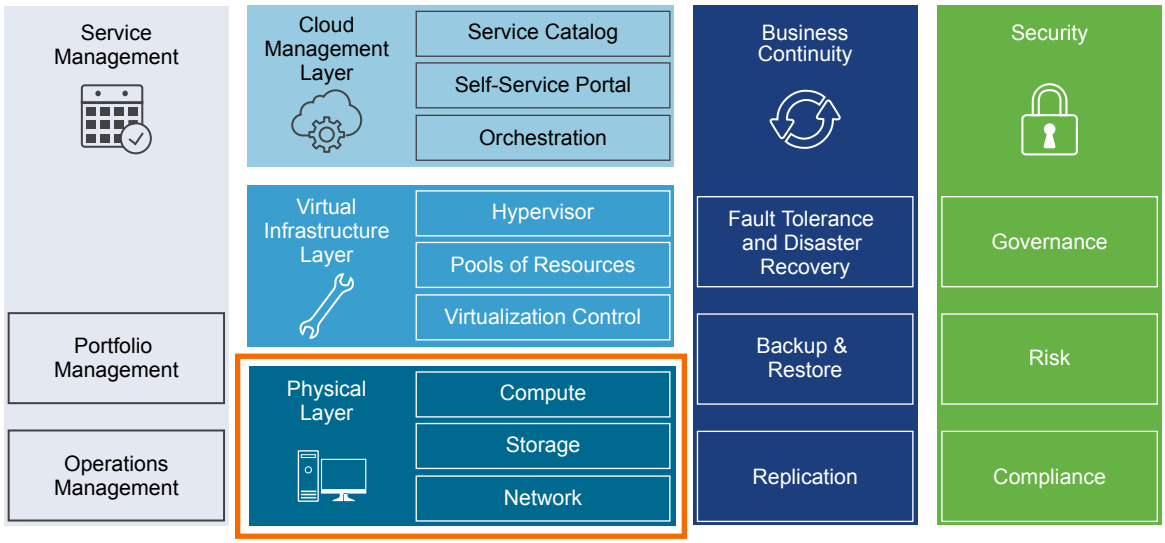
Physical Infrastructure Design

6

The physical infrastructure design includes details on decisions for availability zones and regions and the pod layout within datacenter racks.

Design decisions related to server, networking, and storage hardware are part of the physical infrastructure design.

Figure 6-1. Physical Infrastructure Design



- Physical Design Fundamentals

  Physical design fundamentals include decisions on availability zones and regions and on pod types, pods, and racks. The ESXi host physical design is also part of the design fundamentals.

- Physical Networking Design

  The physical network uses a leaf-and-spine network architecture.

- Physical Storage Design

  This VMware Validated Design relies on both VMware Virtual SAN storage and NFS storage. The Shared Storage Design section explains where the SDDC uses which type of storage and gives background information. The focus of this section is physical storage design.

Physical Design Fundamentals

Physical design fundamentals include decisions on availability zones and regions and on pod types, pods, and racks. The ESXi host physical design is also part of the design fundamentals.

Availability Zones and Regions

Availability zones and regions are used for different purposes.

Availability zones An availability zone is the fault domain of the SDDC. Multiple availability zone scan provide continuous availability of an SDDC, minimize unavailability of services and improve SLAs.

Regions Regions provide disaster recovery across different SDDC instances. This design uses two regions. Each region is a separate SDDC instance. The regions have a similar physical layer design and virtual infrastructure design but different naming. If you are expanding your design to include two regions, see the *Business Continuity / Disaster Recovery Design* chapter in the *VMware Validated Design for the Software-Defined Data Center Reference Architecture* document.

Note This design leverages a single availability zone for a one region deployment, and a single availability zone in each region in the case of a two region deployment.

The two-region design uses the following regions. The region identifier uses United Nations Code for Trade and Transport Locations(UN/LOCODE) along with a numeric instance ID.

Region	Region Identifier	Region-specific Domain Name	Region Description
A	SFO01	sfo01.rainpole.local	San Francisco, CA, USA based data center
B	LAX01	lax01.rainpole.local	Los Angeles, CA, USA based data center

Note Region Identifiers vary based on the locations used in your deployment.

Table 6-1. Availability Zones and Regions Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-001	Per region, a single availability zone that can support all SDDC management components is deployed.	A single availability zone can support all SDDC management and compute components for a region. You can later add another availability zone to extend and scale the management and compute capabilities of the SDDC.	Results in limited redundancy of the overall solution. The single availability zone can become a single point of failure and prevent high-availability design solutions.
SDDC-PHY-002	Use a single region for the initial deployment. Expand the design to two regions if appropriate.	Supports the technical requirement of multi-region failover capability as outlined in the design objectives.	Having multiple regions requires an increased solution footprint and associated costs.

Pods and Racks

The SDDC functionality is split across multiple pods. Each pod can occupy one rack or multiple racks. The total number of pods for each pod type depends on scalability needs.

Figure 6-2. SDDC Pod Architecture

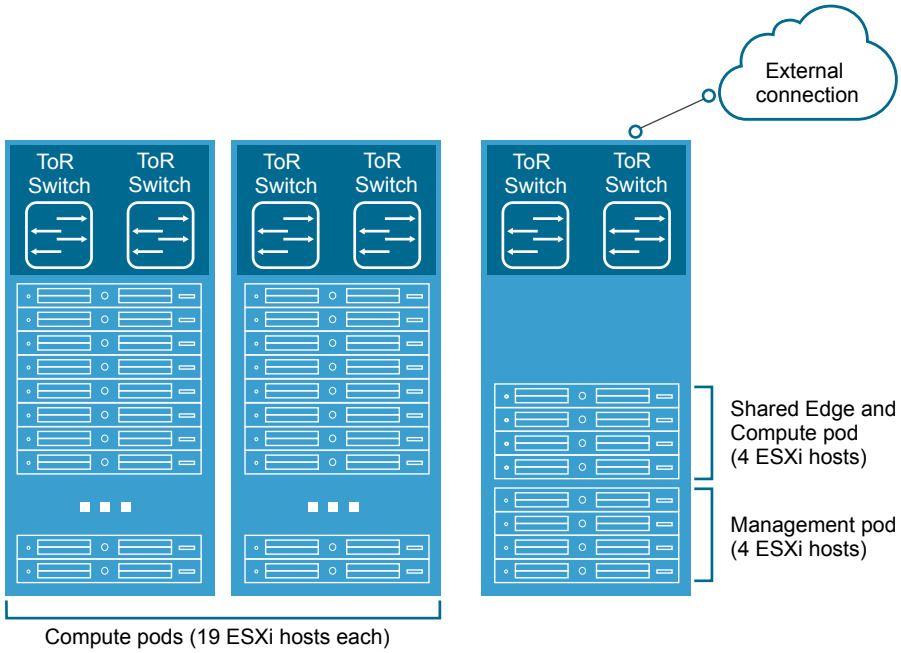


Table 6-2. Required Number of Racks

Pod (Function)	Required Number of Racks (for full scale deployment)	Minimum Number of Racks	Comment
Management pod and shared edge and compute pod	1	1	Two half-racks are sufficient for the management pod and shared edge and compute pod. As the number and resource usage of compute VMs increase adding additional hosts to the cluster will be required, as such extra space in the rack should be reserved for growth.
Compute pods	6	0	With 6 compute racks, 6 compute pods with 19 ESXi hosts each can achieve the target size of 6000 average-sized VMs. If an average size VM has two vCPUs with 4 GB of RAM, 6000 VMs with 20% overhead for bursting workloads require 114 hosts. The quantity and performance varies based on the workloads running within the compute pods.
Storage pods	6	0 (if using VSAN for Compute Pods)	Storage that is not Virtual SAN storage is hosted on isolated storage pods.
Total	13	1	

Table 6-3. POD and Racks Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-003	The management and the shared edge and compute pod occupy the same rack.	The number of required compute resources for the management pod (4 ESXi servers) and shared edge and compute pod (4 ESXi servers) are low and do not justify a dedicated rack for each pod. On-ramp and off-ramp connectivity to physical networks (for example, north-south L3 routing on NSX Edge virtual appliances) can be supplied to both the management and compute pods through this management/edge rack. Edge resources require external connectivity to physical network devices. Placing edge resources for management and compute in the same rack will minimize VLAN spread.	The design must include sufficient power and cooling to operate the server equipment. This depends on the selected vendor and products. If the equipment in this entire rack fails, a second region is needed to mitigate downtime associated with such an event.
SDDC-PHY-004	Storage pods can occupy one or more racks.	To simplify the scale out of the SDDC infrastructure, the storage pod to rack(s) relationship has been standardized. It is possible that the storage system arrives from the manufacturer in dedicated rack or set of racks and a storage system of this type is accommodated for in the design.	The design must include sufficient power and cooling to operate the storage equipment. This depends on the selected vendor and products.
SDDC-PHY-005	Each rack has two separate power feeds.	Redundant power feeds increase availability by ensuring that failure of a power feed does not bring down all equipment in a rack. Combined with redundant network connections into a rack and within a rack, redundant power feeds prevent failure of equipment in an entire rack.	All equipment used must support two separate power feeds. The equipment must keep running if one power feed fails. If the equipment of an entire rack fails, the cause, such as flooding or an earthquake, also affects neighboring racks. A second region is needed to mitigate downtime associated with such an event.
SDDC-PHY-006	Mount the compute resources (minimum of 4 ESXi hosts) for the management pod together in a rack.	Mounting the compute resources for the management pod together can ease physical datacenter design, deployment and troubleshooting.	None.
SDDC-PHY-007	Mount the compute resources for the shared edge and compute pod (minimum of 4 ESXi servers) together in a rack.	Mounting the compute resources for the shared edge and compute pod together can ease physical datacenter design, deployment and troubleshooting.	None.

ESXi Host Physical Design Specifications

The physical design specifications of the ESXi host list the characteristics of the hosts that were used during deployment and testing of this VMware Validated Design.

Physical Design Specification Fundamentals

The configuration and assembly process for each system is standardized, with all components installed the same manner on each host. Standardizing the entire physical configuration of the ESXi hosts is critical to providing an easily manageable and supportable infrastructure because standardization eliminates variability. Consistent PCI card slot location, especially for network controllers, is essential for accurate alignment of physical to virtual I/O resources. Deploy ESXi hosts with identical configuration, including identical storage, and networking configurations, across all cluster members. Identical configurations ensure an even balance of virtual machine storage components across storage and compute resources.

Select all ESXi host hardware, including CPUs following the *VMware Compatibility Guide*.

The sizing of the physical servers for the ESXi hosts for the management and edge pods has special consideration because it is based on the VMware document VMware Virtual SAN Ready Nodes, as these pod type use VMware vSAN.

- An average sized VM has two vCPUs with 4 GB of RAM.
- A standard 2U server can host 60 average-sized VMs on a single ESXi host.

Table 6-4. ESXi Host Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-008	Use vSAN Ready Nodes.	Using a vSAN Ready Node ensures seamless compatibility with vSAN during the deployment.	Might limit hardware choices.
SDDC-PHY-009	All nodes must have uniform configurations across a given cluster.	A balanced cluster delivers more predictable performance even during hardware failures. In addition, performance impact during resync/rebuild is minimal when the cluster is balanced.	Vendor sourcing, budgeting and procurement considerations for uniform server nodes will be applied on a per cluster basis.

ESXi Host Memory

The amount of memory required for compute pods will vary depending on the workloads running in the pod. When sizing memory for compute pod hosts it's important to remember the admission control setting (n+1) which reserves one hosts resources for fail over or maintenance.

Note See the *VMware vSAN 6.5 Design and Sizing Guide* for more information about disk groups, including design and sizing guidance. The number of disk groups and disks that an ESXi host manages determines memory requirements. 32 GB of RAM is required to support the maximum number of disk groups.

Table 6-5. Host Memory Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-010	Set up each ESXi host in the management pod to have a minimum 192 GB RAM.	The management and edge VMs in this pod require a total 424 GB RAM.	None

Host Boot Device Background Considerations

Minimum boot disk size for ESXi in SCSI-based devices (SAS / SATA / SAN) is greater than 5 GB. ESXi can be deployed using stateful local SAN SCSI boot devices, or by using vSphere Auto Deploy.

What is supported depends on the version of vSAN that your are using:

- vSAN does not support stateless vSphere Auto Deploy
- vSAN 5.5 and greater supports USB/SD embedded devices for ESXi boot device (4 GB or greater).
- Since vSAN 6.0, there is an option to use SATADOM as a supported boot device.

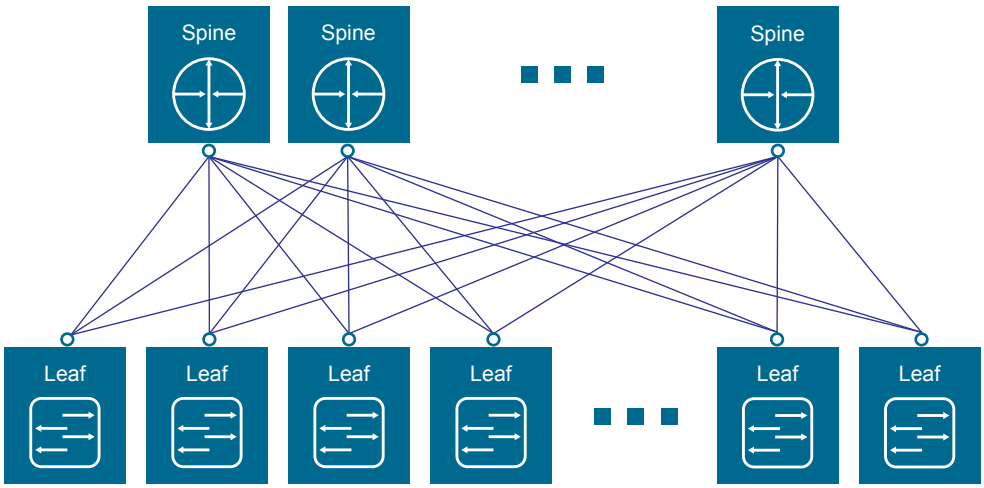
See the *VMware vSAN 6.5 Design and Sizing Guide* to choose the option that best fits your hardware.

Physical Networking Design

The physical network uses a leaf-and-spine network architecture.

Figure 6-3 illustrates a leaf-and-spine network architecture. For additional information, see *Physical Network Architecture*.

Figure 6-3. Leaf-and-Spine Architecture



Leaf-and-Spine and Network Virtualization Architecture

As virtualization, cloud computing, and distributed cloud become more pervasive in the data center, a shift in the traditional three-tier networking model is taking place. This shift addresses simplicity and scalability.

Simplicity

The traditional core-aggregate-access model is efficient for north/south traffic that travels in and out of the data center. This model is usually built for redundancy and resiliency against failure. However, the Spanning Tree Protocol (STP) typically blocks 50 percent of the critical network links to prevent network loops, which means 50 percent of the maximum bandwidth is wasted until something fails.

A core-aggregate-access architecture is still widely used for service-oriented traffic that travels north/south. However, the trends in traffic patterns are changing with the types of workloads. In today's data centers east/west or server-to-server traffic is common. If the servers in a cluster are performing a resource-intensive calculation in parallel, unpredictable latency or lack of bandwidth are undesirable. Powerful servers that perform these calculations can attempt to communicate with each other, but if they cannot communicate efficiently because of a bottleneck in the network architecture, wasted capital expenditure results.

One way to solve the problem is to create a leaf-and-spine architecture, also known as a distributed core.

A leaf-and-spine architecture has two main components: spine switches and leaf switches.

- Spine switches can be thought of as the core, but instead of being a large, chassis-based switching platform, the spine consists of many high-throughput Layer 3 switches with high port density.
- Leaf switches can be treated as the access layer. Leaf switches provide network connection points for servers and uplink to the spine switches.

Every leaf switch connects to every spine switch in the fabric. No matter which leaf switch a server is connected to, it always has to cross the same number of devices to get to another server (unless the other server is located on the same leaf). This design keeps the latency down to a predictable level because a payload has to hop only to a spine switch and another leaf switch to get to its destination.

Instead of relying on one or two large chassis-based switches at the core, the load is distributed across all spine switches, making each individual spine insignificant as the environment scales out.

Scalability

Several factors, including the following, affect scalability.

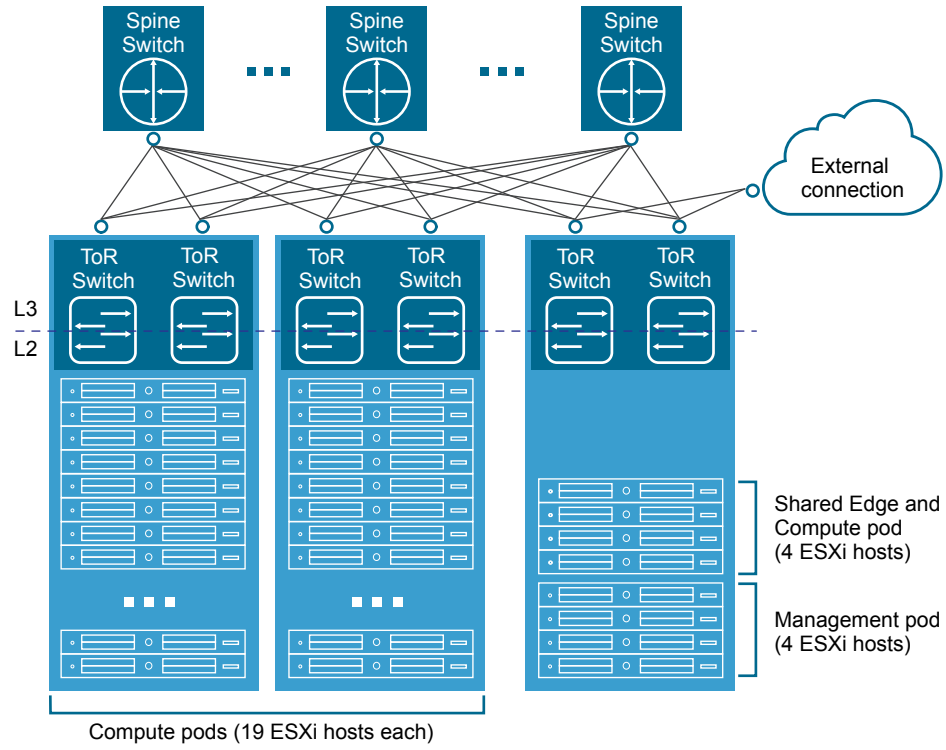
- Number of racks that are supported in a fabric.
- Amount of bandwidth between any two racks in a data center.
- Number of paths a leaf switch can select from when communicating with another rack.

The total number of available ports dictates the number of racks supported in a fabric across all spine switches and the acceptable level of oversubscription.

Different racks might be hosting different types of infrastructure. For example, a rack might host filers or other storage systems, which might attract or source more traffic than other racks in a data center. In addition, traffic levels of compute racks (that is, racks that are hosting hypervisors with workloads or virtual machines) might have different bandwidth requirements than edge racks, which provide connectivity to the outside world. Link speed as well as the number of links vary to satisfy different bandwidth demands.

The number of links to the spine switches dictates how many paths are available for traffic from this rack to another rack. Because the number of hops between any two racks is consistent, equal-cost multipathing (ECMP) can be used. Assuming traffic sourced by the servers carry a TCP or UDP header, traffic distribution can occur on a per-flow basis.

Figure 6-4. Leaf-and-Spine and Network Virtualization



Switch Types and Network Connectivity

Setup of the physical environment requires careful consideration. Follow best practices for physical switches, leaf switch connectivity, VLANs and subnets, and access port settings.

Top of Rack Physical Switches

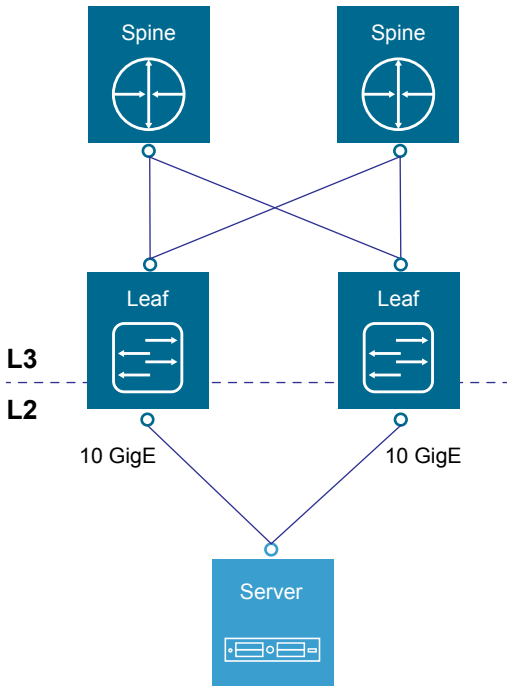
When configuring Top of Rack (ToR) switches, consider the following best practices.

- Configure redundant physical switches to enhance availability.
- Configure switch ports that connect to ESXi hosts manually as trunk ports. Virtual switches are passive devices and do not send or receive trunking protocols, such as Dynamic Trunking Protocol (DTP).
- Modify the Spanning Tree Protocol (STP) on any port that is connected to an ESXi NIC to reduce the time it takes to transition ports over to the forwarding state, for example using the Trunk PortFast feature found in a Cisco physical switch.
- Provide DHCP or DHCP Helper capabilities on all VLANs that are used by Management and VXLAN VMkernel ports. This setup simplifies the configuration by using DHCP to assign IP address based on the IP subnet in use.
- Configure jumbo frames on all switch ports, inter-switch link (ISL) and switched virtual interfaces (SVI's).

Leaf Switch Connectivity and Network Settings

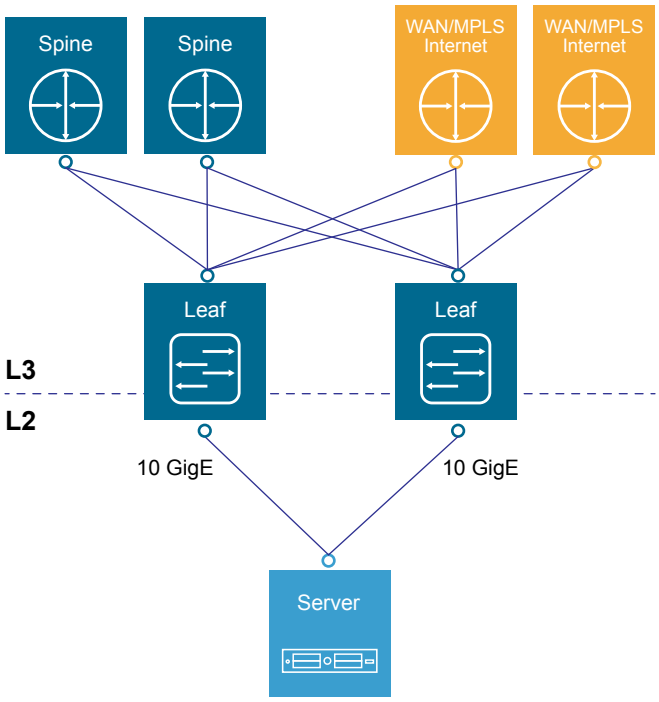
Each ESXi host in the compute rack is connected redundantly to the SDDC network fabric ToR switches by means of two 10 GbE ports, as shown in Figure 6-5. Configure the ToR switches to provide all necessary VLANs via an 802.1Q trunk.

Figure 6-5. Leaf Switch to Server Connection within Compute Racks



Each ESXi host in the management/shared edge and compute rack is connected to the SDDC network fabric and also to the Wide Area Network (WAN) and to the Internet, as shown in Figure 6-6.

Figure 6-6. Leaf Switch to Server Connection within Management/Shared Compute and Edge Rack



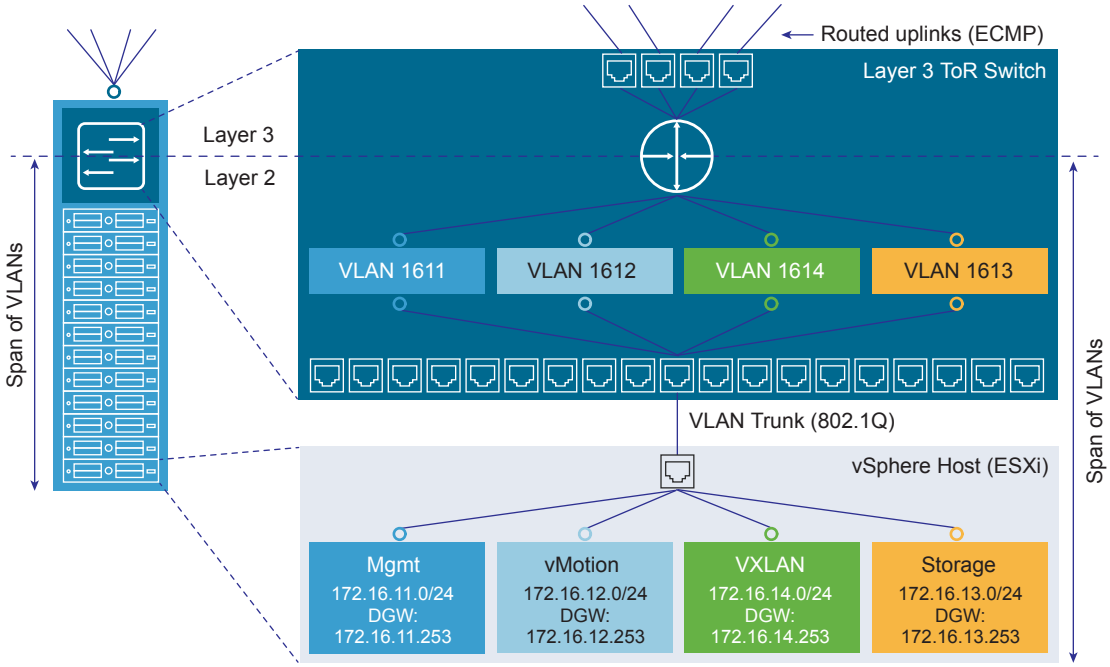
VLANs and Subnets

Each ESXi host in the compute rack and the management/edge rack uses VLANs and corresponding subnets for internal-only traffic, as shown in Figure 6-7.

The leaf switches of each rack act as the Layer 3 interface for the corresponding subnet.

The management/edge rack provides externally accessible VLANs for access to the Internet and/or MPL-based corporate networks.

Figure 6-7. Sample VLANs and Subnets within a Pod



Follow these guidelines.

- Use only /24 subnets to reduce confusion and mistakes when dealing with IPv4 subnetting.
- Use the IP address .1 as the (floating) interface with .2 and .3 for Virtual Router Redundancy Protocol (VRPP) or Hot Standby Routing Protocol (HSRP).
- Use the RFC1918 IPv4 address space for these subnets and allocate one octet by region and another octet by function. For example, the mapping 172.*regionid*.*function*.0/24 results in the following sample subnets.

Note The following VLANs and IP ranges are meant as samples. Your actual implementation depends on your environment.

Pod	Function	Sample VLAN	Sample IP range
Management	Management	1611 (Native)	172.16.11.0/24
Management	vMotion	1612	172.16.12.0/24
Management	VXLAN	1614	172.16.14.0/24
Management	VSAN	1613	172.16.13.0/24
Shared Edge and Compute	Management	1631 (Native)	172.16.31.0/24
Shared Edge and Compute	vMotion	1632	172.16.32.0/24
Shared Edge and Compute	VXLAN	1634	172.16.34.0/24
Shared Edge and Compute	VSAN	1633	172.16.33.0/24

Access Port Network Settings

Configure additional network settings on the access ports that connect the leaf switch to the corresponding servers.

Spanning-Tree Protocol (STP)	Although this design does not use the spanning tree protocol, switches usually come with STP configured by default. Designate the access ports as trunk PortFast.
Trunking	Configure the VLANs as members of a 802.1Q trunk with the management VLAN acting as the native VLAN.
MTU	Set MTU for all VLANS and SVIs (Management, vMotion, VXLAN and Storage) to jumbo frames for consistency purposes.
DHCP helper	Configure the VIF of the Management, vMotion and VXLAN subnet as a DHCP proxy.
Multicast	Configure IGMP snooping on the ToR switches and include an IGMP querier on each VLAN.

Region Interconnectivity

The SDDC management networks, VXLAN kernel ports and the edge and compute VXLAN kernel ports of the two regions must be connected. These connections can be over a VPN tunnel, Point to Point circuits, MPLS, etc. End users must be able to reach the public-facing network segments (public management and tenant networks) of both regions.

The region interconnectivity design must support jumbo frames, and ensure latency is less then 150 ms. For more details on the requirements for region interconnectivity see the *Cross-VC NSX Design Guide*.

The design of a region connection solution is out of scope for this VMware Validated Design.

Physical Network Design Decisions

The physical network design decisions govern the physical layout and use of VLANs. They also include decisions on jumbo frames and on some other network-related requirements such as DNS and NTP.

Physical Network Design Decisions

The design uses 4 spine switches with 40 GbE ports. As a result, each leaf switch must have 4 uplink ports capable of 40 GbE.

The resulting environment supports fault tolerance and compensates for oversubscription, as follows.

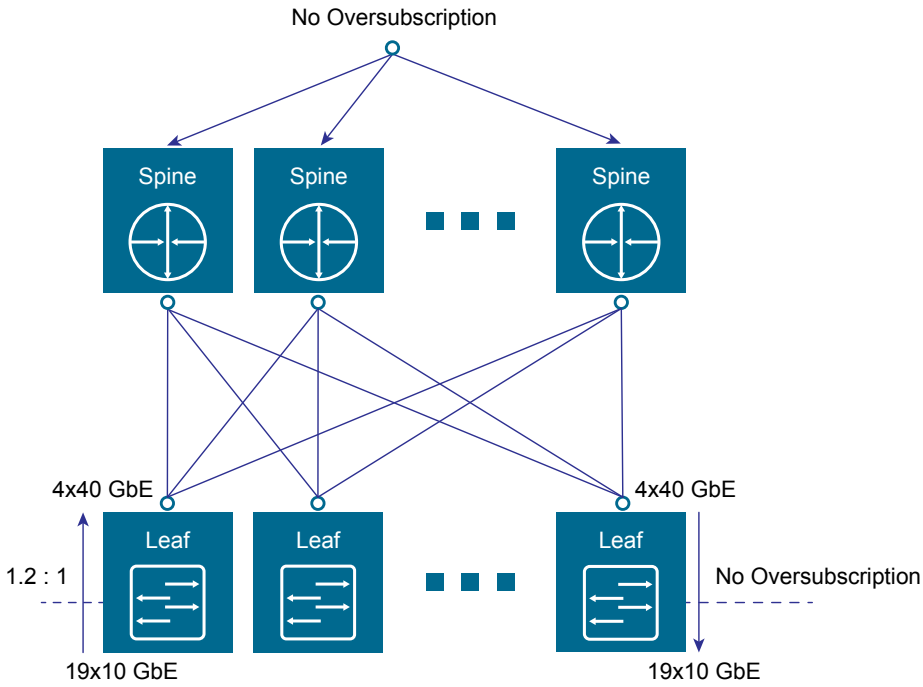
Fault Tolerance	In case of a switch failure or scheduled maintenance, switch fabric capacity reduction is 25% with four spine switches.
Oversubscription	Oversubscription can occur within a leaf switch. To compute the oversubscription for a leaf switch, use this formula.

```
Total bandwidth available to all connected servers /
aggregate amount of uplink bandwidth
```

The compute rack and the management/edge rack have 19 ESXi hosts. Each ESXi host has one 10 GbE port connected to each ToR switch, creating up to 190 Gbps of bandwidth. With four 40 GbE uplinks to the spine, you can compute oversubscription as follows (see Figure 6-8).

```
190 Gbps (total bandwidth)  / 160 Gbps (uplink bandwidth) =  1.2:1
```

Figure 6-8. Oversubscription in the Leaf Switches



Routing protocols

Base the selection of the external routing protocol on your current implementation or on available expertise among the IT staff. Take performance requirements into consideration. Possible options are OSPF, BGP and IS-IS.

DHCP proxy

The DHCP proxy must point to a DHCP server by way of its IPv4 address. See the *Planning and Preparation* documentation for details on the DHCP server.

Table 6-6. Physical Network Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-NET-001	Racks are connected using a leaf-and-spine topology and Layer 3 connectivity.	A Layer 3 leaf-and-spine architecture supports scale out while maintaining failure isolation.	Layer 2 traffic is reduced to within the pod.
SDDC-PHY-NET-002	Only the management and shared edge and compute hosts have physical access to the external network by way of VLANs.	Aggregating physical cabling and network services to the management and shared edge and compute rack reduces costs.	Workloads in compute pods located in compute racks have to use network virtualization (NSX for vSphere) for external network connectivity.
SDDC-PHY-NET-003	Each rack uses two ToR switches. These switches provide connectivity across two 10 GbE links to each server.	This design uses two 10 GbE links to provide redundancy and reduce overall design complexity.	Requires two ToR switches per rack which can increase costs.
SDDC-PHY-NET-004	Use VLANs to segment physical network functions.	Allow for Physical network connectivity without requiring large number of NICs. Segregation is needed for the different network functions that are required in the SDDC. This segregation allows for differentiated services and prioritization of traffic as needed.	Uniform configuration and presentation is required on all the trunks made available to the ESXi hosts.

Additional Design Decisions

Additional design decisions deal with static IP addresses, DNS records, and the required NTP time source.

Table 6-7. Additional Network Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-NET-005	Assign Static IP addresses to all management nodes of the SDDC infrastructure.	Configuration of static IP addresses avoid connection outages due to DHCP availability or misconfiguration.	Accurate IP address management must be in place.
SDDC-PHY-NET-006	Create DNS records for all management nodes to enable forward, reverse, short and FQDN resolution.	Ensures consistent resolution of management nodes using both IP address (reverse lookup) and name resolution.	None
SDDC-PHY-NET-007	Use an NTP time source for all management nodes.	Critical to maintain accurate and synchronized time between management nodes.	None

Jumbo Frames Design Decisions

IP storage throughput can benefit from the configuration of jumbo frames. Increasing the per-frame payload from 1500 bytes to the jumbo frame setting increases the efficiency of data transfer. Jumbo frames must be configured end-to-end, which is easily accomplished in a LAN. When you enable jumbo frames on an ESXi host, you have to select an MTU that matches the MTU of the physical switch ports.

The workload determines whether it makes sense to configure jumbo frames on a virtual machine. If the workload consistently transfers large amounts of network data, configure jumbo frames if possible. In that case, the virtual machine operating systems and the virtual machine NICs must also support jumbo frames.

Using jumbo frames also improves performance of vSphere vMotion.

Note VXLANs need an MTU value of at least 1600 bytes on the switches and routers that carry the transport zone traffic.

Table 6-8. Jumbo Frames Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-NET-008	Configure the MTU size to 9000 bytes (Jumbo Frames) on the portgroups that support the following traffic types. ■ NFS ■ vSAN ■ vMotion ■ VXLAN ■ vSphere Replication	Setting the MTU to 9000 bytes (Jumbo Frames) improves traffic throughput. In order to support VXLAN the MTU setting must be increased to a minimum of 1600 bytes, setting this portgroup also to 9000 bytes has no affect on VXLAN but ensures consistency across portgroups that are adjusted from the default MTU size.	When adjusting the MTU packet size, the entire network path (VMkernel port, distributed switch, physical switches and routers) must also be configured to support the same MTU packet size.

Physical Storage Design

This VMware Validated Design relies on both VMware Virtual SAN storage and NFS storage. The Shared Storage Design section explains where the SDDC uses which type of storage and gives background information. The focus of this section is physical storage design.

vSAN Physical Design

Software-defined storage is a key technology in the SDDC. This design uses VMware Virtual SAN (vSAN) to implement software-defined storage for the management clusters.

vSAN is a fully integrated hypervisor-converged storage software. vSAN creates a cluster of server hard disk drives and solid state drives, and presents a flash-optimized, highly resilient, shared storage datastore to hosts and virtual machines. vSAN allows you to control capacity, performance, and availability on a per virtual machine basis through the use of storage policies.

Requirements and Dependencies

The software-defined storage module has the following requirements and options.

- Minimum of 3 hosts providing storage resources to the vSAN cluster.
- vSAN is configured as hybrid storage or all-flash storage.
  - A vSAN hybrid storage configuration requires both magnetic devices and flash caching devices.
  - An All-Flash vSAN configuration requires vSphere 6.0 or later.

- Each ESXi host that provides storage resources to the cluster must meet the following requirements.
  - Minimum of one SSD. The SSD flash cache tier should be at least 10% of the size of the HDD capacity tier.
  - Minimum of two HHDs.
  - RAID controller compatible with vSAN.
  - 10 Gbps network for vSAN traffic \ with Multicast enabled.
  - vSphere High Availability Isolation Response set to power off virtual machines. With this setting, no possibility of split brain conditions in case of isolation or network partition exists. In a split-brain condition, the virtual machine might be powered on by two hosts by mistake. See design decision SDDC-VI-VC-012 for more details.

Table 6-9. vSAN Physical Storage Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-STO-001	Use one or more 200 GB or greater SSD and two or more traditional 1 TB or greater HDDs to create at least a single disk group in the management cluster.	Using a 200 GB SSD and two 1 TB HDD's allows enough capacity for the management VMs with a minimum of 10% flash-based caching.	When using only a single disk group you limit the amount of striping (performance) capability and increase the size of the fault domain.

Hybrid Mode and All-Flash Mode

vSphere offers two different vSAN modes of operation, all-flash or hybrid.

Hybrid Mode In a hybrid storage architecture, vSAN pools server-attached capacity devices (in this case magnetic devices) and caching devices, typically SSDs or PCI-e devices to create a distributed shared datastore.

All-Flash Mode vSAN can be deployed as all-flash storage. All-flash storage uses flash-based devices (SSD or PCI-e) only as a write cache while other flash-based devices provide high endurance for capacity and data persistence.

Table 6-10. vSAN Mode Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-STO-002	Configure vSAN in hybrid mode in the management cluster.	The VMs in the management cluster, which are hosted within vSAN, do not require the performance or expense of an all-flash vSAN configuration.	vSAN hybrid mode does not provide the potential performance or additional capabilities such as deduplication of an all-flash configuration.

Hardware Considerations

You can build your own VMware vSAN cluster or choose from a list of vSAN Ready Nodes.

Build Your Own

Be sure to use hardware from the VMware Compatibly Guide for the following vSAN components:

- Solid state disks (SSDs)
- Magnetic hard drives (HDDs)
- I/O controllers, including vSAN certified driver/firmware combinations

Use VMware vSAN Ready Nodes

A vSAN Ready Node is a validated server configuration in a tested, certified hardware form factor for vSAN deployment, jointly recommended by the server OEM and VMware. See the VMware Compatibility Guide. The vSAN Ready Node documentation provides examples of standardized configurations, including the numbers of VMs supported and estimated number of 4K IOPS delivered.

As per design decision SDDC-PHY-009, the VMware Validated Design uses vSAN Ready Nodes.

Solid State Disk (SSD) Characteristics

In a VMware vSAN configuration, the SSDs are used for the vSAN caching layer for hybrid deployments and for the capacity layer for all flash.

- For a hybrid deployment, the use of the SSD is split between a non-volatile write cache (approximately 30%) and a read buffer (approximately 70%). As a result, the endurance and the number of I/O operations per second that the SSD can sustain are important performance factors.
- For an all-flash model, endurance and performance have the same criteria. However many more write operations are held by the caching tier, thus elongating or extending the life of the SSD capacity-tier.

SSD Endurance

This VMware Validated Design uses class D endurance class SSDs for the caching tier.

SDDC Endurance Design Decision Background

For endurance of the SSDs used for vSAN, standard industry write metrics are the primary measurements used to gauge the reliability of the drive. No standard metric exists across all vendors, however, Drive Writes per Day (DWPD) or Petabytes Written (PBW) are the measurements normally used.

For vSphere 5.5, the endurance class was based on Drive Writes Per Day (DWPD). For VMware vSAN 6.0 and later, the endurance class has been updated to use Terabytes Written (TBW), based on the vendor's drive warranty. TBW can be used for VMware vSAN 5.5, VMware vSAN 6.0 and VMware vSAN 6.5 and is reflected in the *VMware Compatibility Guide*.

The reasoning behind using TBW is that VMware provides the flexibility to use larger capacity drives with lower DWPD specifications.

If a SSD vendor uses Drive Writes Per Day as a measurement, you can calculate endurance in Terabytes Written (TBW) with the following equation.

```
TBW (over 5 years) = Drive Size x DWPD x 365 x 5
```

For example, if a vendor specified DWPD = 10 for a 800 GB capacity SSD, you can compute TBW with the following equation.

```
TBW = 0.4TB X 10DWPD X 365days X 5yrs
TBW = 7300TBW
```

That means the SSD supports 7300TB writes over 5 years (The higher the TBW number, the greater the endurance class.).

For SSDs that are designated for caching and all-flash capacity layers, the following table outlines which endurance class to use for hybrid and for all-flash VMware vSAN.

Endurance Class	TBW	Hybrid Caching Tier	All-Flash Caching Tier	All-Flash Capacity Tier
Class A	>=365	No	No	Yes
Class B	>=1825	Yes	No	Yes
Class C	>=3650	Yes	Yes	Yes
Class D	>=7300	Yes	Yes	Yes

Note This VMware Validated Design does not use All-Flash vSAN.

Table 6-11. SSD Endurance Class Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-STO-003	Use Class D (>=7300TBW) SSDs for the caching tier of the management cluster.	If a SSD designated for the caching tier fails due to wear-out, the entire VMware vSAN disk group becomes unavailable. The result is potential data loss or operational impact.	SSDs with higher endurance may be more expensive than lower endurance classes.

SSD Performance

There is a direct correlation between the SSD performance class and the level of vSAN performance. The highest-performing hardware results in the best performance of the solution. Cost is therefore the determining factor. A lower class of hardware that is more cost effective might be attractive even if the performance or size is not ideal.

For optimal performance of vSAN, select class E or greater SSDs. See the VMware Compatibility Guide for detail on the different classes.

SSD Performance Design Decision Background

Select a high class of SSD for optimal performance of VMware vSAN. Before selecting a drive size, consider disk groups and sizing as well as expected future growth. VMware defines classes of performance in the VMware Compatibility Guide as follows.

Table 6-12. SSD Performance Classes

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

Select an SSD size that is, at a minimum, 10% of the anticipated size of the consumed HDD storage capacity, before failures to tolerate are considered. For example, select an SSD of at least 100 GB for 1 TB of HDD storage consumed in a 2 TB disk group.

Caching Algorithm

Both hybrid clusters and all-flash configurations adhere to the recommendation that 10% of consumed capacity for the flash cache layer. However, there are differences between the two configurations.

Hybrid vSAN — 70% of the available cache is allocated for storing frequently read disk blocks, minimizing accesses to the slower magnetic disks. 30% of available cache is allocated to writes.

All-Flash vSAN — All-flash clusters have two types of flash: very fast and durable write cache, and cost-effective capacity flash. Here cache is 100% allocated for writes, as read performance from capacity flash is more than sufficient.

Use Class E SSDs or greater for the highest possible level of performance from the VMware vSAN volume.

Table 6-13. SSD Performance Class Selection

Design Quality	Option 1 Class E	Option 2 Class C	Comments
Availability	o	o	Neither design option impacts availability.
Manageability	o	o	Neither design option impacts manageability.
Performance	↑	↓	The higher the storage class that is used, the better the performance.
Recover-ability	o	o	Neither design option impacts recoverability.
Security	o	o	Neither design option impacts security.

Legend: ↑ = positive impact on quality; ↓ = negative impact on quality; o = no impact on quality.

Table 6-14. SSD Performance Class Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-STO-004	Use Class E SSDs (30,000-100,000 writes per second) for the management cluster.	The storage I/O performance requirements within the Management cluster dictate the need for at least Class E SSDs.	Class E SSDs might be more expensive than lower class drives.

Magnetic Hard Disk Drives (HDD) Characteristics

The HDDs in a VMware vSAN environment have two different purposes, capacity and object stripe width.

Capacity Magnetic disks, or HDDs, unlike caching-tier SSDs, make up the capacity of a vSAN datastore

Stripe Width You can define stripe width at the virtual machine policy layer. vSAN might use additional stripes when making capacity and placement decisions outside a storage policy.

vSAN supports these disk types:

- Serial Attached SCSI (SAS)
- Near Line Serial Attached SCSI (NL-SCSI). NL-SAS can be thought of as enterprise SATA drives but with a SAS interface.
- Serial Advanced Technology Attachment (SATA). Use SATA magnetic disks only in capacity-centric environments where performance is not prioritized.

SAS and NL-SAS get you the best results. This VMware Validated Design uses 10,000 RPM drives to achieve a balance between cost and availability.

HDD Capacity, Cost, and Availability Background Considerations

You can achieve the best results with SAS and NL-SAS.

The VMware vSAN design must consider the number of magnetic disks required for the capacity layer, and how well the capacity layer will perform.

- SATA disks typically provide more capacity per individual drive, and tend to be less expensive than SAS drives. However the trade off is performance, because SATA performance is not as good as SAS performance due to lower rotational speeds (typically 7200RPM)
- Choose SAS magnetic disks instead of SATA magnetic disks in environments where performance is critical.

Consider that failure of a larger capacity drive has operational impact on the availability and recovery of more components.

Rotational Speed (RPM) Background Considerations

HDDs tend to be more reliable, but that comes at a cost. SAS disks can be available up to 15,000 RPM speeds.

Table 6-15. vSAN HDD Environmental Characteristics

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

Cache-friendly workloads are less sensitive to disk performance characteristics; however, workloads can change over time. HDDs with 10,000 RPM are the accepted norm when selecting a capacity tier.

For the software-defined storage module, VMware recommends that you use an HDD configuration that is suited to the characteristics of the environment. If there are no specific requirements, selecting 10,000 RPM drives achieves a balance between cost and availability.

Table 6-16. HDD Selection Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-STO-005	Use 10,000 RPM HDDs for the management cluster.	10,000 RPM HDDs achieve a balance between performance and availability for the VMware vSAN configuration. The performance of 10,000 RPM HDDs avoids disk drain issues. In vSAN hybrid mode, the vSAN periodically flushes uncommitted writes to the capacity tier.	Slower and potentially cheaper HDDs are not available.

I/O Controllers

The I/O controllers are as important to a VMware vSAN configuration as the selection of disk drives. vSAN supports SAS, SATA, and SCSI adapters in either pass-through or RAID 0 mode. vSAN supports multiple controllers per host.

- Multiple controllers can improve performance and mitigate a controller or SSD failure to a smaller number of drives or vSAN disk groups.
- With a single controller, all disks are controlled by one device. A controller failure impacts all storage, including the boot media (if configured).

Controller queue depth is possibly the most important aspect for performance. All I/O controllers in the *VMware vSAN Hardware Compatibility Guide* have a minimum queue depth of 256. Consider normal day-to-day operations and increase of I/O due to Virtual Machine deployment operations or re-sync I/O activity as a result of automatic or manual fault remediation.

About SAS Expanders

SAS expanders are a storage technology that lets you maximize the storage capability of your SAS controller card. Like switches of an ethernet network, SAS expanders enable you to connect a larger number of devices, that is, more SAS/SATA devices to a single SAS controller. Many SAS controllers support up to 128 or more hard drives.

Caution VMware has not extensively tested SAS expanders, as a result performance and operational predictability are relatively unknown at this point. For this reason, you should avoid configurations with SAS expanders.

NFS Physical Storage Design

Network File System (NFS) is a distributed file system protocol that allows a user on a client computer to access files over a network much like local storage is accessed. In this case the client computer is an ESXi host, and the storage is provided by a NFS-capable external storage array.

The management cluster uses VMware vSAN for primary storage and NFS for secondary storage. The compute clusters are not restricted to any particular storage technology. For compute clusters, the decision on which technology to use is based on the performance, capacity, and capabilities (replication, deduplication, compression, etc.) required by the workloads that are running in the clusters.

Table 6-17. NFS Usage Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-STO-006	NFS storage is presented to provide the following features. ■ A datastore for backup data ■ An export for archive data ■ A datastore for templates and ISOs	Separate primary virtual machine storage from backup data in case of primary storage failure. vRealize Log Insight archiving requires a NFS export.	An NFS capable external array is required.

Requirements

Your environment must meet the following requirements to use NFS storage in the VMware Validated Design.

- Storage arrays are connected directly to the leaf switches.
- All connections are made using 10 Gb Ethernet.
- Jumbo Frames are enabled.
- 10K SAS (or faster) drives are used in the storage array.

Different disk speeds and disk types can be combined in an array to create different performance and capacity tiers. The management cluster uses 10K SAS drives in the RAID configuration recommended by the array vendor to achieve the required capacity and performance.

Table 6-18. NFS Hardware Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-STO-007	Use 10K SAS drives for NFS volumes.	10K SAS drives achieve a balance between performance and capacity. Faster drives can be used if desired. vRealize Log Insight uses NFS datastores for its archive storage which, depending on compliance regulations, can use a large amount of disk space.	10K SAS drives are generally more expensive than other alternatives.

Volumes

A volume consists of multiple disks in a storage array that RAID is applied to.

Multiple datastores can be created on a single volume, but for applications that do not have a high I/O footprint a single volume with multiple datastores is sufficient.

- For high I/O applications, such as backup applications, use a dedicated volume to avoid performance issues.
- For other applications, set up Storage I/O Control (SIOC) to impose limits on high I/O applications so that other applications get the I/O they are requesting.

Table 6-19. Volume Assignment Design Decisions

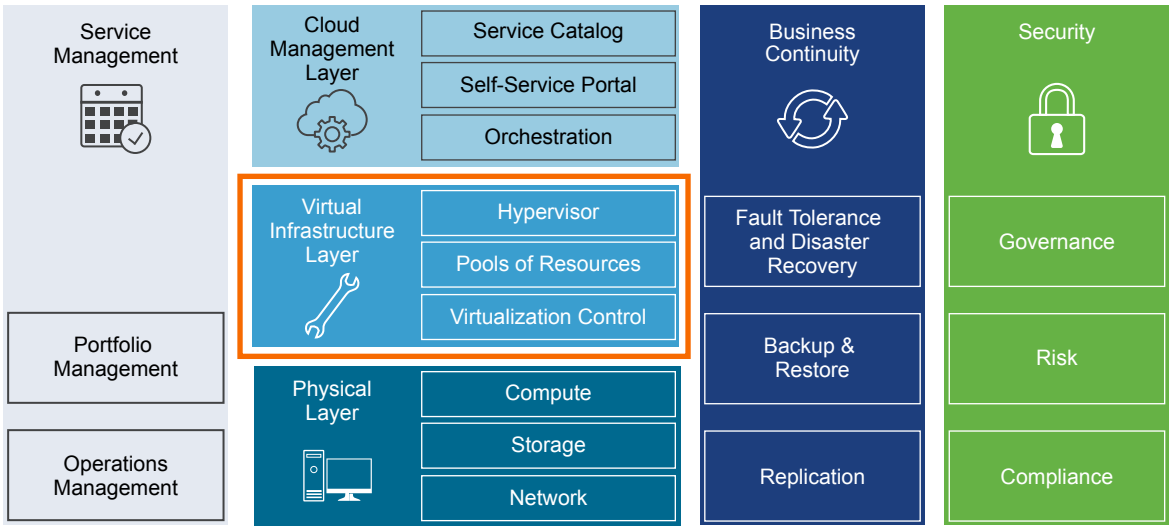
Decision ID	Design Decision	Design Justification	Design Implication
SDDC-PHY-STO-008	Use a dedicated NFS volume to support backup requirements.	The backup and restore process is I/O intensive. Using a dedicated NFS volume ensures that the process does not impact the performance of other management components.	Dedicated volumes add management overhead to storage administrators. Dedicated volumes might use more disks, depending on the array and type of RAID.
SDDC-PHY-STO-009	Use a shared volume for other management component datastores.	Non-backup related management applications can share a common volume due to the lower I/O profile of these applications.	Enough storage space for shared volumes and their associated application data must be available.

7 Virtual Infrastructure Design

The virtual infrastructure design includes the software components that make up the virtual infrastructure layer and that support the business continuity of the SDDC.

These components include the software products that provide the virtualization platform hypervisor, virtualization management, storage virtualization, network virtualization, backup and disaster recovery. VMware products in this layer include VMware vSphere, VMware Virtual SAN, and VMware NSX. If you expand the design to dual region, it also includes vSphere Data Protection, and VMware Site Recovery Manager.

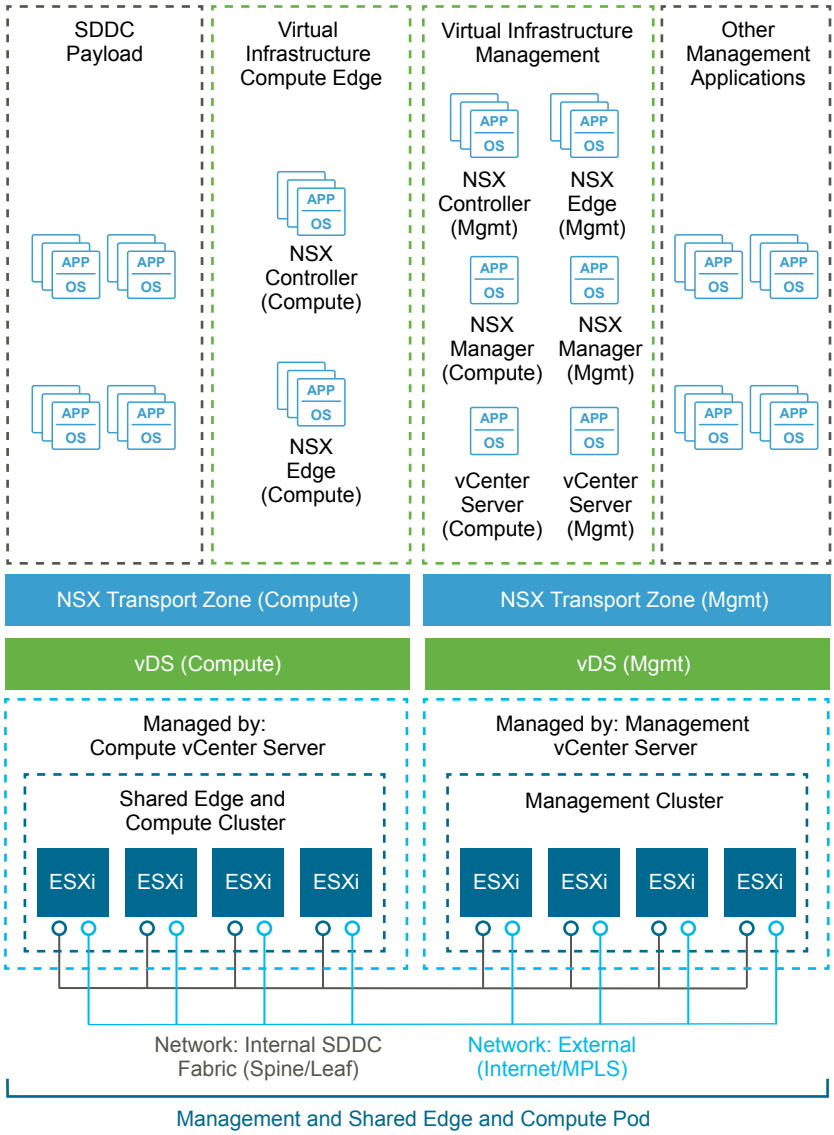
Figure 7-1. Virtual Infrastructure Layer in the SDDC



Virtual Infrastructure Design Overview

Each region includes a management pod, and a shared edge and compute pod.

Figure 7-2. Logical Design



Management Pod

Management pods run the virtual machines that manage the SDDC. These virtual machines host vCenter Server, NSX Manager, NSX Controller, vRealize Log Insight, and other shared management components. All management, monitoring, and infrastructure services are provisioned to a vSphere cluster which provides high availability for these critical services. Permissions on the management cluster limit access to only administrators. This protects the virtual machines running the management, monitoring, and infrastructure services.

Shared Edge and Compute Pod

The virtual infrastructure design uses a shared edge and compute pod. The shared pod combines the characteristics of typical edge and compute pods into a single pod. It is possible to separate these in the future if required.

This pod provides the following main functions:

- Supports on-ramp and off-ramp connectivity to physical networks
- Connects with VLANs in the physical world
- Hosts the SDDC tenant virtual machines

The shared edge and compute pod connects the virtual networks (overlay networks) provided by NSX for vSphere and the external networks. An SDDC can mix different types of compute-only pods and provide separate compute pools for different types of SLAs.

This chapter includes the following topics:

ESXi Design

The ESXi design includes design decisions for boot options, user access, and the virtual machine swap configuration.

ESXi Hardware Requirements

You can find the ESXi hardware requirements in Physical Design Fundamentals. The following design outlines the design of the ESXi configuration.

ESXi Manual Install and Boot Options

You can install or boot ESXi 6.5 from the following storage systems:

SATA disk drives	SATA disk drives connected behind supported SAS controllers or supported on-board SATA controllers.
Serial-attached SCSI (SAS) disk drives	Supported for installing ESXi.
SAN	Dedicated SAN disk on Fibre Channel or iSCSI.
USB devices	Supported for installing ESXi. 16 GB or larger SD card is recommended.
FCoE	(Software Fibre Channel over Ethernet)

ESXi can boot from a disk larger than 2 TB if the system firmware and the firmware on any add-in card support it. See the vendor documentation.

ESXi Boot Disk and Scratch Configuration

For new installations of ESXi, the installer creates a 4 GB VFAT scratch partition. ESXi uses this scratch partition to store log files persistently. By default, vm-support output, which is used by VMware to troubleshoot issues on the ESXi host, is also stored on the scratch partition.

An ESXi installation on USB media does not configure a default scratch partition. VMware recommends that you specify a scratch partition on a shared datastore and configure remote syslog logging for the host.

Table 7-1. ESXi Boot Disk Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-ESXi-001	Install and configure all ESXi hosts to boot using a SD device of 16 GB or greater.	SD cards are an inexpensive and easy to configure option for installing ESXi. Using SD cards allows allocation of all local HDDs to a VMware vSAN storage system.	When you use SD cards ESXi logs are not retained locally.

ESXi Host Access

After installation, ESXi hosts are added to a VMware vCenter Server system and managed through that vCenter Server system.

Direct access to the host console is still available and most commonly used for troubleshooting purposes. You can access ESXi hosts directly using one of these three methods:

Direct Console User Interface (DCUI)	Graphical interface on the console. Allows basic administrative controls and troubleshooting options.
ESXi Shell	A Linux-style bash login on the ESXi console itself.
Secure Shell (SSH) Access	Remote command-line console access.

You can enable or disable each method. By default the ESXi Shell and SSH are disabled to secure the ESXi host. The DCUI is disabled only if Strict Lockdown Mode is enabled.

ESXi User Access

By default, root is the only user who can log in to an ESXi host directly, however, you can add ESXi hosts to an Active Directory domain. After the host has been added to an Active Directory domain, access can be granted through Active Directory groups. Auditing who has logged into the host also becomes easier.

Table 7-2. ESXi User Access Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-ESXi-002	Add each host to the Active Directory domain for the region in which it will reside.	Using Active Directory membership allows greater flexibility in granting access to ESXi hosts. Ensuring that users log in with a unique user account allows greater visibility for auditing.	Adding hosts to the domain can add some administrative overhead.
SDDC-VI-ESXi-003	Change the default ESX Admins group to the SDDC-Admins Active Directory group. Add ESXi administrators to the SDDC-Admins group following standard access procedures.	Having an SDDC-Admins group is more secure because it removes a known administrative access point. In addition different groups allow for separation of management tasks.	Additional changes to the host's advanced settings are required.

Virtual Machine Swap Configuration

When a virtual machine is powered on, the system creates a VMkernel swap file to serve as a backing store for the virtual machine's RAM contents. The default swap file is stored in the same location as the virtual machine's configuration file. This simplifies the configuration, however it can cause an excess of replication traffic that is not needed.

You can reduce the amount of traffic that is replicated by changing the swap file location to a user-configured location on the host. However, it can take longer to perform VMware vSphere vMotion® operations when the swap file has to be recreated.

Table 7-3. Other ESXi Host Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-ESXi-004	Configure all ESXi hosts to synchronize time with the central NTP servers.	Required because deployment of vCenter Server Appliance on an ESXi host might fail if the host is not using NTP.	All firewalls located between the ESXi host and the NTP servers have to allow NTP traffic on the required network ports.

vCenter Server Design

The vCenter Server design includes both the design for the vCenter Server instance and the VMware Platform Services Controller instance.

A Platform Services Controller groups a set of infrastructure services including vCenter Single Sign-On, License service, Lookup Service, and VMware Certificate Authority (VMCA). You can deploy the Platform Services Controller and the associated vCenter Server system on the same virtual machine(embedded Platform Services Controller) or on different virtual machines (external Platform Services Controller).

vCenter Server Deployment

The design decisions for vCenter Server deployment discuss the number of vCenter Server and Platform Services Controller instances, the type of installation, and the topology.

Table 7-4. vCenter Server Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-001	Deploy two vCenter Server systems in the first availability zone of each region. ■ One vCenter Server supporting the SDDC management components. ■ One vCenter Server supporting the edge components and compute workloads.	Isolates vCenter Server failures to management or compute workloads. Isolates vCenter Server operations between management and compute. Supports a scalable cluster design where the management components may be re-used as additional compute needs to be added to the SDDC. Simplifies capacity planning for compute workloads by eliminating management workloads from consideration in the Compute vCenter Server. Improves the ability to upgrade the vSphere environment and related components by providing for explicit separation of maintenance windows: ■ Management workloads remain available while workloads in compute are being addressed ■ Compute workloads remain available while workloads in management are being addressed Ability to have clear separation of roles and responsibilities to ensure that only those administrators with proper authorization can attend to the management workloads. Facilitates quicker troubleshooting and problem resolution. Simplifies Disaster Recovery operations by supporting a clear demarcation between recovery of the management components and compute workloads. Enables the use of two NSX managers, one for the management pod and the other for the shared edge and compute pod. Network separation of the pods in the SDDC allows for isolation of potential network issues.	Requires licenses for each vCenter Server instance.

You can install vCenter Server as a Windows-based system or deploy the Linux-based VMware vCenter Server Appliance. The Linux-based vCenter Server Appliance is preconfigured, enables fast deployment, and potentially results in reduced Microsoft licensing costs.

Table 7-5. vCenter Server Platform Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-002	Deploy all vCenter Server instances as Linux-based vCenter Server Appliances.	Allows for rapid deployment, enables scalability, and reduces Microsoft licensing costs.	Operational staff might need Linux experience to troubleshoot the Linux-based appliances.

Platform Services Controller Design Decision Background

vCenter Server supports installation with an embedded Platform Services Controller (embedded deployment) or with an external Platform Services Controller.

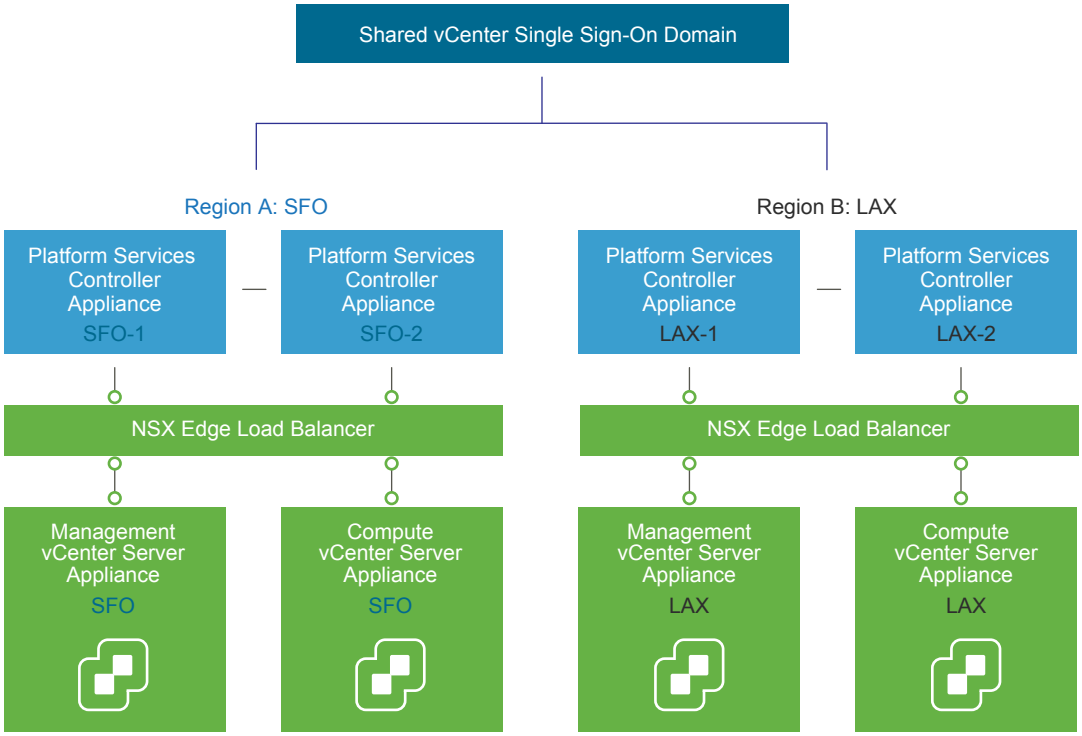
- In an embedded deployment, vCenter Server and the Platform Services Controller run on the same virtual machine. Embedded deployments are recommended for standalone environments with only one vCenter Server system.
- Environments with an external Platform Services Controller can have multiple vCenter Server systems. The vCenter Server systems can use the same Platform Services Controller services. For example, several vCenter Server systems can use the same instance of vCenter Single Sign-On for authentication.
- If there is a need to replicate with other Platform Services Controller instances, or if the solution includes more than one vCenter Single Sign-On instance, you can deploy multiple external Platform Services Controller instances on separate virtual machines.

Table 7-6. Platform Service Controller Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-003	Deploy each vCenter Server with an external Platform Services Controller.	External Platform Services Controllers are required for replication between Platform Services Controller instances.	The number of VMs that have to be managed increases.
SDDC-VI-VC-004	Join all Platform Services Controller instances to a single vCenter Single Sign-On domain.	When all Platform Services Controller instances are joined into a single vCenter Single Sign-On domain, they can share authentication and license data across all components and regions.	Only one Single Sign-On domain will exist.
SDDC-VI-VC-005	Create a ring topology for the Platform Service Controllers.	By default Platform Service Controllers only replicate with one other Platform Services Controller, that creates a single point of failure for replication. A ring topology ensures each Platform Service Controller has two replication partners and eliminates any single point of failure.	Command-line interface commands must be used to configure the ring replication topology.
SDDC-VI-VC-006	Use an NSX Edge Services Gateway as a load balancer for the Platform Services Controllers.	Using a load balancer increases the availability of the PSC's for all applications.	Configuring the load balancer and repointing vCenter Server to the load balancers Virtual IP (VIP) creates administrative overhead.

The following illustration shows the deployment model for a two-region design. For some of the use cases, you initially use a single-region design.

Figure 7-3. vCenter Server and Platform Services Controller Deployment Model



vCenter Server Networking

As specified in the physical networking design, all vCenter Server systems must use static IP addresses and host names. The IP addresses must have valid (internal) DNS registration including reverse name resolution.

The vCenter Server systems must maintain network connections to the following components:

- All VMware vSphere Client and vSphere Web Client user interfaces.
- Systems running vCenter Server add-on modules.
- Each ESXi host.

vCenter Server Redundancy

Protecting the vCenter Server system is important because it is the central point of management and monitoring for the SDDC. How you protect vCenter Server depends on maximum downtime tolerated, and on whether failover automation is required.

The following table lists methods available for protecting the vCenter Server system and the vCenter Server Appliance.

Table 7-7. Methods for Protecting vCenter Server System and the vCenter Server Appliance

Redundancy Method	Protects vCenter Server system (Windows)	Protects Platform Services Controller (Windows)	Protects vCenter Server (Appliance)	Protects Platform Services Controller (Appliance)
Automated protection using vSphere HA.	Yes	Yes	Yes	Yes
Manual configuration and manual failover. For example, using a cold standby.	Yes	Yes	Yes	Yes
HA Cluster with external load balancer	Not Available	Yes	Not Available	Yes
vCenter Server HA	Not Available	Not Available	Yes	Not Available

Table 7-8. vCenter Server Protection Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-007	Protect all vCenter Server and Platform Services Controller appliances by using vSphere HA.	Supports availability objectives for vCenter Server appliances without a required manual intervention during a failure event.	vCenter Server will be unavailable during a vSphere HA failover.

vCenter Server Appliance Sizing

The following tables outline minimum hardware requirements for the management vCenter Server appliance and the compute vCenter Server appliance.

Table 7-9. Logical Specification for Management vCenter Server Appliance

Attribute	Specification
vCenter Server version	6.5 (vCenter Server Appliance)
Physical or virtual system	Virtual (appliance)
Appliance Size	Small (up to 100 hosts / 1,000 VMs)
Platform Services Controller	External
Number of CPUs	4
Memory	16 GB
Disk Space	290 GB

Table 7-10. Logical Specification for Compute vCenter Server Appliance

Attribute	Specification
vCenter Server version	6.5 (vCenter Server Appliance)
Physical or virtual system	Virtual (appliance)
Appliance Size	Large (up to 1,000 hosts / 10,000 VMs)
Platform Services Controller	External
Number of CPUs	16

Table 7-10. Logical Specification for Compute vCenter Server Appliance (Continued)

Attribute	Specification
Memory	32 GB
Disk Space	640 GB

Table 7-11. vCenter Server Appliance Sizing Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-008	Configure the management vCenter Server Appliances with at least the small size setting.	Based on the number of management VMs that are running, a vCenter Server Appliance installed with the small size setting is sufficient.	If the size of the management environment changes, the vCenter Server Appliance size might need to be increased.
SDDC-VI-VC-009	Configure the compute vCenter Server Appliances with at least the large size setting.	Based on the number of compute workloads and NSX edge devices running, a vCenter Server Appliance installed with the large size setting is recommended.	As the compute environment grows resizing to X-Large or adding additional vCenter Server instances may be required.

vSphere Cluster Design

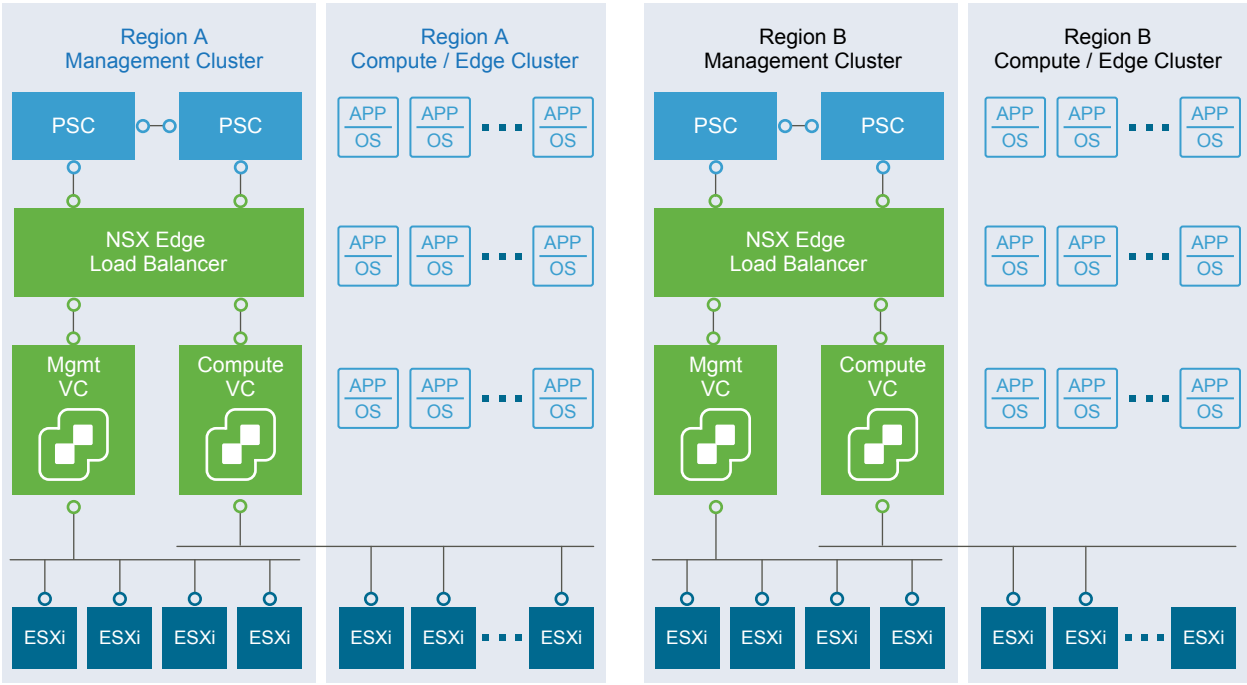
The cluster design must take into account the workload that the cluster handles. Different cluster types in this design have different characteristics.

vSphere Cluster Design Decision Background

The following heuristics help with cluster design decisions.

- Decide to use fewer, larger hosts or more, smaller hosts.
  - A scale-up cluster has fewer, larger hosts.
  - A scale-out cluster has more, smaller hosts.
  - A virtualized server cluster typically has more hosts with fewer virtual machines per host.
- Compare the capital costs of purchasing fewer, larger hosts with the costs of purchasing more, smaller hosts. Costs vary between vendors and models.
- Evaluate the operational costs of managing a few hosts with the costs of managing more hosts.
- Consider the purpose of the cluster.
- Consider the total number of hosts and cluster limits.

Figure 7-4. vSphere Logical Cluster Layout



vSphere High Availability Design

VMware vSphere High Availability (vSphere HA) protects your virtual machines in case of host failure by restarting virtual machines on other hosts in the cluster when a host fails.

vSphere HA Design Basics

During configuration of the cluster, the hosts elect a master host. The master host communicates with the vCenter Server system and monitors the virtual machines and secondary hosts in the cluster.

The master hosts detects different types of failure:

- Host failure, for example an unexpected power failure
- Host network isolation or connectivity failure
- Loss of storage connectivity
- Problems with virtual machine OS availability

Table 7-12. vSphere HA Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-010	Use vSphere HA to protect all clusters against failures.	vSphere HA supports a robust level of protection for both host and virtual machine availability.	Sufficient resources on the remaining host are required to so that virtual machines can be migrated to those hosts in the event of a host outage.
SDDC-VI-VC-011	Set vSphere HA Host Isolation Response to Power Off.	Virtual SAN requires that the HA Isolation Response be set to Power Off and to restart VMs on available hosts.	VMs are powered off in case of a false positive and a host is declared isolated incorrectly.

vSphere HA Admission Control Policy Configuration

The vSphere HA Admission Control Policy allows an administrator to configure how the cluster judges available resources. In a smaller vSphere HA cluster, a larger proportion of the cluster resources are reserved to accommodate host failures, based on the selected policy.

The following policies are available:

Host failures the cluster tolerates. vSphere HA ensures that a specified number of hosts can fail and sufficient resources remain in the cluster to fail over all the virtual machines from those hosts.

Percentage of cluster resources reserved. Percentage of cluster resources reserved. vSphere HA ensures that a specified percentage of aggregate CPU and memory resources are reserved for failover.

Specify Failover Hosts. When a host fails, vSphere HA attempts to restart its virtual machines on any of the specified failover hosts. If restart is not possible, for example the failover hosts have insufficient resources or have failed as well, then vSphere HA attempts to restart the virtual machines on other hosts in the cluster.

vSphere Cluster Workload Design

This design defines the following vSphere clusters and the workloads that they handle.

Table 7-13. vSphere Cluster Workload Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-012	Create a single management cluster containing all management hosts.	Simplifies configuration by isolating management workloads from compute workloads. Ensures that compute workloads have no impact on the management stack. You can add ESXi hosts to the cluster as needed.	Management of multiple clusters and vCenter Server instances increases operational overhead.
SDDC-VI-VC-013	Create a shared edge and compute cluster that hosts compute workloads, NSX Controllers and associated NSX Edge gateway devices used for compute workloads.	Simplifies configuration and minimizes the number of hosts required for initial deployment. Ensures that the management stack has no impact on compute workloads. You can add ESXi hosts to the cluster as needed.	Management of multiple clusters and vCenter Server instances increases operational overhead. Due to the shared nature of the cluster, when compute workloads are added, the cluster must be scaled out to keep high level of network performance. Due to the shared nature of the cluster, resource pools are required to ensure edge components receive all required resources.

Management Cluster Design

The management cluster design determines the number of hosts and vSphere HA settings for the management cluster.

Table 7-14. Management Cluster Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-014	Create a management cluster with 4 hosts.	Three hosts are used to provide n+1 redundancy for the Virtual SAN cluster. The fourth host is used to guarantee n+1 for Virtual SAN redundancy during maintenance operations.	Additional host resources are required for redundancy.
SDDC-VI-VC-015	Configure Admission Control for 1 host failure and percentage based failover capacity.	Using the percentage-based reservation works well in situations where virtual machines have varying and sometime significant CPU or memory reservations. vSphere 6.5 automatically calculates the reserved percentage based on host failures to tolerate and the number of hosts in the cluster.	In a four host management cluster only the resources of three hosts are available for use.
SDDC-VI-VC-016	Create a host profile for the Management Cluster.	Utilizing host profiles simplifies configuration of hosts and ensures settings are uniform across the cluster.	Anytime an authorized change to a host is made the host profile must be updated to reflect the change or the status will show non-compliant.

The following table summarizes the attributes of the management cluster logical design.

Table 7-15. Management Cluster Logical Design Background

Attribute	Specification
Number of hosts required to support management hosts with no over commitment .	2
Number of hosts recommended due to operational constraints (Ability to take a host offline without sacrificing High Availability capabilities) .	3
Number of hosts recommended due to operational constraints, while using Virtual SAN (Ability to take a host offline without sacrificing High Availability capabilities) .	4
Capacity for host failures per cluster.	25% reserved CPU RAM

Shared Edge and Compute Cluster Design

Tenant workloads run on the ESXi hosts in the shared edge and compute cluster. Due to the shared nature of the cluster, NSX Controllers and Edge devices run in this cluster. The design decisions determine the number of hosts and vSphere HA settings and several other characteristics of the shared edge and compute cluster.

Table 7-16. Shared Edge and Compute Cluster Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-017	Create a shared edge and compute cluster for the NSX Controllers and NSX Edge gateway devices.	NSX Manager requires a 1:1 relationship with a vCenter Server system.	Each time you provision a Compute vCenter Server system, a new NSX Manager is required. Set anti-affinity rules to keep each Controller on a separate host. A 4-node cluster allows maintenance while ensuring that the 3 Controllers remain on separate hosts.
SDDC-VI-VC-018	Configure Admission Control for 1 host failure and percentage based failover capacity.	vSphere HA protects the NSX Controller instances and edge services gateway devices in the event of a host failure. vSphere HA powers on virtual machines from the failed hosts on any remaining hosts.	Only a single host failure is tolerated before potential resource contention.
SDDC-VI-VC-019	Create shared edge and compute cluster with a minimum of 4 hosts.	■ 3 NSX Controllers are required for sufficient redundancy and majority decisions. ■ One host is available for failover and to allow for scheduled maintenance.	4 hosts is the smallest starting point for the shared edge and compute cluster for redundancy and performance thus increasing cost over a 3 node cluster.

Table 7-16. Shared Edge and Compute Cluster Design Decisions (Continued)

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-020	Set up VLAN-backed port groups for external access and management on the shared edge and compute cluster hosts.	Edge gateways need access to the external network in addition to the management network.	VLAN-backed port groups must be configured with the correct number of ports, or with elastic port allocation.
SDDC-VI-VC-021	Create a resource pool for the required SDDC NSX Controllers and edge appliances with a CPU share level of High, a memory share of normal, and 16 GB memory reservation.	The NSX components control all network traffic in and out of the SDDC as well as update route information for inter-SDDC communication. In a contention situation it is imperative that these virtual machines receive all the resources required.	During contention SDDC NSX components receive more resources then all other workloads as such monitoring and capacity management must be a proactive activity.

The following table summarizes the attributes of the shared edge and compute cluster logical design. The number of VMs on the shared edge and compute cluster will start low but will grow quickly as user workloads are created.

Table 7-17. Shared Edge and Compute Cluster Logical Design Background

Attribute	Specification
Minimum number of hosts required to support the shared edge and compute cluster	4
Capacity for host failures per cluster	1
Number of usable hosts per cluster	3

Compute Cluster Design

As the SDDC grows, additional compute-only clusters can be configured. Tenant workloads run on the ESXi hosts in the compute cluster instances. Multiple compute clusters are managed by the Compute vCenter Server instance. The design determines host-to-rack relationship and vSphere HA settings for the compute cluster.

Table 7-18. Compute Cluster Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-025	Configure vSphere HA to use percentage-based failover capacity to ensure n+1 availability.	Using explicit host failover limits the total available resources in a cluster.	The resources of one host in the cluster is reserved which can cause provisioning to fail if resources are exhausted.

vCenter Server Customization

vCenter Server supports a rich set of customization options, including monitoring, virtual machine fault tolerance, and so on. For each feature, this VMware Validated Design specifies the design decisions.

VM and Application Monitoring Service

When VM and Application Monitoring is enabled, the VM and Application Monitoring service, which uses VMware Tools, evaluates whether each virtual machine in the cluster is running. The service checks for regular heartbeats and I/O activity from the VMware Tools process running on guests. If the service receives no heartbeats or I/O activity, it is likely that the guest operating system has failed or that VMware Tools is not being allocated time for heartbeats or I/O activity. In this case, the service determines that the virtual machine has failed and reboots the virtual machine.

Enable Virtual Machine Monitoring for automatic restart of a failed virtual machine. The application or service that is running on the virtual machine must be capable of restarting successfully after a reboot or the VM restart is not sufficient.

Table 7-19. Monitor Virtual Machines Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-026	Enable Virtual Machine Monitoring for each cluster.	Virtual Machine Monitoring provides adequate in-guest protection for most VM workloads.	There is no downside to enabling Virtual Machine Monitoring.
SDDC-VI-VC-027	Create Virtual Machine Groups for use in startup rules in the management and shared edge and compute clusters.	By creating Virtual Machine groups, rules can be created to configure the startup order of the SDDC management components.	Creating the groups is a manual task and adds administrative overhead.
SDDC-VI-VC-028	Create Virtual Machine rules to specify the startup order of the SDDC management components.	The rules enforce the startup order of virtual machine groups to ensure the correct startup order of the SDDC management components.	Creating the rules is a manual task and adds administrative overhead.

VMware vSphere Distributed Resource Scheduling (DRS)

vSphere Distributed Resource Scheduling provides load balancing of a cluster by migrating workloads from heavily loaded hosts to less utilized hosts in the cluster. DRS supports manual and automatic modes.

Manual Recommendations are made but an administrator needs to confirm the changes

Automatic Automatic management can be set to five different levels. At the lowest setting, workloads are placed automatically at power on and only migrated to fulfill certain criteria, such as entering maintenance mode. At the highest level, any migration that would provide a slight improvement in balancing will be executed.

Table 7-20. vSphere Distributed Resource Scheduling Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-029	Enable DRS on all clusters and set it to Fully Automated, with the default setting (medium).	The default settings provide the best trade-off between load balancing and excessive migration with vMotion events.	In the event of a vCenter outage, mapping from virtual machines to ESXi hosts might be more difficult to determine.

Enhanced vMotion Compatibility (EVC)

EVC works by masking certain features of newer CPUs to allow migration between hosts containing older CPUs. EVC works only with CPUs from the same manufacturer and there are limits to the version difference gaps between the CPU families.

If you set EVC during cluster creation, you can add hosts with newer CPUs at a later date without disruption. You can use EVC for a rolling upgrade of all hardware with zero downtime.

Set EVC to the highest level possible with the current CPUs in use.

Table 7-21. VMware Enhanced vMotion Compatibility Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-030	Enable Enhanced vMotion Compatibility on all clusters. Set EVC mode to the lowest available setting supported for the hosts in the cluster.	Allows cluster upgrades without virtual machine downtime.	You can enable EVC only if clusters contain hosts with CPUs from the same vendor.

Use of Transport Layer Security (TLS) Certificates

By default vSphere 6.5 uses TLS/SSL certificates that are signed by VMCA (VMware Certificate Authority). By default, these certificates are not trusted by end-user devices or browsers. It is a security best practice to replace at least user-facing certificates with certificates that are signed by a third-party or enterprise Certificate Authority (CA). Certificates for machine-to-machine communication can remain as VMCA-signed certificates.

Table 7-22. vCenter Server TLS Certificate Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-VC-031	Replace the vCenter Server machine certificate and Platform Services Controller machine certificate with a certificate signed by a 3rd party Public Key Infrastructure.	Infrastructure administrators connect to both vCenter Server and the Platform Services Controller by way of s Web browser to perform configuration, management and troubleshooting activities. Certificate warnings result with the default certificate.	Replacing and managing certificates is an operational overhead.
SDDC-VI-VC-032	Use a SHA-2 or higher algorithm when signing certificates.	The SHA-1 algorithm is considered less secure and has been deprecated.	Not all certificate authorities support SHA-2.

Virtualization Network Design

A well-designed network helps the organization meet its business goals. It prevents unauthorized access, and provides timely access to business data.

This network virtualization design uses vSphere and VMware NSX for vSphere to implement virtual networking.

Virtual Network Design Guidelines

This VMware Validated Design follows high-level network design guidelines and networking best practices.

Design Goals

The high-level design goals apply regardless of your environment.

- Meet diverse needs. The network must meet the diverse needs of many different entities in an organization. These entities include applications, services, storage, administrators, and users.
- Reduce costs. Reducing costs is one of the simpler goals to achieve in the vSphere infrastructure. Server consolidation alone reduces network costs by reducing the number of required network ports and NICs, but a more efficient network design is desirable. For example, configuring two 10 GbE NICs with VLANs might be more cost effective than configuring a dozen 1 GbE NICs on separate physical networks.
- Boost performance. You can achieve performance improvement and decrease the time that is required to perform maintenance by providing sufficient bandwidth, which reduces contention and latency.
- Improve availability. A well-designed network improves availability, typically by providing network redundancy.
- Support security. A well-designed network supports an acceptable level of security through controlled access (where required) and isolation (where necessary).
- Enhance infrastructure functionality. You can configure the network to support vSphere features such as vSphere vMotion, vSphere High Availability, and vSphere Fault Tolerance.

Best Practices

Follow networking best practices throughout your environment.

- Separate network services from one another to achieve greater security and better performance.
- Use Network I/O Control and traffic shaping to guarantee bandwidth to critical virtual machines. During network contention these critical virtual machines will receive a higher percentage of the bandwidth.
- Separate network services on a single vSphere Distributed Switch by attaching them to port groups with different VLAN IDs.

- Keep vSphere vMotion traffic on a separate network. When migration with vMotion occurs, the contents of the guest operating system's memory is transmitted over the network. You can put vSphere vMotion on a separate network by using a dedicated vSphere vMotion VLAN.
- When using passthrough devices with a Linux kernel version 2.6.20 or earlier guest OS, avoid MSI and MSI-X modes because these modes have significant performance impact.
- For best performance, use VMXNET3 virtual NICs.
- Ensure that physical network adapters that are connected to the same vSphere Standard Switch or vSphere Distributed Switch are also connected to the same physical network.

Network Segmentation and VLANs

Separating different types of traffic is required to reduce contention and latency. Separate networks are also required for access security.

High latency on any network can negatively affect performance. Some components are more sensitive to high latency than others. For example, reducing latency is important on the IP storage and the vSphere Fault Tolerance logging network because latency on these networks can negatively affect the performance of multiple virtual machines.

Depending on the application or service, high latency on specific virtual machine networks can also negatively affect performance. Use information gathered from the current state analysis and from interviews with key stakeholder and SMEs to determine which workloads and networks are especially sensitive to high latency.

Virtual Networks

Determine the number of networks or VLANs that are required depending on the type of traffic.

- vSphere operational traffic.
  - Management
  - vMotion
  - Virtual SAN
  - NFS Storage
  - VXLAN
- Traffic that supports the organization's services and applications.

Virtual Switches

Virtual switches simplify the configuration process by providing one single pane of glass view for performing virtual network management tasks.

Virtual Switch Design Background

A vSphere Distributed Switch (distributed switch) offers several enhancements over standard virtual switches.

Centralized management	Because distributed switches are created and managed centrally on a vCenter Server system, they make the switch configuration more consistent across ESXi hosts. Centralized management saves time, reduces mistakes, and lowers operational costs.
Additional features	Distributed switches offer features that are not available on standard virtual switches. Some of these features can be useful to the applications and services that are running in the organization's infrastructure. For example, NetFlow and port mirroring provide monitoring and troubleshooting capabilities to the virtual infrastructure.

Consider the following caveats for distributed switches.

- Distributed switches are not manageable when vCenter Server is unavailable. vCenter Server therefore becomes a tier one application.

Health Check

The health check service helps identify and troubleshoot configuration errors in vSphere distributed switches.

Health check helps identify the following common configuration errors.

- Mismatched VLAN trunks between an ESXi host and the physical switches it's connected to.
- Mismatched MTU settings between physical network adapters, distributed switches, and physical switch ports.
- Mismatched virtual switch teaming policies for the physical switch port-channel settings.

Health check monitors VLAN, MTU, and teaming policies.

VLANs	Checks whether the VLAN settings on the distributed switch match the trunk port configuration on the connected physical switch ports.
MTU	For each VLAN, health check determines whether the physical access switch port's MTU jumbo frame setting matches the distributed switch MTU setting.
Teaming policies	Health check determines whether the connected access ports of the physical switch that participate in an EtherChannel are paired with distributed ports whose teaming policy is IP hash.

Health check is limited to the access switch port to which the ESXi hosts' NICs connects.

Design ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Net-001	Enable vSphere Distributed Switch Health Check on all virtual distributed switches.	vSphere Distirubted Switch Health Check ensures all VLANS are trunked to all hosts attached to the vSphere Distributed Switch and ensures MTU sizes match the physical network.	You must have a minimum of two physical uplinks to use this feature.

Note For VLAN and MTU checks, at least two physical NICs for the distributed switch are required. For a teaming policy check, at least two physical NICs and two hosts are required when applying the policy.

Number of Virtual Switches

Create fewer virtual switches, preferably just one. For each type of network traffic, configure a single portgroup to simplify configuration and monitoring.

Table 7-23. Virtual Switch Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Net-002	Use vSphere Distributed Switches (VDS).	vSphere Distributed Switches simplify management.	Migration from a VSS to a VDS requires a minimum of two physical NICs to maintain redundancy.
SDDC-VI-Net-003	Use a single VDS per cluster.	Reduces complexity of the network design. Reduces the size of the fault domain.	Increases the number of vSphere Distributed Switches that must be managed.

Management Cluster Distributed Switches

The management cluster uses a single vSphere Distributed Switch with the following configuration settings.

Table 7-24. Virtual Switch for the Management Cluster

vSphere Distributed Switch Name	Function	Network I/O Control	Number of Physical NIC Ports	MTU
vDS-Mgmt	■ ESXi Management ■ Network IP Storage (NFS) ■ Virtual SAN ■ vSphere vMotion ■ VXLAN Tunnel Endpoint (VTEP) ■ Uplinks (2) to enable ECMP ■ External management connectivity	Enabled	2	9000

Table 7-25. vDS-MgmtPort Group Configuration Settings

Parameter	Setting
Failover detection	Link status only
Notify switches	Enabled

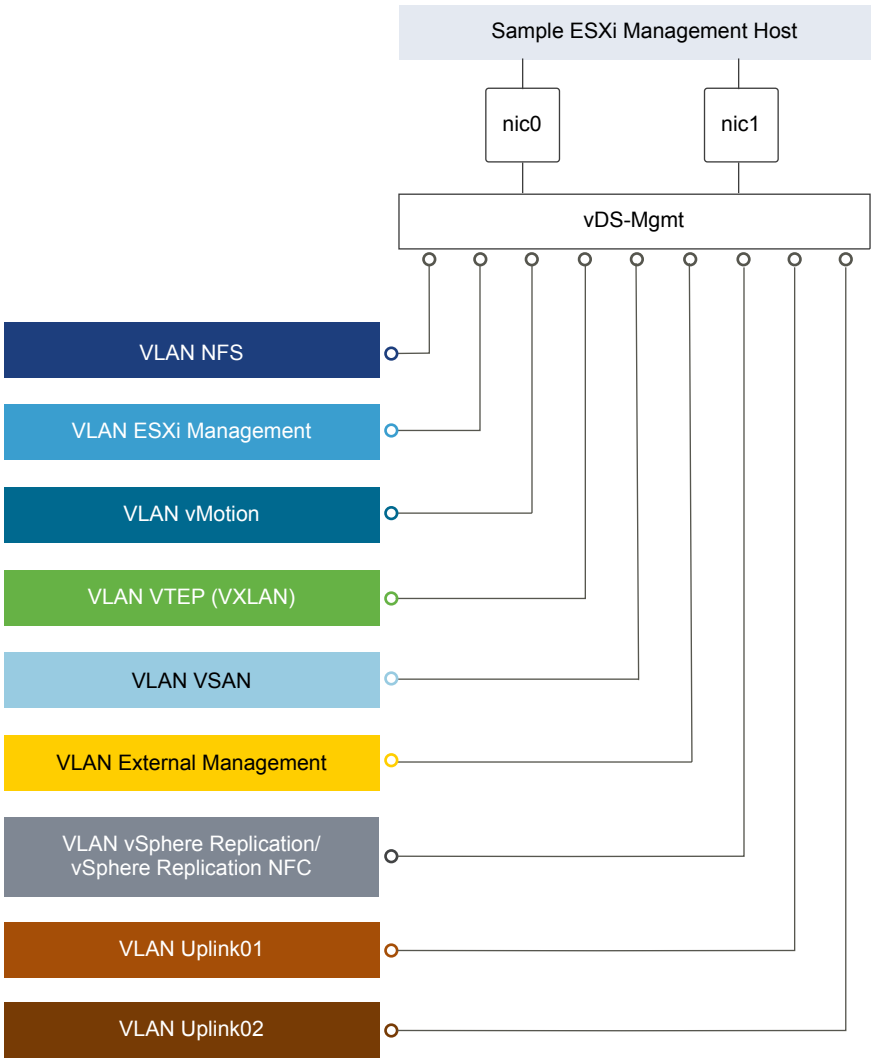
Table 7-25. vDS-MgmtPort Group Configuration Settings (Continued)

Parameter	Setting
Failback	No
Failover order	Active uplinks: Uplink1, Uplink2

The following figure illustrates the network switch design.

Note The illustration includes the vSphere Replication VLAN. That VLAN is not needed for a single-region implementation.

Figure 7-5. Network Switch Design for Management Hosts



This section expands on the logical network design by providing details on the physical NIC layout and physical network attributes.

Table 7-26. Management Virtual Switches by Physical/Virtual NIC

vSphere Distributed Switch	vmnic	Function
vDS-Mgmt	0	Uplink
vDS-Mgmt	1	Uplink

Note The following VLANs are meant as samples. Your actual implementation depends on your environment.

Table 7-27. Management Virtual Switch Port Groups and VLANs

vSphere Distributed Switch	Port Group Name	Teaming Policy	Active Uplinks	VLAN ID
vDS-Mgmt	vDS-Mgmt-Management	Route based on physical NIC load	0, 1	1611
vDS-Mgmt	vDS-Mgmt-vMotion	Route based on physical NIC load	0, 1	1612
vDS-Mgmt	vDS-Mgmt-VSAN	Route based on physical NIC load	0, 1	1613
vDS-Mgmt	Auto Generated (NSX VTEP)	Route based on SRC-ID	0, 1	1614
vDS-Mgmt	vDS-Mgmt-Uplink01	Route based on physical NIC load	0, 1	2711
vDS-Mgmt	vDS-Mgmt-Uplink02	Route based on physical NIC load	0, 1	2712
vDS-Mgmt	vDS-Mgmt-NFS	Route based on physical NIC load	0, 1	1615
vDS-Mgmt	vDS-Mgmt-Ext-Management	Route based on physical NIC load	0, 1	130

Table 7-28. Management VMkernel Adapter

vSphere Distributed Switch	Network Label	Connected Port Group	Enabled Services	MTU
vDS-Mgmt	Management	vDS-Mgmt-Management	Management Traffic	1500 (Default)
vDS-Mgmt	vMotion	vDS-Mgmt-vMotion	vMotion Traffic	9000
vDS-Mgmt	VSAN	vDS-Mgmt-VSAN	VSAN	9000
vDS-Mgmt	NFS	vDS-Mgmt-NFS	-	9000
vDS-Mgmt	VTEP	Auto Generated (NSX VTEP)	-	9000

For more information on the physical network design specifications, see Physical Networking Design.

Shared Edge and Compute Cluster Distributed Switches

The shared edge and compute cluster uses a single vSphere Distributed Switch with the following configuration settings.

Table 7-29. Virtual Switch for the Shared Edge and Compute Cluster

vSphere Distributed Switch Name	Function	Network I/O Control	Number of Physical NIC Ports	MTU
vDS-Comp01	■ ESXi Management ■ Network IP Storage (NFS) ■ vSphere vMotion ■ VXLAN Tunnel Endpoint (VTEP) ■ Uplinks (2) to enable ECMP ■ vSAN ■ External customer/tenant connectivity	Enabled	2	9000

Table 7-30. vDS-Comp01 Port Group Configuration Settings

Parameter	Setting
Failoverdetection	Link status only
Notify switches	Enabled
Failback	Yes
Failover order	Active uplinks: Uplink1, Uplink2

Network Switch Design for Shared Edge and Compute Hosts

This section expands on the logical network design by providing details on the physical NIC layout and physical network attributes.

Figure 7-6. Network Switch Design for Shared Edge and Compute Hosts

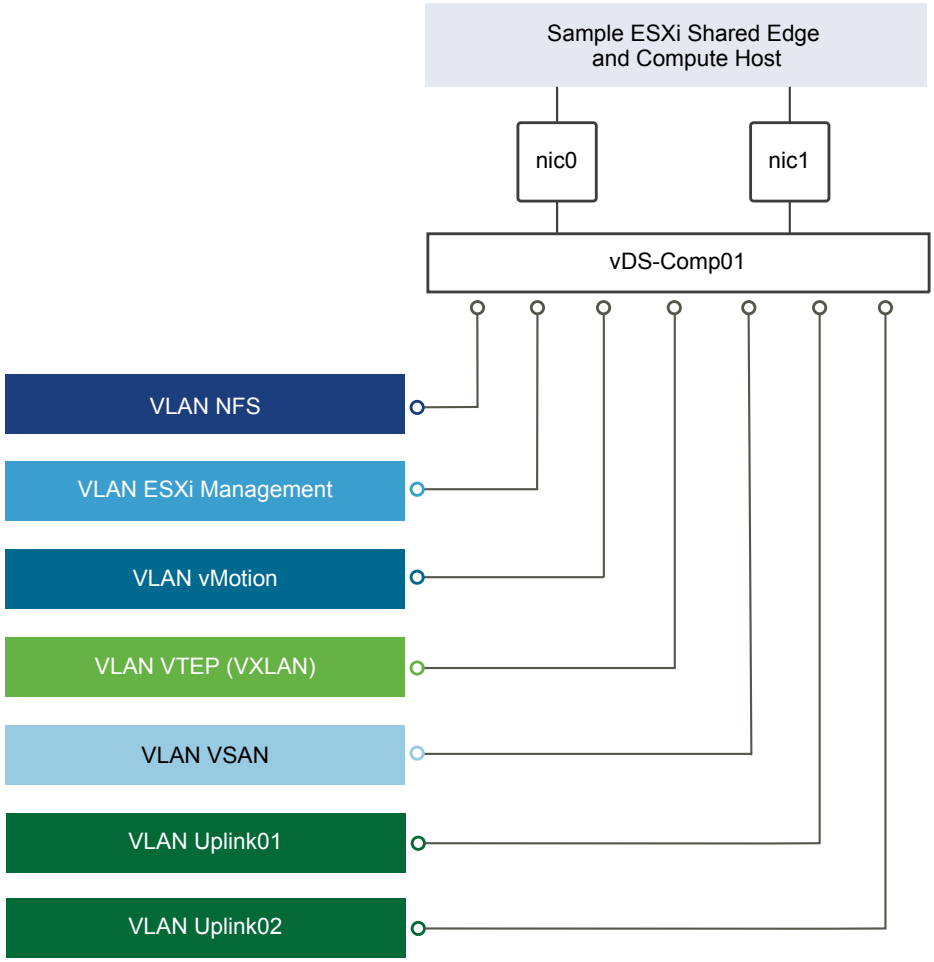


Table 7-31. Shared Edge and Compute Cluster Virtual Switches by Physical/Virtual NIC

vSphere Distributed Switch	vmnic	Function
vDS-Comp01	0	Uplink
vDS-Comp01	1	Uplink

Note The following VLANs are meant as samples. Your actual implementation depends on your environment.

Table 7-32. Shared Edge and Compute Cluster Virtual Switch Port Groups and VLANs

vSphere Distributed Switch	Port Group Name	Teaming Policy	Active Uplinks	VLAN ID
vDS-Comp01	vDS-Comp01-Management	Route based on physical NIC load	0, 1	1631
vDS-Comp01	vDS-Comp01-vMotion	Route based on physical NIC load	0, 1	1632
vDS-Comp01	vDS-Comp01-VSAN	Route based on physical NIC load	0, 1	1633
vDS-Comp01	vDS-Comp01-NFS	Route based on physical NIC load	0, 1	1615
vDS-Comp01	Auto Generated (NSX VTEP)	Route based on SRC-ID	0, 1	1634

Table 7-32. Shared Edge and Compute Cluster Virtual Switch Port Groups and VLANs (Continued)

vSphere Distributed Switch	Port Group Name	Teaming Policy	Active Uplinks	VLAN ID
vDS-Comp01	vDS-Comp01-Uplink01	Route based on physical NIC load	0, 1	1635
vDS-Comp01	vDS-Comp01-Uplink02	Route based on physical NIC load	0, 1	2713

Table 7-33. Shared Edge and Compute Cluster VMkernel Adapter

vSphere Distributed Switch	Network Label	Connected Port Group	Enabled Services	MTU
vDS-Comp01	Management	vDS-Comp01-Management	Management Traffic	1500 (Default)
vDS-Comp01	vMotion	vDS-Comp01-vMotion	vMotion Traffic	9000
vDS-Comp01	VSAN	vDS-Comp01-VSAN	VSAN	9000
vDS-Comp01	NFS	vDS-Comp01-NFS	-	9000
vDS-Comp01	VTEP	Auto Generated (NSX VTEP)	-	9000

For more information on the physical network design, see *Physical Networking Design*.

Compute Cluster Distributed Switches

A compute cluster vSphere Distributed Switch uses the following configuration settings.

Table 7-34. Virtual Switch for a dedicated Compute Cluster

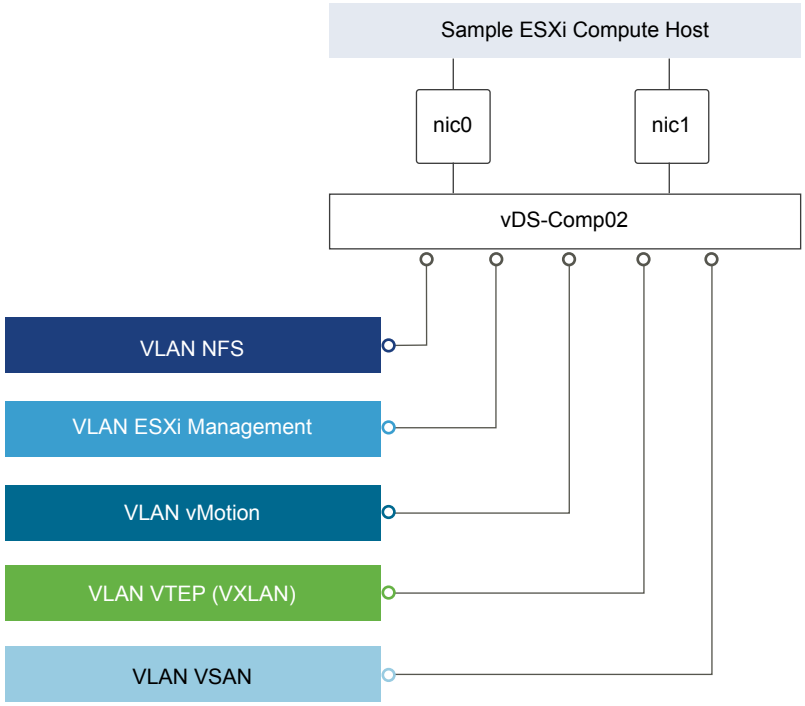
vSphere Distributed Switch Name	Function	Network I/O Control	Number of Physical NIC Ports	MTU
vDS-Comp02	■ ESXi Management ■ Network IP Storage (NFS) ■ vSphere vMotion ■ VXLAN Tunnel Endpoint (VTEP)	Enabled	2	9000

Table 7-35. vDS-Comp02 Port Group Configuration Settings

Parameter	Setting
Failover detection	Link status only
Notify switches	Enabled
Failback	Yes
Failover order	Active uplinks: Uplink1, Uplink2

Network Switch Design for Compute Hosts

Figure 7-7. Network Switch Design for Compute Hosts



This section expands on the logical network design by providing details on the physical NIC layout and physical network attributes.

Table 7-36. Compute Cluster Virtual Switches by Physical/Virtual NIC

vSphere Distributed Switch	vmnic	Function
vDS-Comp02	0	Uplink
vDS-Comp02	1	Uplink

Note The following VLANs are meant as samples. Your actual implementation depends on your environment.

Table 7-37. Compute Cluster Virtual Switch Port Groups and VLANs

vSphere Distributed Switch	Port Group Name	Teaming Policy	Active Uplinks	VLAN ID
vDS-Comp02	vDS-Comp02-Management	Route based on physical NIC load	0, 1	1621
vDS-Comp02	vDS-Comp02-vMotion	Route based on physical NIC load	0, 1	1622
vDS-Comp02	Auto Generated (NSX VTEP)	Route based on SRC-ID	0, 1	1624
vDS-Comp02	vDS-Comp02-NFS	Route based on physical NIC load	0, 1	1625

Table 7-38. Compute Cluster VMkernel Adapter

vSphere Distributed Switch	Network Label	Connected Port Group	Enabled Services	MTU
vDS-Comp02	Management	vDS-Comp02-Management	Management traffic	1500 (Default)
vDS-Comp02	vMotion	vDS-Comp02-vMotion	vMotion traffic	9000
vDS-Comp02	NFS	vDS-Comp02-NFS	-	9000
vDS-Comp02	VTEP	Auto Generated (NSX VTEP)	-	9000

For more information on the physical network design specifications, see *Physical Networking Design*.

NIC Teaming

You can use NIC teaming to increase the network bandwidth available in a network path, and to provide the redundancy that supports higher availability.

NIC teaming helps avoid a single point of failure and provides options for load balancing of traffic. To further reduce the risk of a single point of failure, build NIC teams by using ports from multiple NIC and motherboard interfaces.

Create a single virtual switch with teamed NICs across separate physical switches.

This VMware Validated Design uses an active-active configuration using the route that is based on physical NIC load algorithm for teaming. In this configuration, idle network cards do not wait for a failure to occur, and they aggregate bandwidth.

Benefits and Overview

NIC teaming helps avoid a single point of failure and provides options for load balancing of traffic. To further reduce the risk of a single point of failure, build NIC teams by using ports from multiple NIC and motherboard interfaces.

Create a single virtual switch with teamed NICs across separate physical switches.

This VMware Validated Design uses an active-active configuration using the route that is based on physical NIC load algorithm for teaming. In this configuration, idle network cards do not wait for a failure to occur, and they aggregate bandwidth.

NIC Teaming Design Background

For a predictable level of performance, use multiple network adapters in one of the following configurations.

- An active-passive configuration that uses explicit failover when connected to two separate switches.
- An active-active configuration in which two or more physical NICs in the server are assigned the active role.

This validated design uses an active-active configuration.

Table 7-39. NIC Teaming and Policy

Design Quality	Active-Active	Active-Passive	Comments
Availability	↑	↑	Using teaming regardless of the option increases the availability of the environment.
Manageability	o	o	Neither design option impacts manageability.
Performance	↑	o	An active-active configuration can send traffic across either NIC, thereby increasing the available bandwidth. This configuration provides a benefit if the NICs are being shared among traffic types and Network I/O Control is used.
Recoverability	o	o	Neither design option impacts recoverability.
Security	o	o	Neither design option impacts security.

Legend: ↑ = positive impact on quality; ↓ = negative impact on quality; o = no impact on quality.

Table 7-40. NIC Teaming Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Net-004	Use the Route based on physical NIC load teaming algorithm for all port groups except for ones that carry VXLAN traffic. VTEP kernel ports and VXLAN traffic will use Route based on SRC-ID.	Reduce complexity of the network design and increase resiliency and performance.	Because NSX does not support Route based on physical NIC load two different algorithms are necessary.

Network I/O Control

When Network I/O Control is enabled, the distributed switch allocates bandwidth for the following system traffic types.

- Fault tolerance traffic
- iSCSI traffic
- vSphere vMotion traffic
- Management traffic
- NFS traffic
- VMware Virtual SAN traffic
- Virtual machine traffic

How Network I/O Control Works

Network I/O Control enforces the share value specified for the different traffic types only when there is network contention. When contention occurs Network I/O Control applies the share values set to each traffic type. As a result, less important traffic, as defined by the share percentage, will be throttled, allowing more important traffic types to gain access to more network resources.

Network I/O Control also allows the reservation of bandwidth for system traffic based on the capacity of the physical adapters on a host, and enables fine-grained resource control at the virtual machine network adapter level. Resource control is similar to the model for vCenter CPU and memory reservations.

Network I/O Control Heuristics

The following heuristics can help with design decisions.

Shares vs. Limits — When you use bandwidth allocation, consider using shares instead of limits. Limits impose hard limits on the amount of bandwidth used by a traffic flow even when network bandwidth is available.

Limits on Certain Resource Pools — Consider imposing limits on a given resource pool. For example, if you put a limit on vSphere vMotion traffic, you can benefit in situations where multiple vSphere vMotion data transfers, initiated on different hosts at the same time, result in oversubscription at the physical network level. By limiting the available bandwidth for vSphere vMotion at the ESXi host level, you can prevent performance degradation for other traffic.

Teaming Policy — When you use Network I/O Control, use Route based on physical NIC load teaming as a distributed switch teaming policy to maximize the networking capacity utilization. With load-based teaming, traffic might move among uplinks, and reordering of packets at the receiver can result occasionally.

Traffic Shaping — Use distributed port groups to apply configuration policies to different traffic types. Traffic shaping can help in situations where multiple vSphere vMotion migrations initiated on different hosts converge on the same destination host. The actual limit and reservation also depend on the traffic shaping policy for the distributed port group where the adapter is connected to.

Network I/O Control Design Decisions

Based on the heuristics, this design has the following decisions.

Table 7-41. Network I/O Control Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-NET-005	Enable Network I/O Control on all distributed switches.	Increase resiliency and performance of the network.	If configured incorrectly Network I/O Control could impact network performance for critical traffic types.
SDDC-VI-NET-006	Set the share value for vMotion traffic to Low.	During times of contention vMotion traffic is not as important as virtual machine or storage traffic.	During times of network contention vMotion migrations take longer then usual to complete.

Table 7-41. Network I/O Control Design Decisions (Continued)

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-NET-008	Set the share value for vSAN to High.	During times of contention vSAN traffic needs guaranteed bandwidth so virtual machine performance does not suffer.	None.
SDDC-VI-NET-009	Set the share value for Management to Normal.	By keeping the default setting of Normal management traffic is prioritized higher than vMotion but lower then vSAN traffic. Management traffic is important as it ensures the hosts can still be managed during times of network contention.	None.
SDDC-VI-NET-010	Set the share value for NFS Traffic to Low.	Because NFS is used for secondary storage, such as VDP backups and vRealize Log Insight archives it is not as important as vSAN traffic, by prioritizing it lower vSAN is not impacted.	None
SDDC-VI-NET-012	Set the share value for virtual machines to High.	Virtual machines are the most important asset in the SDDC. Leaving the default setting of High ensures that they will always have access to the network resources they need.	None.
SDDC-VI-NET-013	Set the share value for Fault Tolerance to Low.	Fault Tolerance is not used in this design therefore it can be set to the lowest priority.	None.
SDDC-VI-NET-014	Set the share value for iSCSI traffic to Low.	iSCSI is not used in this design therefore it can be set to the lowest priority.	None.

VXLAN

VXLAN provides the capability to create isolated, multi-tenant broadcast domains across data center fabrics and enables customers to create elastic, logical networks that span physical network boundaries.

The first step in creating these logical networks is to abstract and pool the networking resources. Just as vSphere abstracts compute capacity from the server hardware to create virtual pools of resources that can be consumed as a service, vSphere Distributed Switch and VXLAN abstract the network into a generalized pool of network capacity and separate the consumption of these services from the underlying physical infrastructure. A network capacity pool can span physical boundaries, optimizing compute resource utilization across clusters, pods, and geographically-separated data centers. The unified pool of network capacity can then be optimally segmented into logical networks that are directly attached to specific applications.

VXLAN works by creating Layer 2 logical networks that are encapsulated in standard Layer 3 IP packets. A Segment ID in every frame differentiates the VXLAN logical networks from each other without any need for VLAN tags. As a result, large numbers of isolated Layer 2 VXLAN networks can coexist on a common Layer 3 infrastructure.

In the vSphere architecture, the encapsulation is performed between the virtual NIC of the guest VM and the logical port on the virtual switch, making VXLAN transparent to both the guest virtual machines and the underlying Layer 3 network. Gateway services between VXLAN and non-VXLAN hosts (for example, a physical server or the Internet router) are performed by the NSX for vSphereEdge gateway appliance. The Edge gateway translates VXLAN segment IDs to VLAN IDs, so that non-VXLAN hosts can communicate with virtual machines on a VXLAN network.

The dedicated edge cluster hosts all NSX Edge instances and all Universal Distributed Logical Router instances that are connect to the Internet or to corporate VLANs, so that the network administrator can manage the environment in a more secure and centralized way.

Table 7-42. VXLAN Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Net-015	Use NSX for vSphere to introduce VXLANs for the use of virtual application networks and tenants networks.	Simplify the network configuration for each tenant via centralized virtual network management.	Requires additonal compute and storage resources to deploy NSX components. Additional training may be needed on NSX for vSphere.
SDDC-VI-Net-016	Use VXLAN along with NSX Edge gateways, the Universal Distributed Logical Router (UDLR) and Distributed Logical Router (DLR) to provide customer/tenant network capabilities.	Create isolated, multi-tenant broadcast domains across data center fabrics to create elastic, logical networks that span physical network boundaries.	Transport networks and MTU greater than 1600 bytes has to be configured in the reachability radius.
SDDC-VI-Net-017	Use VXLAN along with NSX Edge gateways and the Universal Distributed Logical Router (UDLR) to provide management application network capabilities.	Leverage benefits of network virtualization in the management pod.	Requires installation and configuration of the NSX for vSphere instance in the management pod.

NSX Design

This design implements software-defined networking by using VMware NSX™ for vSphere®.With NSX for vSphere, virtualization delivers for networking what it has already delivered for compute and storage.

In much the same way that server virtualization programmatically creates, snapshots, deletes, and restores software-based virtual machines (VMs), NSX network virtualization programmatically creates, snapshots, deletes, and restores software-based virtual networks. The result is a transformative approach to networking that not only enables data center managers to achieve orders of magnitude better agility and economics, but also supports a vastly simplified operational model for the underlying physical network. NSX for vSphere is a nondisruptive solution because it can be deployed on any IP network, including existing traditional networking models and next-generation fabric architectures, from any vendor.

When administrators provision workloads, network management is one of the most time-consuming tasks. Most of the time spent provisioning networks is consumed configuring individual components in the physical infrastructure and verifying that network changes do not affect other devices that are using the same networking infrastructure.

The need to pre-provision and configure networks is a major constraint to cloud deployments where speed, agility, and flexibility are critical requirements. Pre-provisioned physical networks can allow for the rapid creation of virtual networks and faster deployment times of workloads utilizing the virtual network. As long as the physical network that you need is already available on the host where the workload is to be deployed, this works well. However, if the network is not available on a given host, you must find a host with the available network and spare capacity to run your workload in your environment.

To get around this bottleneck requires a decoupling of virtual networks from their physical counterparts. This, in turn, requires that you can programmatically recreate all physical networking attributes that are required by workloads in the virtualized environment. Because network virtualization supports the creation of virtual networks without modification of the physical network infrastructure, it allows more rapid network provisioning.

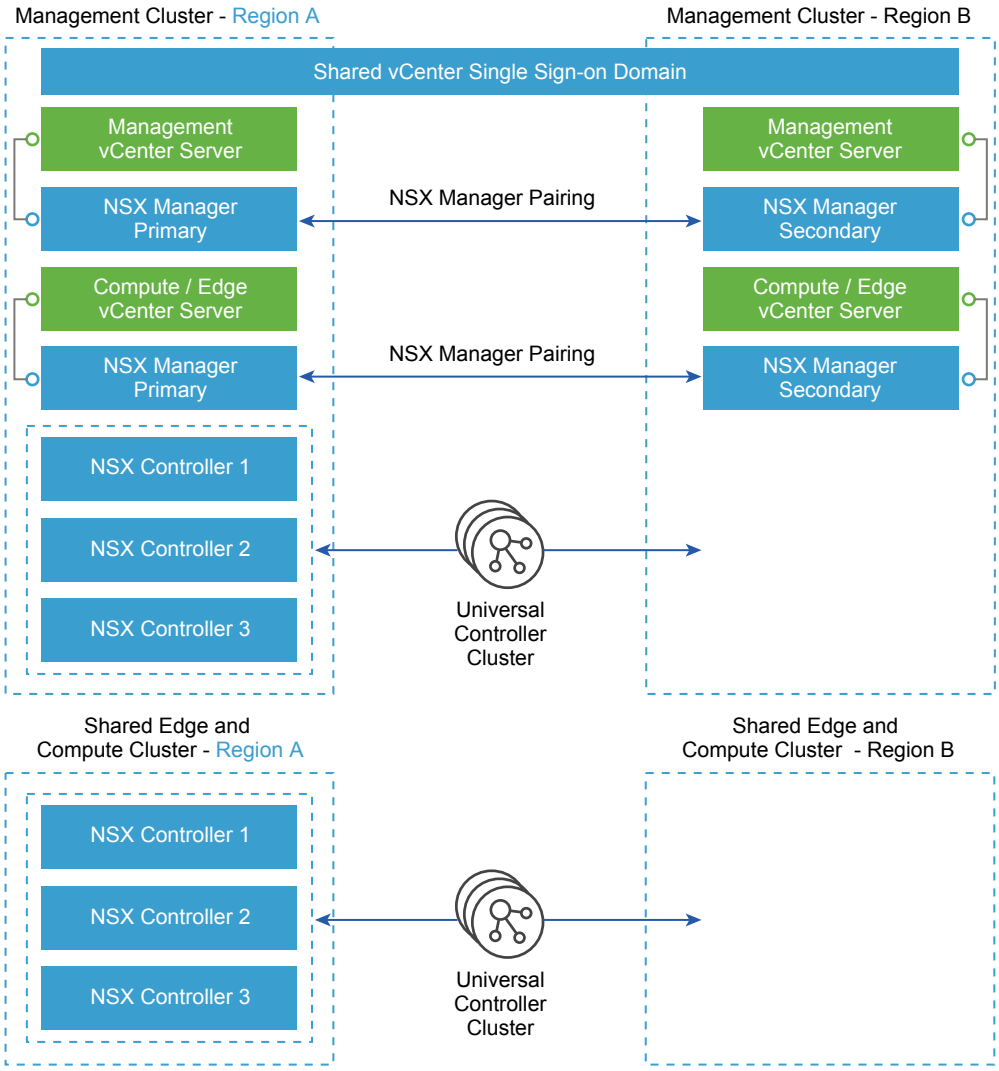
NSX for vSphere Design

Each NSX instance is tied to a vCenter Server instance. The design decision to deploy two vCenter Server instances per region(SDDC-VI-VC-001) requires deployment of two separate NSX instances per region.

Table 7-43. NSX for vSphere Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-001	Use two separate NSX instances per region. One instance is tied to the Management vCenter Server, and the other instance is tied to the Compute vCenter Server.	Software-defined Networking (SDN) capabilities offered by NSX, such as load balancing and firewalls, are crucial for the compute/edge layer to support the cloud management platform operations, and also for the management applications in the management stack that need these capabilities.	You must install and perform initial configuration of multiple NSX instances separately.
SDDC-VI-SDN-002	Pair NSX Manager instances in a primary-secondary relationship across regions for both management and compute workloads.	NSX can extend the logical boundaries of the networking and security services across regions. As a result, workloads can be live-migrated and failed over between regions without reconfiguring the network and security constructs.	You must consider that you can pair up to eight NSX Manager instances.

Figure 7-8. Architecture of NSX for vSphere



NSX Components

The following sections describe the components in the solution and how they are relevant to the network virtualization design.

Consumption Layer

If the design is expanded, a cloud management platform such as vRealize Automation can consume NSX for vSphere by using the NSX REST API, the vSphere Web Client, or both.

API

NSX for vSphere offers a powerful management interface through its REST API.

- A client can read an object by making an HTTP GET request to the object's resource URL.

- A client can write (create or modify) an object with an HTTP PUT or POST request that includes a new or changed XML document for the object.
- A client can delete an object with an HTTP DELETE request.

Note Some of the use case designs do not include a cloud platform.

vSphere Web Client

The NSX Manager component provides a networking and security plug-in in the vSphere Web Client. This plug-in provides an interface to consuming virtualized networking from the NSX Manager for users that have sufficient privileges.

Table 7-44. Consumption Method Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-003	For the shared edge and compute cluster NSX instance, administrators can use both the vSphere Web Client and the NSX REST API.	The vSphere Web Client consumes NSX for vSphere resources through the Network and Security plug-in. The NSX REST API offers the potential of scripting repeating actions and operations.	None.
SDDC-VI-SDN-004	For the management cluster NSX instance, consumption is only by provider staff via the vSphere Web Client and the API.	Ensures that infrastructure components are not modified by tenants and/or non-provider staff.	Tenants do not have access to the management stack workloads.

NSX Manager

NSX Manager provides the centralized management plane for NSX for vSphere and has a one-to-one mapping to vCenter Server workloads.

NSX Manager performs the following functions.

- Provides the single point of configuration and the REST API entry-points for NSX in a vSphere environment.
- Deploys NSX Controller clusters, Edge distributed routers, and Edge service gateways in the form of OVF appliances, guest introspection services, and so on.
- Prepares ESXi hosts for NSX by installing VXLAN, distributed routing and firewall kernel modules, and the User World Agent (UWA).
- Communicates with NSX Controller clusters over REST and with hosts over the RabbitMQ message bus. This internal message bus is specific to NSX for vSphere and does not require setup of additional services.
- Generates certificates for the NSX Controller instances and ESXi hosts to secure control plane communications with mutual authentication.

NSX Controller

An NSX Controller performs the following functions.

- Provides the control plane to distribute VXLAN and logical routing information to ESXi hosts.
- Includes nodes that are clustered for scale-out and high availability.
- Slices network information across cluster nodes for redundancy.
- Removes requirement of VXLAN Layer 3 multicast in the physical network.
- Provides ARP suppression of broadcast traffic in VXLAN networks.

NSX control plane communication occurs over the management network.

Table 7-45. NSX Controller Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-005	Deploy NSX Controller instances in Universal Cluster mode with three members to provide high availability and scale. Provision these three nodes through the primary NSX Manager instance.	The high availability of NSX Controller reduces the downtime period in case of failure of one physical host.	The secondary NSX Manager will not have active controllers but will automatically import the configuration of the Universal Controllers that are created in the primary NSX Manager

NSX Virtual Switch

The NSX data plane consists of the NSX virtual switch. This virtual switch is based on the vSphere Distributed Switch (VDS)with additional components to enable rich services. The add-on NSX components include kernel modules (VIBs) which run within the hypervisor kernel and provide services such as distributed logical router (DLR) and distributed firewall (DFW), and VXLAN capabilities.

The NSX virtual switch abstracts the physical network and provides access-level switching in the hypervisor. It is central to network virtualization because it enables logical networks that are independent of physical constructs such as VLAN. Using an NSX virtual switch includes several benefits.

- Supports overlay networking and centralized network configuration. Overlay networking enables the following capabilities.
- Facilitates massive scale of hypervisors.
- Because the NSX virtual switch is based on VDS, it provides a comprehensive toolkit for traffic management, monitoring, and troubleshooting within a virtual network through features such as port mirroring, NetFlow/IPFIX, configuration backup and restore, network health check, QoS, and more.

Logical Switching

NSX logical switches create logically abstracted segments to which tenant virtual machines can be connected. A single logical switch is mapped to a unique VXLAN segment and is distributed across the ESXi hypervisors within a transport zone. The logical switch allows line-rate switching in the hypervisor without the constraints of VLAN sprawl or spanning tree issues.

Distributed Logical Router

The NSX distributed logical router (DLR) is optimized for forwarding in the virtualized space, that is, forwarding between VMs on VXLAN- or VLAN-backed port groups. DLR has the following characteristics.

- High performance, low overhead first hop routing
- Scales with number of hosts
- Up to 1,000 Logical Interfaces (LIFs) on each DLR

Distributed LogicalRouter Control Virtual Machine

The distributed logical router control virtual machine is the control plane component of the routing process, providing communication between NSX Manager and the NSX Controller cluster through the User World Agent (UWA). NSX Manager sends logical interface information to the control virtual machine and the NSX Controller cluster, and the control virtual machine sends routing updates to the NSX Controller cluster.

User World Agent

The User World Agent (UWA) is a TCP (SSL) client that facilitates communication between the ESXi hosts and the NSX Controller instances as well as the retrieval of information from the NSX Manager via interaction with the message bus agent.

VXLAN Tunnel Endpoint

VXLAN Tunnel Endpoints (VTEPs) are instantiated within the vSphere Distributed Switch to which the ESXi hosts that are prepared for NSX for vSphere are connected. VTEPs are responsible for encapsulating VXLAN traffic as frames in UDP packets and for the corresponding decapsulation. VTEPs take the form of one or more VMkernel ports with IP addresses and are used both to exchange packets with other VTEPs and to join IP multicast groups via Internet Group Membership Protocol (IGMP). If you use multiple VTEPs, then you must select a teaming method.

Edge Services Gateway

The NSX Edge services gateways (ESGs) primary function is north/south communication, but it also offers support for Layer 2, Layer 3, perimeter firewall, load balancing and other services such as SSL-VPN and DHCP-relay.

Distributed Firewall

NSX includes a distributed kernel-level firewall known as the distributed firewall. Security enforcement is done at the kernel and VM network adapter level. The security enforcement implementation enables firewall rule enforcement in a highly scalable manner without creating bottlenecks on physical appliances. The distributed firewall has minimal CPU overhead and can perform at line rate.

The flow monitoring feature of the distributed firewall displays network activity between virtual machines at the application protocol level. This information can be used to audit network traffic, define and refine firewall policies, and identify botnets.

Logical Load Balancer

The NSX logical load balancer provides load balancing services up to Layer 7, allowing distribution of traffic across multiple servers to achieve optimal resource utilization and availability. The logical load balancer is a service provided by the NSX Edge service gateway.

NSX for vSphere Requirements

NSX for vSphere requirements impact both physical and virtual networks.

Physical Network Requirements

Physical requirements determine the MTU size for networks that carry VLAN traffic, dynamic routing support, type synchronization through an NTP server, and forward and reverse DNS resolution.

Requirement	Comments
Any network that carries VXLAN traffic must have an MTU size of 1600 or greater.	VXLAN packets cannot be fragmented. The MTU size must be large enough to support extra encapsulation overhead. This design uses jumbo frames, MTU size of 9000, for VXLAN traffic.
For the hybrid replication mode, Internet Group Management Protocol (IGMP) snooping must be enabled on the Layer 2 switches to which ESXi hosts that participate in VXLAN are attached. IGMP querier must be enabled on the connected router or Layer 3 switch.	IGMP snooping on Layer 2 switches is a requirement of the hybrid replication mode. Hybrid replication mode is the recommended replication mode for broadcast, unknown unicast, and multicast (BUM) traffic when deploying into an environment with large scale-out potential. The traditional requirement for Protocol Independent Multicast (PIM) is removed.
Dynamic routing support on the upstream Layer 3 data center switches must be enabled.	Enable a dynamic routing protocol supported by NSX on the upstream data center switches to establish dynamic routing adjacency with the ESGs.
NTP server must be available.	The NSX Manager requires NTP settings that synchronize it with the rest of the vSphere environment. Drift can cause problems with authentication. The NSX Manager must be in sync with the vCenter Single Sign-On service on the Platform Services Controller.
Forward and reverse DNS resolution for all management VMs must be established.	The NSX Controller nodes do not require DNS entries.

NSX Component Specifications

The following table lists the components involved in the NSX for vSphere solution and the requirements for installing and running them. The compute and storage requirements have been taken into account when sizing resources to support the NSX for vSphere solution.

Note NSX ESG sizing can vary with tenant requirements, so all options are listed.

VM	vCPU	Memory	Storage	Quantity per Stack Instance
NSX Manager	4	16 GB	60 GB	1
NSX Controller	4	4 GB	20 GB	3

VM	vCPU	Memory	Storage	Quantity per Stack Instance
NSX ESG	1 (Compact) 2 (Large) 4 (Quad Large) 6 (X-Large)	512 MB (Compact) 1 GB (Large) 1 GB (Quad Large) 8 GB (X-Large)	512 MB 512 MB 512 MB 4.5 GB (X-Large) (+4 GB with swap)	Optional component. Deployment of the NSX ESG varies per use case.
DLR control VM	1	512 MB	512 MB	Optional component. Varies with use case. Typically 2 per HA pair.
Guest introspection	2	1 GB	4 GB	Optional component. 1 per ESXi host.
NSX data security	1	512 MB	6 GB	Optional component.1 per ESXi host.

NSX Edge Service Gateway Sizing

The Quad Large model is suitable for high performance firewall abilities and the X-Large is suitable for both high performance load balancing and routing.

You can convert between NSX Edge service gateway sizes upon demand using a non-disruptive upgrade process, so the recommendation is to begin with the Large model and scale up if necessary. A Large NSX Edge service gateway is suitable for medium firewall performance but as detailed later, the NSX Edge service gateway does not perform the majority of firewall functions.

Note Edge service gateway throughput is influenced by the WAN circuit. An adaptable approach, that is, converting as necessary, is recommended.

Table 7-46. NSX Edge Service Gateway Sizing Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-006	Use large size NSX Edge service gateways.	The large size provides all the performance characteristics needed even in the event of a failure. A larger size would also provide the performance required but at the expense of extra resources that wouldn't be used.	None.

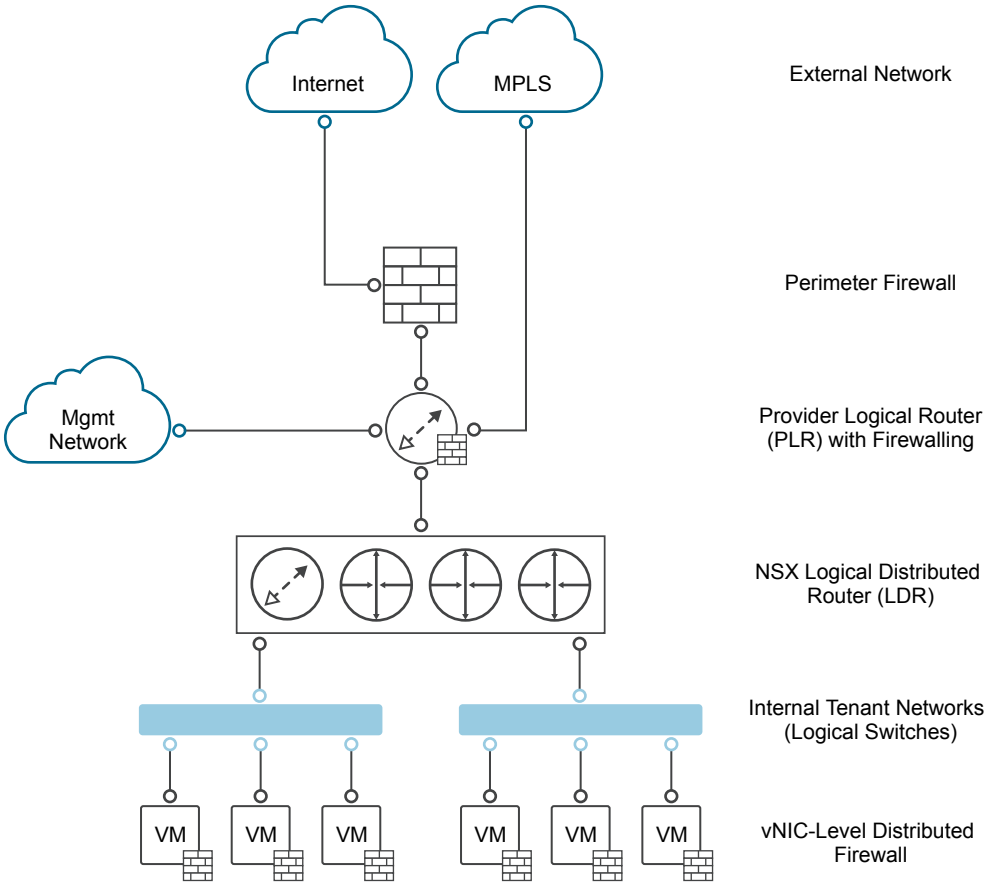
Network Virtualization Conceptual Design

This conceptual design provides you with an understanding of the network virtualization design.

The network virtualization conceptual design includes a perimeter firewall, a provider logical router, and the NSX for vSphere Logical Router. It also includes the external network, internal tenant network, and internal non-tenant network.

Note In this document, tenant refers to a tenant of the cloud management platform within the compute/edge stack, or to a management application within the management stack.

Figure 7-9. Conceptual Tenant Overview



The conceptual design has the following key components.

External Networks	Connectivity to and from external networks is through the perimeter firewall. The main external network is the Internet.
Perimeter Firewall	The physical firewall exists at the perimeter of the data center. Each tenant receives either a full instance or partition of an instance to filter external traffic.
Provider Logical Router (PLR)	The PLR exists behind the perimeter firewall and handles north/south traffic that is entering and leaving tenant workloads.
NSX for vSphereDistributed Logical Router (DLR)	This logical router is optimized for forwarding in the virtualized space, that is, between VMs, on VXLAN port groups or VLAN-backed port groups.

Internal Non-Tenant Network	A single management network, which sits behind the perimeter firewall but not behind the PLR. Enables customers to manage the tenant environments.
Internal Tenant Networks	Connectivity for the main tenant workload. These networks are connected to a DLR, which sits behind the PLR. These networks take the form of VXLAN-based NSX for vSphere logical switches. Tenant virtual machine workloads will be directly attached to these networks.

Cluster Design for NSX for vSphere

Following the vSphere design, the NSX for vSphere design consists of a management stack and a compute/edge stack in each region.

Management Stack

In the management stack, the underlying hosts are prepared for NSX for vSphere. The management stack has these components.

- NSX Manager instances for both stacks (management stack and compute/edge stack)
- NSX Controller cluster for the management stack
- NSX ESG and DLR control VMs for the management stack

Compute/Edge Stack

In the compute/edge stack, the underlying hosts are prepared for NSX for vSphere. The compute/edge stack has these components.

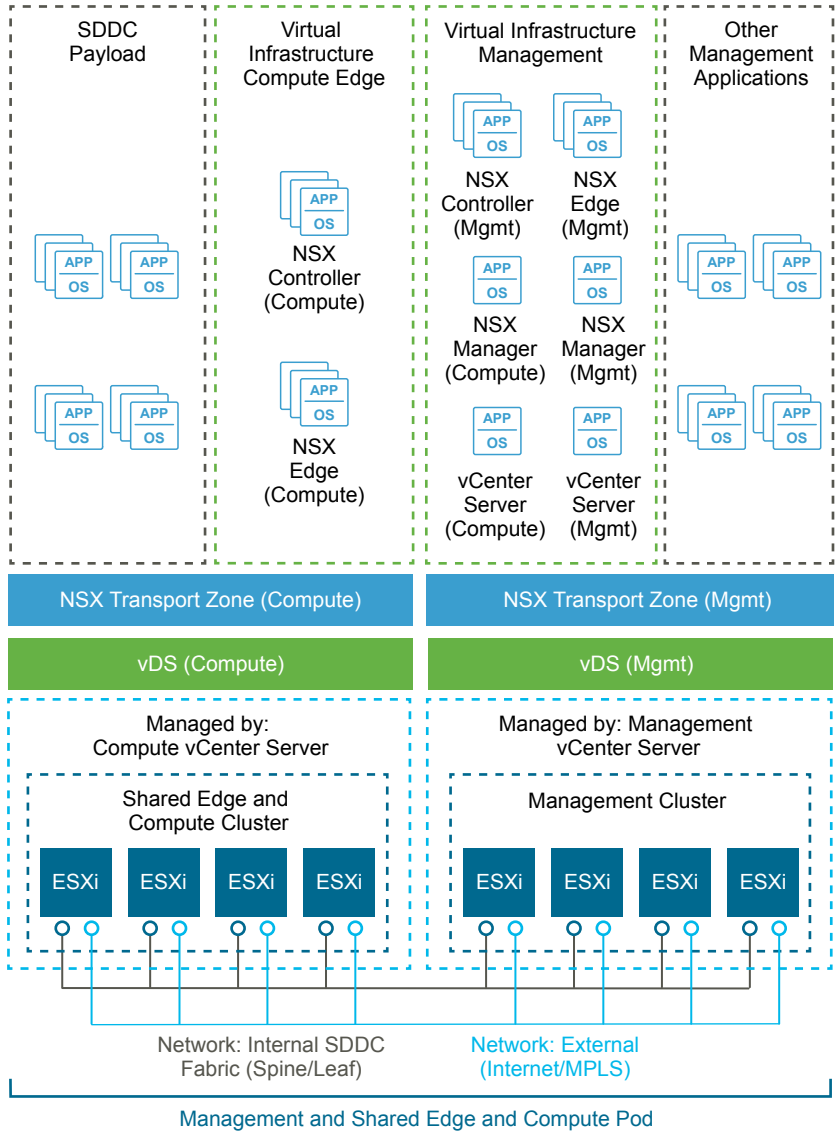
- NSX Controller cluster for the compute stack.
- All NSX Edge service gateways and DLR control VMs of the compute stack that are dedicated to handling the north/south traffic in the data center. A shared edge and compute stack helps prevent VLAN sprawl because any external VLANs need only be trunked to the hosts in this cluster.

Table 7-47. vSphere Cluster Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-007	For the compute stack, do not use a dedicated edge cluster.	Simplifies configuration and minimizes the number of hosts required for initial deployment.	The NSX Controller instances, NSX Edge services gateways, and DLR control VMs of the compute stack are deployed in the shared edge and compute cluster. The shared nature of the cluster will require the cluster to be scaled out as compute workloads are added so as to not impact network performance.
SDDC-VI-SDN-008	For the management stack, do not use a dedicated edge cluster.	The number of supported management applications does not justify the cost of a dedicated edge cluster in the management stack.	The NSX Controller instances, NSX Edge service gateways, and DLR control VMs of the management stack are deployed in the management cluster.
SDDC-VI-SDN-009	Apply vSphere Distributed Resource Scheduler (DRS) anti-affinity rules to the NSX components in both stacks.	Using DRS prevents controllers from running on the same ESXi host and thereby risking their high availability capability.	Additional configuration is required to set up anti-affinity rules.

The logical design of NSX considers the vCenter Server clusters and define the place where each NSX component runs.

Figure 7-10. Cluster Design for NSX for vSphere



High Availability of NSX for vSphere Components

The NSX Manager instances of both stacks run on the management cluster. vSphere HA protects the NSX Manager instances by ensuring that the NSX Manager VM is restarted on a different host in the event of primary host failure.

The NSX Controller nodes of the management stack run on the management cluster. The NSX for vSphere Controller nodes of the compute stack run on the shared edge and compute cluster. In both clusters, vSphere Distributed Resource Scheduler (DRS) rules ensure that NSX for vSphere Controller nodes do not run on the same host.

The data plane remains active during outages in the management and control planes although the provisioning and modification of virtual networks is impaired until those planes become available again.

The NSX Edge service gateways and DLR control VMs of the compute stack are deployed on the shared edge and compute cluster. The NSX Edge service gateways and DLR control VMs of the management stack run on the management cluster.

NSX Edge components that are deployed for north/south traffic are configured in equal-cost multi-path (ECMP) mode that supports route failover in seconds. NSX Edge components deployed for load balancing utilize NSX HA. NSX HA provides faster recovery than vSphere HA alone because NSX HA uses an active/passive pair of NSX Edge devices. By default the passive Edge device becomes active within 15 seconds. All NSX Edge devices are also protected by vSphere HA.

Scalability of NSX Components

A one-to-one mapping between NSX Manager instances and vCenter Server instances exists. If the inventory of either the management stack or the compute stack exceeds the limits supported by a single vCenter Server, then you can deploy a new vCenter Server instance, and must also deploy a new NSX Manager instance. You can extend transport zones by adding more shared edge and compute and compute clusters until you reach the vCenter Server limits. Consider the limit of 100 DLRs per ESXi host although the environment usually would exceed other vCenter Server limits before the DLR limit.

vSphere Distributed Switch Uplink Configuration

Each ESXi host utilizes two physical 10 Gb Ethernet adapters, associated with the uplinks on the vSphere Distributed Switches to which it is connected. Each uplink is connected to a different top-of-rack switch to mitigate the impact of a single top-of-rack switch failure and to provide two paths in and out of the SDDC.

Table 7-48. VTEP Teaming and Failover Configuration Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-010	Set up VXLAN Tunnel Endpoints (VTEPs) to use Route based on SRC-ID for teaming and failover configuration.	Allows for the utilization of the two uplinks of the vDS resulting in better bandwidth utilization and faster recovery from network path failures.	Link aggregation such as LACP between the top-of-rack (ToR) switches and ESXi host must not be configured in order to allow dynamic routing to peer between the ESGs and the upstream switches.

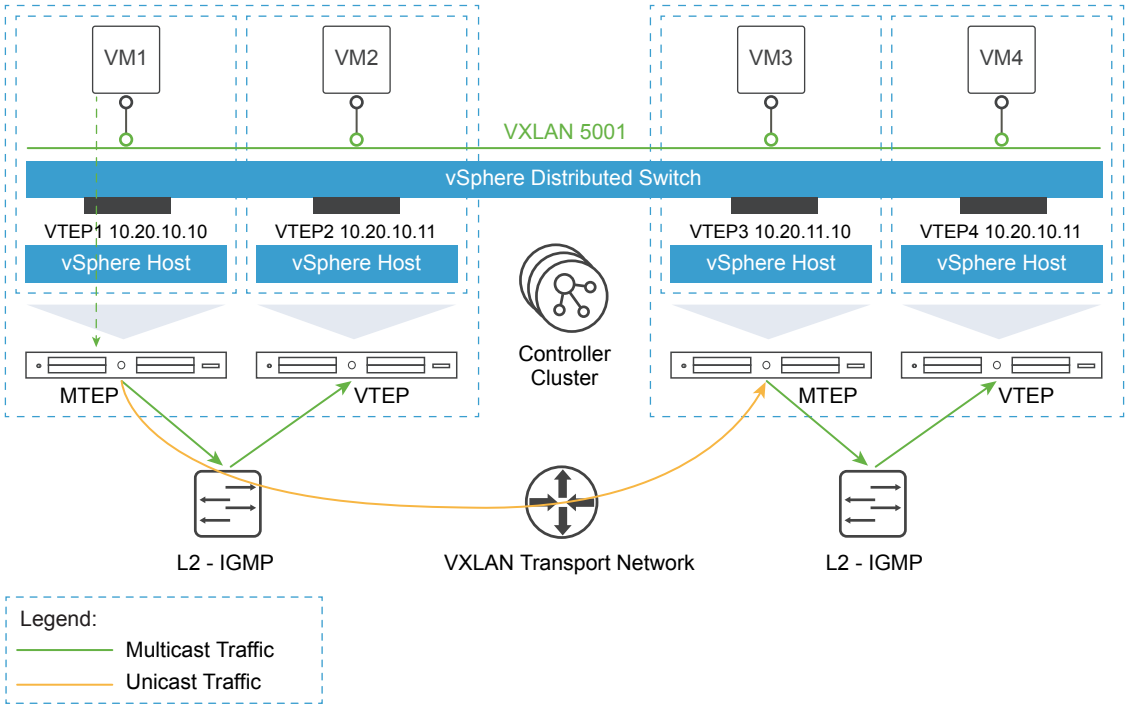
Logical Switch Control Plane Mode Design

The control plane decouples NSX for vSphere from the physical network and handles the broadcast, unknown unicast, and multicast (BUM) traffic within the logical switches. The control plane is on top of the transport zone and is inherited by all logical switches that are created within it. It is possible to override aspects of the control plane.

The following options are available.

Multicast Mode	The control plane uses multicast IP addresses on the physical network. Use multicast mode only when upgrading from existing VXLAN deployments. In this mode, you must configure PIM/IGMP on the physical network.
Unicast Mode	The control plane is handled by the NSX Controllers and all replication occurs locally on the host. This mode does not require multicast IP addresses or physical network configuration.
Hybrid Mode	This mode is an optimized version of the unicast mode where local traffic replication for the subnet is offloaded to the physical network. Hybrid mode requires IGMP snooping on the first-hop switch and access to an IGMP querier in each VTEP subnet. Hybrid mode does not require PIM.

Figure 7-11. Logical Switch Control Plane in Hybrid Mode



This design uses hybrid mode for control plane replication.

Table 7-49. Logical Switch Control Plane Mode Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-011	Use hybrid mode for control plane replication.	Offloading multicast processing to the physical network reduces pressure on VTEPs as the environment scales out. For large environments, hybrid mode is preferable to unicast mode. Multicast mode is used only when migrating from existing VXLAN solutions.	IGMP snooping must be enabled on the ToR physical switch and an IGMP querier must be available.

Transport Zone Design

A transport zone is used to define the scope of a VXLAN overlay network and can span one or more clusters within one vCenter Server domain. One or more transport zones can be configured in an NSX for vSphere solution. A transport zone is not meant to delineate a security boundary.

Note The design decisions for transport zones are forward looking to include expansion to dual-region design and management applications such as vRealize Automation.

Table 7-50. Transport Zones Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-012	For the compute stack, use a universal transport zone that encompasses all shared edge and compute, and compute clusters from all regions for workloads that require mobility between regions.	A Universal Transport zone supports extending networks and security policies across regions. This greater facilitates later expansion to a dual-region design because it allows seamless migration of applications across regions.	vRealize Automation is not able to deploy on demand network objects against a Secondary NSX Manager. You must consider that you can pair up to eight NSX Manager instances. If the solution grows past eight NSX Manager instances, you must deploy a new primary manager and new transport zone.
SDDC-VI-SDN-013	For the compute stack, use a global transport zone in each region that encompasses all shared edge and compute, and compute clusters for use with vRealize Automation on demand network provisioning.	NSX Managers with a role of Secondary can not deploy Universal objects. To allow all regions to deploy on demand network objects a global transport zone is required.	Shared Edge and Compute, and Compute Pods have two transport zones.
SDDC-VI-SDN-014	For the management stack, use a single universal transport zone that encompasses all management clusters.	A single Universal Transport zone supports extending networks and security policies across regions if you expand the design. This allows seamless migration of the management applications across regions either by cross-vCenter vMotion or by failover recovery with Site Recovery Manager.	You can pair up to eight NSX Manager instances. If the solution grows past eight NSX Manager instances, you must deploy a new primary manager and new transport zone.

Routing Design

The routing design considers different levels of routing within the environment from which to define a set of principles for designing a scalable routing solution.

Note The routing design decisions were made to facilitate easy expansion to a dual-region design and management applications such as vRealize Automation.

North/south The Provider Logical Router (PLR) handles the north/south traffic to and from a tenant and management applications inside of application virtual networks.

East/west Internal east/west routing at the layer beneath the PLR deals with the application workloads.

Table 7-51. Routing Model Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-017	Deploy NSX Edge Services Gateways in an ECMP configuration for north/south routing in both management and shared edge and compute clusters.	The NSX ESG is the recommended device for managing north/south traffic. Using ECMP provides multiple paths in and out of the SDDC. This results in faster failover times than deploying Edge service gateways in HA mode.	ECMP requires 2 VLANS for uplinks which adds an additional VLAN over traditional HA ESG configurations.
SDDC-VI-SDN-018	Deploy a single NSX UDLR for the management cluster to provide east/west routing across all regions.	Using the UDLR reduces the hop count between nodes attached to it to 1. This reduces latency and improves performance.	UDLRs are limited to 1,000 logical interfaces. When that limit is reached, a new UDLR must be deployed.
SDDC-VI-SDN-019	Deploy a single NSX UDLR for the shared edge and compute, and compute clusters to provide east/west routing across all regions for workloads that require mobility across regions.	Using the UDLR reduces the hop count between nodes attached to it to 1. This reduces latency and improves performance.	UDLRs are limited to 1,000 logical interfaces. When that limit is reached a new UDLR must be deployed.
SDDC-VI-SDN-020	Deploy a DLR for the shared edge and compute and compute clusters to provide east/west routing for workloads that require on demand network objects from vRealize Automation.	Using the DLR reduces the hop count between nodes attached to it to 1. This reduces latency and improves performance.	DLRs are limited to 1,000 logical interfaces. When that limit is reached a new DLR must be deployed.
SDDC-VI-SDN-021	Deploy all NSX UDLRs without the local egress option enabled.	When local egress is enabled, control of ingress traffic, is also necessary (for example using NAT). This becomes hard to manage for little to no benefit.	All north/south traffic is routed through Region A until those routes are no longer available. At that time, all traffic dynamically changes to Region B.

Table 7-51. Routing Model Design Decisions (Continued)

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-022	Use BGP as the dynamic routing protocol inside the SDDC.	Using BGP as opposed to OSPF eases the implementation of dynamic routing. There is no need to plan and design access to OSPF area 0 inside the SDDC. OSPF area 0 varies based on customer configuration.	BGP requires configuring each ESG and UDLR with the remote router that it exchanges routes with.
SDDC-VI-SDN-023	Configure BGP Keep Alive Timer to 1 and Hold Down Timer to 3 between the UDLR and all ESGs that provide north/south routing.	With Keep Alive and Hold Timers between the UDLR and ECMP ESGs set low, a failure is detected quicker, and the routing table is updated faster.	If an ESXi host becomes resource constrained, the ESG running on that host might no longer be used even though it is still up.
SDDC-VI-SDN-024	Configure BGP Keep Alive Timer to 4 and Hold Down Timer to 12 between the ToR switches and all ESGs providing north/south routing.	This provides a good balance between failure detection between the ToRs and the ESGs and overburdening the ToRs with keep alive traffic.	By using longer timers to detect when a router is dead, a dead router stays in the routing table longer and continues to send traffic to a dead router.
SDDC-VI-SDN-025	Create one or more static routes on ECMP enabled edges for subnets behind the UDLR and DLR with a higer admin cost then the dynamically learned routes.	When the UDLR or DLR control VM fails over router adjacency is lost and routes from upstream devices such as ToR's to subnets behind the UDLR are lost.	This requires each ECMP edge device be configured with static routes to the UDLR or DLR. If any new subnets are added behind the UDLR or DLR the routes must be updated on the ECMP edges.

Transit Network and Dynamic Routing

Dedicated networks are needed to facilitate traffic between the universal dynamic routers and edge gateways, and to facilitate traffic between edge gateways and the top of rack switches. These networks are used for exchanging routing tables and for carrying transit traffic.

Table 7-52. Transit Network Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-026	Create a universal virtual switch for use as the transit network between the UDLR and ESGs. The UDLR provides east/west routing in both compute and management stacks while the ESG's provide north/south routing.	The universal virtual switch allows the UDLR and all ESGs across regions to exchange routing information if you expand to a dual region design.	Only the primary NSX Manager can create and manage universal objects including this UDLR.
SDDC-VI-SDN-027	Create a global virtual switch in each region for use as the transit network between the DLR and ESG's. The DLR provides east/west routing in the compute stack while the ESG's provide north/south routing.	The global virtual switch allows the DLR and ESGs in each region to exchange routing information.	A global virtual switch for use as a transit network is required in each region.
SDDC-VI-SDN-028	Create two VLANs in each region. Use those VLANs to enable ECMP between the north/south ESGs and the ToR switches. The ToR's have an SVI on one of the two VLANS and each north/south ESG has an interface on each VLAN.	This enables the ESGs to have multiple equal-cost routes and provides more resiliency and better bandwidth utilization in the network.	Extra VLANs are required.

Firewall Logical Design

The NSX Distributed Firewall is used to protect all management applications attached to application virtual networks. To secure the SDDC, only other solutions in the SDDC and approved administration IPs can directly communicate with individual components. External facing portals are accessible via a load balancer virtual IP (VIP). This simplifies the design by having a single point of administration for all firewall rules. The firewall on individual ESGs is set to allow all traffic. An exception are ESGs that provide ECMP services, which require the firewall to be disabled.

Table 7-53. Firewall Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-029	For all ESGs deployed as load balancers, set the default firewall rule to allow all traffic.	Restricting and granting access is handled by the distributed firewall. The default firewall rule does not have to do it.	Explicit rules to allow access to management applications must be defined in the distributed firewall.
SDDC-VI-SDN-030	For all ESGs deployed as ECMP north/south routers, disable the firewall.	Use of ECMP on the ESGs is a requirement. Leaving the firewall enabled, even in allow all traffic mode, results in sporadic network connectivity.	Services such as NAT and load balancing can not be used when the firewall is disabled.
SDDC-VI-SDN-031	Configure the Distributed Firewall to limit access to administrative interfaces in the management cluster.	To ensure only authorized administrators can access the administrative interfaces of management applications.	Maintaining firewall rules adds administrative overhead.

Load Balancer Design

The ESG implements load balancing within NSX for vSphere.

The ESG has both a Layer 4 and a Layer 7 engine that offer different features, which are summarized in the following table.

Feature	Layer 4 Engine	Layer 7 Engine
Protocols	TCP	TCP HTTP HTTPS (SSL Pass-through) HTTPS (SSL Offload)
Load balancing method	Round Robin Source IP Hash Least Connection	Round Robin Source IP Hash Least Connection URI
Health checks	TCP	TCP HTTP (GET, OPTION, POST) HTTPS (GET, OPTION, POST)

Feature	Layer 4 Engine	Layer 7 Engine
Persistence (keeping client connections to the same back-end server)	TCP: SourceIP	TCP: SourceIP, MSRDP HTTP: SourceIP, Cookie HTTPS: SourceIP, Cookie, ssl_session_id
Connection throttling	No	Client Side: Maximum concurrent connections, Maximum new connections per second Server Side: Maximum concurrent connections
High availability	Yes	Yes
Monitoring	View VIP (Virtual IP), Pool and Server objects and stats via CLI and API View global stats for VIP sessions from the vSphere Web Client	View VIP, Pool and Server objects and statistics by using CLI and API View global statistics about VIP sessions from the vSphere Web Client
Layer 7 manipulation	No	URL block, URL rewrite, content rewrite

Table 7-54. NSX for vSphere Load Balancer Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-032	Use the NSX load balancer.	The NSX load balancer can support the needs of the management applications. Using another load balancer would increase cost and add another component to be managed as part of the SDDC.	None.
SDDC-VI-SDN-033	Use an NSX load balancer in HA mode for all management applications.	All management applications that require a load balancer are on a single virtual wire, having a single load balancer keeps the design simple.	One management application owner could make changes to the load balancer that impact another application.
SDDC-VI-SDN-034	Use an NSX load balancer in HA mode for the Platform Services Controllers.	Using a load balancer increases the availability of the PSC's for all applications.	Configuring the Platform Services Controllers and the NSX load balancer adds administrative overhead.

Bridging Physical Workloads

NSX for vSphere offers VXLAN to Layer 2 VLAN bridging capabilities with the data path contained entirely in the ESXi hypervisor. The bridge runs on the ESXi host where the DLR control VM is located. Multiple bridges per DLR are supported.

Table 7-55. Virtual to Physical Interface Type Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-037	Place all virtual machines, both management and tenant, on VXLAN-backed networks unless you must satisfy an explicit requirement to use VLAN-backed port groups for these virtual machines. If VLAN-backed port groups are required, connect physical workloads that need to communicate to virtualized workloads to routed VLAN LIFs on a DLR.	Bridging and routing are not possible on the same logical switch. As a result, it makes sense to attach a VLAN LIF to a distributed router or ESG and route between the physical and virtual machines. Use bridging only where virtual machines need access only to the physical machines on the same Layer 2.	Access to physical workloads is routed via the DLR or ESG.

Application Virtual Network

Management applications, leverage a traditional 3-tier client/server architecture with a presentation tier (user interface), functional process logic tier, and data tier. The micro-segmentation use case includes only vRealize Log Insight, though other applications such as vRealize Automation can be added as the designs grows to a full software-defined data center.

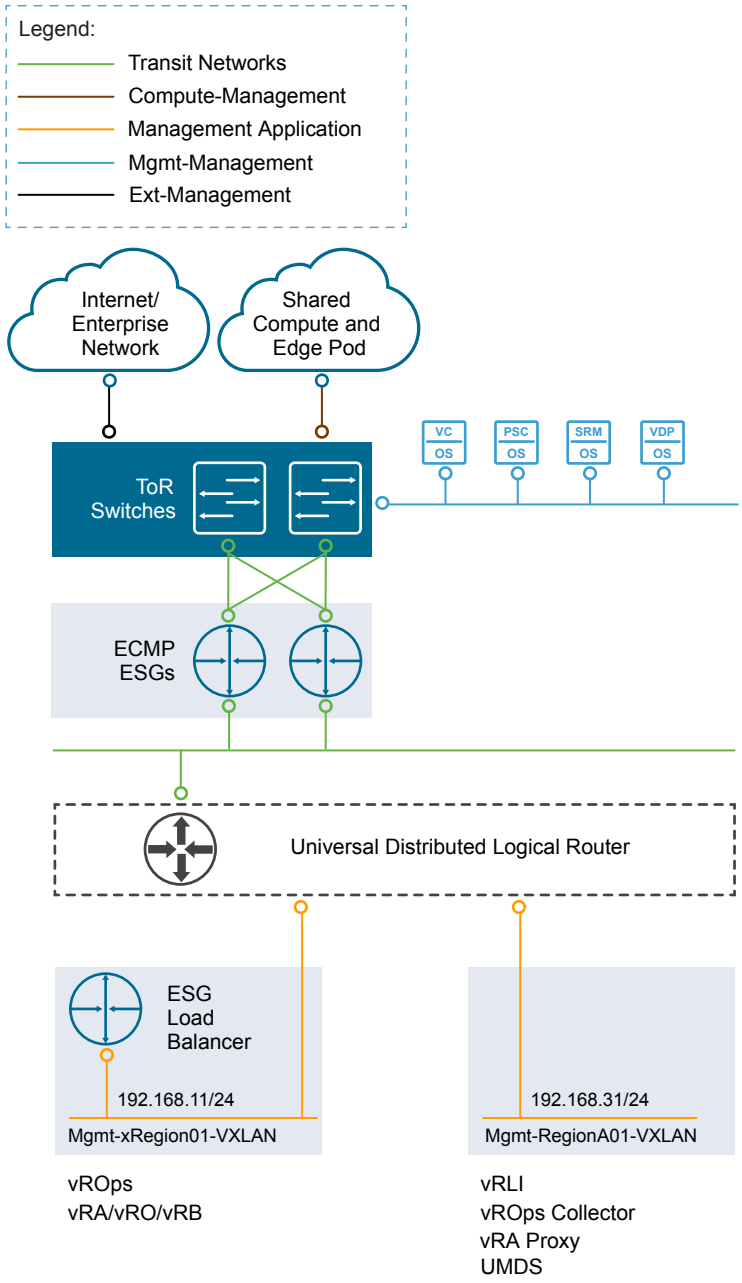
Table 7-56. Isolated Management Applications Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
SDDC-VI-SDN-038	Place vRealize Log Insight on an application virtual network.	Access to any management applications is only through published access points.	The application virtual network is fronted by an NSX Edge device for load balancing and the distributed firewall to isolate applications from each other and external users. Direct access to application virtual networks is controlled by distributed firewall rules.

Having software-defined networking based on NSX in the management stack makes all NSX features available to the management applications.

This approach to network virtualization service design improves security and mobility of the management applications, and reduces the integration effort with existing customer networks.

Figure 7-12. Virtual Application Network Components and Design



Certain configuration choices might later facilitate the tenant onboarding process.

- Create the primary NSX ESG to act as the tenant PLR and the logical switch that forms the transit network for use in connecting to the UDLR.
- Connect the primary NSX ESG uplinks to the external networks
- Connect the primary NSX ESG internal interface to the transit network.
- Create the NSX UDLR to provide routing capabilities for tenant internal networks and connect the UDLR uplink to the transit network.
- Create any tenant networks that are known up front and connect them to the UDLR.

Use of Secure Sockets Layer (SSL) Certificates

By default NSX Manager uses a self signed SSL certificate. By default, this certificate is not trusted by end-user devices or browsers. It is a security best practice to replace these certificates with certificates that are signed by a third-party or enterprise Certificate Authority (CA).

Design ID	Design Decision	Design Justification	Design Implication
SDDC-VI-SDN-043	Replace the NSX Manager certificate with a certificate signed by a 3rd party Public Key Infrastructure.	Ensures communication between NSX admins and the NSX Manager are encrypted by a trusted certificate.	Replacing and managing certificates is an operational overhead.

Shared Storage Design

The shared storage design includes design decisions for vSAN storage and NFS storage.

Well-designed shared storage provides the basis for an SDDC and has the following benefits.

- Prevents unauthorized access to business data
- Protects data from hardware and software failures
- Protects data from malicious or accidental corruption

Follow these guidelines when designing shared storage for your environment.

- Optimize the storage design to meet the diverse needs of applications, services, administrators, and users.
- Strategically align business applications and the storage infrastructure to reduce costs, boost performance, improve availability, provide security, and enhance functionality.
- Provide multiple tiers of storage to match application data access to application requirements.
- Design each tier of storage with different performance, capacity, and availability characteristics. Because not every application requires expensive, high-performance, highly available storage, designing different storage tiers reduces cost.

Shared Storage Platform

You can choose between traditional storage, VMware vSphere Virtual Volumes, and VMware vSAN storage.

Storage Types

Traditional Storage	Fibre Channel, NFS, and iSCSI are mature and viable options to support virtual machine needs.
VMware vSAN Storage	vSAN is a software-based distributed storage platform that combines the compute and storage resources of VMware ESXi hosts. When you design and size a vSAN cluster, hardware choices are more limited than for traditional storage.
VMware vSphere Virtual Volumes	This design does not leverage VMware vSphere Virtual Volumes because Virtual Volumes does not support Site Recovery Manager.

Traditional Storage and vSAN Storage

Fibre Channel, NFS, and iSCSI are mature and viable options to support virtual machine needs.

Your decision to implement one technology or another can be based on performance and functionality, and on considerations like the following:

- The organization’s current in-house expertise and installation base
- The cost, including both capital and long-term operational expenses
- The organization’s current relationship with a storage vendor

vSAN is a software-based distributed storage platform that combines the compute and storage resources of ESXi hosts. It provides a simple storage management experience for the user. This solution makes software-defined storage a reality for VMware customers. However, you must carefully consider supported hardware options when sizing and designing a vSAN cluster.

Storage Type Comparison

ESXi hosts support a variety of storage types. Each storage type supports different vSphere features.

Table 7-57. Network Shared Storage Supported by ESXi Hosts

Technology	Protocols	Transfers	Interface
Fibre Channel	FC/SCSI	Block access of data/LUN	Fibre Channel HBA
Fibre Channel over Ethernet	FCoE/SCSI	Block access of data/LUN	Converged network adapter (hardware FCoE) NIC with FCoE support (software FCoE)
iSCSI	IP/SCSI	Block access of data/LUN	iSCSI HBA or iSCSI enabled NIC (hardware iSCSI) Network Adapter (software iSCSI)
NAS	IP/NFS	File (no direct LUN access)	Network adapter
vSAN	IP	Block access of data	Network adapter

Table 7-58. vSphere Features Supported by Storage Type

Type	vSphere vMotion	Datastore	Raw Device Mapping (RDM)	Application or Block-level Clustering	HA/DRS	Storage APIs Data Protection
Local Storage	Yes	VMFS	No	Yes	No	Yes
Fibre Channel / Fibre Channel over Ethernet	Yes	VMFS	Yes	Yes	Yes	Yes
iSCSI	Yes	VMFS	Yes	Yes	Yes	Yes
NAS over NFS	Yes	NFS	No	No	Yes	Yes
vSAN	Yes	vSAN	No	Yes (via iSCSI Initiator)	Yes	Yes

Shared Storage Logical Design

The shared storage design selects the appropriate storage device for each type of cluster.

The storage devices for use by each type of cluster are as follows.

- Management clusters use vSAN for primary storage and NFS for secondary storage.
- Shared edge and compute clusters can use FC/FCoE, iSCSI, NFS, or vSAN storage. No specific guidance is given as user workloads and other factors determine storage type and SLA for user workloads.

Figure 7-13. Logical Storage Design

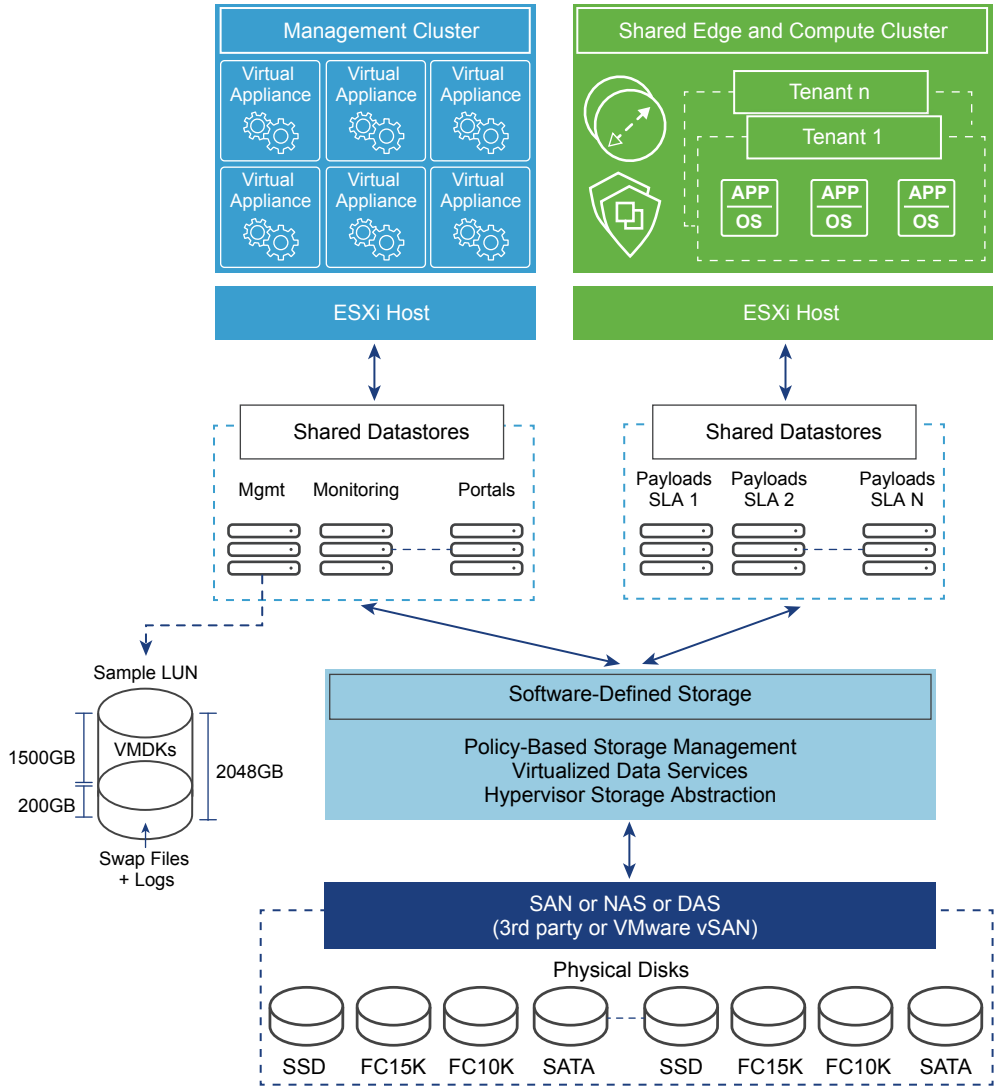


Table 7-59. Storage Type Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-001	In the management cluster, use vSAN and NFS shared storage: ■ Use vSAN as the primary shared storage platform. ■ Use NFS as the secondary shared storage platform for the management cluster.	vSAN as the primary shared storage solution can take advantage of more cost-effective local storage. NFS is used primarily for archival and the need to maintain historical data. Leveraging NFS provides large, low cost volumes that have the flexibility to be expanded on a regular basis depending on capacity needs.	The use of two different storage technologies increases the complexity and operational overhead.
SDDC-VI-Storage-002	In all clusters, ensure that at least 20% of free space is always available on all non-vSAN datastores.	If the datastore runs out of free space, applications and services within the SDDC, including but not limited to the NSX Edge core network services, the provisioning portal and VDP backups, will fail. To prevent this, maintain adequate free space.	Monitoring and capacity management are critical, and must be proactively performed.

Storage Tiering

Today's enterprise-class storage arrays contain multiple drive types and protection mechanisms. The storage, server, and application administrators face challenges when selecting the correct storage configuration for each application being deployed in the environment. Virtualization can make this problem more challenging by consolidating many different application workloads onto a small number of large devices. Given this challenge, administrators might use single storage type for every type of workload without regard to the needs of the particular workload. However, not all application workloads have the same requirements, and storage tiering allows for these differences by creating multiple levels of storage with varying degrees of performance, reliability and cost, depending on the application workload needs.

The most mission-critical data typically represents the smallest amount of data and offline data represents the largest amount. Details differ for different organizations.

To determine the storage tier for application data, determine the storage characteristics of the application or service.

- I/O operations per second (IOPS) requirements
- Megabytes per second (MBps) requirements
- Capacity requirements
- Availability requirements
- Latency requirements

After you determine the information for each application, you can move the application to the storage tier with matching characteristics.

- Consider any existing service-level agreements (SLAs).
- Move data between storage tiers during the application life cycle as needed.

VMware Hardware Acceleration API/CLI for Storage

The VMware Hardware Acceleration API/CLI for storage (previously known as vStorage APIs for Array Integration or VAAI), supports a set of ESXCLI commands for enabling communication between ESXi hosts and storage devices. The APIs define a set of storage primitives that enable the ESXi host to offload certain storage operations to the array. Offloading the operations reduces resource overhead on the ESXi hosts and can significantly improve performance for storage-intensive operations such as storage cloning, zeroing, and so on. The goal of hardware acceleration is to help storage vendors provide hardware assistance to speed up VMware I/O operations that are more efficiently accomplished in the storage hardware.

Without the use of VAAI, cloning or migration of virtual machines by the VMkernel data mover involves software data movement. The data mover issues I/O to read and write blocks to and from the source and destination datastores. With VAAI, the data mover can use the API primitives to offload operations to the array when possible. For example, when you copy a virtual machine disk file (VMDK file) from one

datastore to another inside the same array, the data mover directs the array to make the copy completely inside the array. If you invoke a data movement operation and the corresponding hardware offload operation is enabled, the data mover first attempts to use hardware offload. If the hardware offload operation fails, the data mover reverts to the traditional software method of data movement.

In nearly all cases, hardware data movement performs significantly better than software data movement. It consumes fewer CPU cycles and less bandwidth on the storage fabric. Timing operations that use the VAAI primitives and use `esxtop` to track values such as CMDS/s, READS/s, WRITES/s, MBREAD/s, and MBWRTN/s of storage adapters during the operation show performance improvements.

Table 7-60. vStorage APIs for Array Integration Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-003	Select an array that supports VAAI over NAS (NFS).	VAAI offloads tasks to the array itself, enabling the ESXi hypervisor to use its resources for application workloads and not become a bottleneck in the storage subsystem. VAAI is required to support the desired number of virtual machine lifecycle operations.	Not all VAAI arrays support VAAI over NFS. A plugin from the array vendor is required to enable this functionality.

Virtual Machine Storage Policies

You can create a storage policy for a virtual machine to specify which storage capabilities and characteristics are the best match for this virtual machine.

Note vSAN uses storage policies to allow specification of the characteristics of virtual machines, so you can define the policy on an individual disk level rather than at the volume level for vSAN.

You can identify the storage subsystem capabilities by using the VMware vSphere API for Storage Awareness or by using a user-defined storage policy.

VMware vSphere API for Storage Awareness (VASA)	With vSphere API for Storage Awareness, storage vendors can publish the capabilities of their storage to VMware vCenter Server, which can display these capabilities in its user interface.
User-defined storage policy	Defined by using the VMware Storage Policy SDK or VMware vSphere PowerCL, or from the vSphere Web Client.

You can assign a storage policy to a virtual machine and periodically check for compliance so that the virtual machine continues to run on storage with the correct performance and availability characteristics.

You can associate a virtual machine with a virtual machine storage policy when you create, clone, or migrate that virtual machine. If a virtual machine is associated with a storage policy, the vSphere Web Client shows the datastores that are compatible with the policy. You can select a datastore or datastore cluster. If you select a datastore that does not match the virtual machine storage policy, the vSphere Web Client shows that the virtual machine is using non-compliant storage. See *Creating and Managing vSphere Storage Policies* in the vSphere 6.5 documentation.

Table 7-61. Virtual Machine Storage Policy Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-004	Use the default vSAN storage policy for all virtual machines in the management cluster.	The default vSAN storage policy is adequate for the management cluster VMs.	If third party or additional VMs have different storage requirements, additional VM storage policies may be required.

vSphere Storage I/O Control Design

VMware vSphere Storage I/O Control allows cluster-wide storage I/O prioritization, which results in better workload consolidation and helps reduce extra costs associated with over provisioning.

vSphere Storage I/O Control extends the constructs of shares and limits to storage I/O resources. You can control the amount of storage I/O that is allocated to virtual machines during periods of I/O congestion, so that more important virtual machines get preference over less important virtual machines for I/O resource allocation.

When vSphere Storage I/O Control is enabled on a datastore, the ESXi host monitors the device latency when communicating with that datastore. When device latency exceeds a threshold, the datastore is considered to be congested and each virtual machine that accesses that datastore is allocated I/O resources in proportion to their shares. Shares are set on a per-virtual machine basis and can be adjusted.

vSphere Storage I/O Control has several requirements, limitations, and constraints.

- Datastores that are enabled with vSphere Storage I/O Control must be managed by a single vCenter Server system.
- Storage I/O Control is supported on Fibre Channel-connected, iSCSI-connected, and NFS-connected storage. RDM is not supported.
- Storage I/O Control does not support datastores with multiple extents.
- Before using vSphere Storage I/O Control on datastores that are backed by arrays with automated storage tiering capabilities, check the *VMware Compatibility Guide* whether the storage array has been certified a compatible with vSphere Storage I/O Control.

Table 7-62. Storage I/O Control Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-005	Enable Storage I/O Control with the default values on all non vSAN datastores.	Storage I/O Control ensures that all virtual machines on a datastore receive an equal amount of I/O.	Virtual machines that use more I/O are throttled to allow other virtual machines access to the datastore only when contention occurs on the datastore.

Datastore Cluster Design

A datastore cluster is a collection of datastores with shared resources and a shared management interface. Datastore clusters are to datastores what clusters are to ESXi hosts. After you create a datastore cluster, you can use vSphere Storage DRS to manage storage resources.

vSphere datastore clusters group similar datastores into a pool of storage resources. When vSphere Storage DRS is enabled on a datastore cluster, vSphere automates the process of initial virtual machine file placement and balances storage resources across the cluster to avoid bottlenecks. vSphere Storage DRS considers datastore space usage and I/O load when making migration recommendations.

When you add a datastore to a datastore cluster, the datastore's resources become part of the datastore cluster's resources. The following resource management capabilities are also available for each datastore cluster.

Capability	Description
Space utilization load balancing	You can set a threshold for space use. When space use on a datastore exceeds the threshold, vSphere Storage DRS generates recommendations or performs migrations with vSphere Storage vMotion to balance space use across the datastore cluster.
I/O latency load balancing	You can configure the I/O latency threshold to avoid bottlenecks. When I/O latency on a datastore exceeds the threshold, vSphere Storage DRS generates recommendations or performs vSphere Storage vMotion migrations to help alleviate high I/O load.
Anti-affinity rules	You can configure anti-affinity rules for virtual machine disks to ensure that the virtual disks of a virtual machine are kept on different datastores. By default, all virtual disks for a virtual machine are placed on the same datastore.

You can enable vSphere Storage I/O Control or vSphere Storage DRS for a datastore cluster. You can enable the two features separately, even though vSphere Storage I/O control is enabled by default when you enable vSphere Storage DRS.

vSphere Storage DRS Background Information

vSphere Storage DRS supports automating the management of datastores based on latency and storage utilization. When configuring vSphere Storage DRS, verify that all datastores use the same version of VMFS and are on the same storage subsystem. Because vSphere Storage vMotion performs the migration of the virtual machines, confirm that all prerequisites are met.

vSphere Storage DRS provides a way of balancing usage and IOPS among datastores in a storage cluster:

- Initial placement of virtual machines is based on storage capacity.
- vSphere Storage DRS uses vSphere Storage vMotion to migrate virtual machines based on storage capacity.
- vSphere Storage DRS uses vSphere Storage vMotion to migrate virtual machines based on I/O latency.
- You can configure vSphere Storage DRS to run in either manual mode or in fully automated mode.

vSphere vStorage I/O Control and vSphere Storage DRS manage latency differently.

- vSphere Storage I/O Control distributes the resources based on virtual disk share value after a latency threshold is reached.
- vSphere Storage DRS measures latency over a period of time. If the latency threshold of vSphere Storage DRS is met in that time frame, vSphere Storage DRS migrates virtual machines to balance latency across the datastores that are part of the cluster.

When making a vSphere Storage design decision, consider these points:

- Use vSphere Storage DRS where possible.
- vSphere Storage DRS provides a way of balancing usage and IOPS among datastores in a storage cluster:
    - Initial placement of virtual machines is based on storage capacity.

- vSphere Storage vMotion is used to migrate virtual machines based on storage capacity.
- vSphere Storage vMotion is used to migrate virtual machines based on I/O latency.
- vSphere Storage DRS can be configured in either manual or fully automated modes

vSAN Storage Design

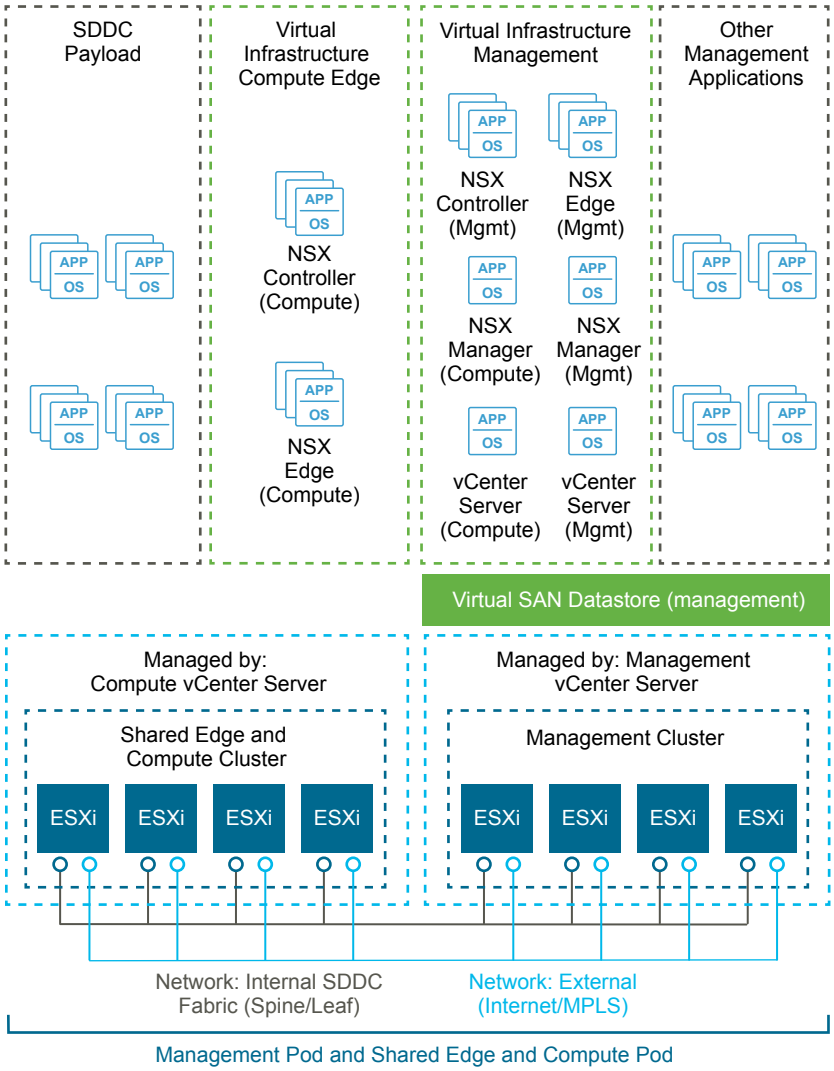
VMware vSAN Storage design in this VMware Validated Design includes conceptual design, logical design, network design, cluster and disk group design, and policy design.

VMware vSAN Conceptual Design and Logical Design

This vSAN design is limited to the management cluster. The design uses the default storage policy to achieve redundancy and performance within the cluster. While VMware vSAN can also be used within the shared edge and compute cluster, this design currently gives no guidance for the implementation.

In a cluster that is managed by vCenter Server, you can manage software-defined storage resources just as you can manage compute resources. Instead of CPU or memory reservations, limits, and shares, you can define storage policies and assign them to virtual machines. The policies specify the characteristics of the storage and can be changed as business requirements change.

Figure 7-14. Conceptual vSAN Design



VMware vSAN Network Design

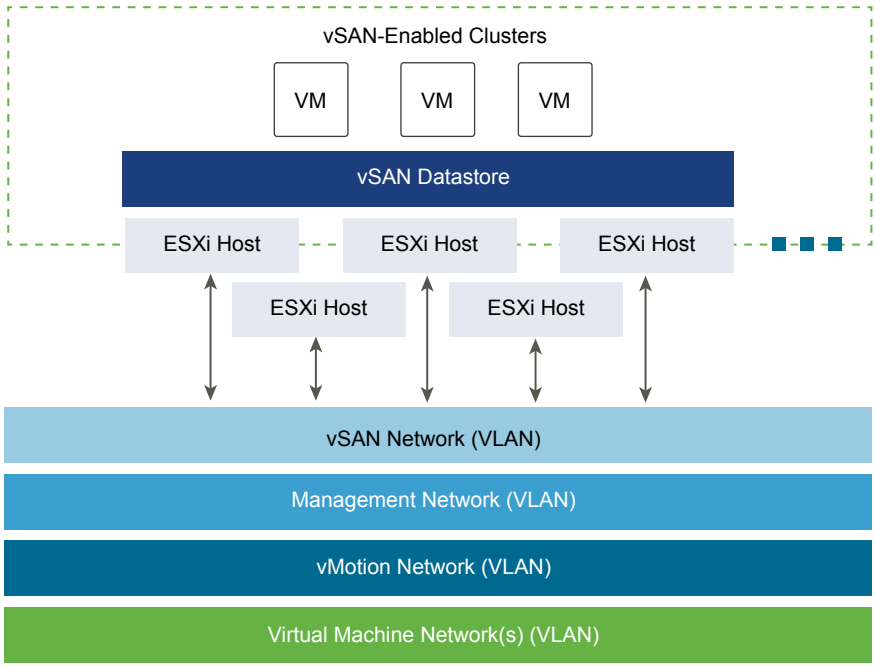
When performing network configuration, you have to consider the traffic and decide how to isolate vSAN traffic.

- Consider how much replication and communication traffic is running between hosts. With VMware vSAN, the amount of traffic depends on the number of VMs that are running in the cluster, and on how write-intensive the I/O is for the applications running in the VMs.
- Isolate vSAN traffic on its own Layer 2 network segment. You can do this with dedicated switches or ports, or by using a VLAN.

The vSAN VMkernel port group is created as part of cluster creation. Configure this port group on all hosts in a cluster, even for hosts that are not contributing storage resources to the cluster.

The following diagram illustrates the logical design of the network.

Figure 7-15. VMware vSAN Conceptual Network



Network Bandwidth Requirements

VMware recommends that solutions use a 10 Gb Ethernet connection for use with vSAN to ensure the best and most predictable performance (IOPS) for the environment. Without it, a significant decrease in array performance results.

Note vSAN all-flash configurations are supported only with 10 GbE.

Table 7-63. Network Speed Selection

Design Quality	1Gb	10Gb	Comments
Availability	o	o	Neither design option impacts availability.
Manageability	o	o	Neither design option impacts manageability.
Performance	↓	↑	Faster network speeds increase vSAN performance (especially in I/O intensive situations).
Recoverability	↓	↑	Faster network speeds increase the performance of rebuilds and synchronizations in the environment. This ensures that VMs are properly protected from failures.
Security	o	o	Neither design option impacts security.

Legend: ↑ = positive impact on quality; ↓ = negative impact on quality; o = no impact on quality.

Table 7-64. Network Bandwidth Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-SDS-001	Use only 10 GbE for VMware vSAN traffic.	Performance with 10 GbE is optimal. Without it, a significant decrease in array performance results.	The physical network must support 10 Gb networking between every host in the vSAN clusters.

VMware vSAN Virtual Switch Type

vSAN supports the use of vSphere Standard Switch or vSphere Distributed Switch. The benefit of using vSphere Distributed Switch is that it supports Network I/O Control which allows for prioritization of bandwidth in case of contention in an environment.

This design uses a vSphere Distributed Switch for the vSAN port group to ensure that priority can be assigned using Network I/O Control to separate and guarantee the bandwidth for vSAN traffic.

Virtual Switch Design Background

Virtual switch type affects performance and security of the environment.

Table 7-65. Virtual Switch Types

Design Quality	vSphere Standard Switch	vSphere Distributed Switch	Comments
Availability	o	o	Neither design option impacts availability.
Manageability	↓	↑	The vSphere Distributed Switch is centrally managed across all hosts, unlike the standard switch which is managed on each host individually.
Performance	↓	↑	The vSphere Distributed Switch has added controls, such as Network I/O Control, which you can use to guarantee performance for vSAN traffic.
Recoverability	↓	↑	The vSphere Distributed Switch configuration can be backed up and restored, the standard switch does not have this functionality.
Security	↓	↑	The vSphere Distributed Switch has added built-in security controls to help protect traffic.

Legend: ↑ = positive impact on quality; ↓ = negative impact on quality; o = no impact on quality.

Table 7-66. Virtual Switch Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-SDS-002	Use the existing vSphere Distributed Switch instances in the management clusters.	Provide guaranteed performance for vSAN traffic in case of contention by using existing networking components.	All traffic paths are shared over common uplinks.

Jumbo Frames

VMware vSAN supports jumbo frames for vSAN traffic.

A VMware vSAN design should use jumbo frames only if the physical environment is already configured to support them, they are part of the existing design, or if the underlying configuration does not create a significant amount of added complexity to the design.

Table 7-67. Jumbo Frames Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-SDS-003	Configure jumbo frames on the VLAN dedicated to vSAN traffic.	Jumbo frames are already used to improve performance of vSphere vMotion and NFS storage traffic.	Every device in the network must support jumbo frames.

VLANs

VMware recommends isolating VMware vSAN traffic on its own VLAN. When a design uses multiple vSAN clusters, each cluster should use a dedicated VLAN or segment for its traffic. This approach prevents interference between clusters and helps with troubleshooting cluster configuration.

Table 7-68. VLAN Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-SDS-004	Use a dedicated VLAN for vSAN traffic for each vSAN enabled cluster.	VLANs ensure traffic isolation.	VLANs span only a single pod. A sufficient number of VLANs are available within each pod and should be used for traffic segregation.

Multicast Requirements

VMware vSAN requires that IP multicast is enabled on the Layer 2 physical network segment that is used for intra-cluster communication. All VMkernel ports on the vSAN network subscribe to a multicast group using Internet Group Management Protocol (IGMP).

A default multicast address is assigned to each vSAN cluster at the time of creation. IGMP (v3) snooping is used to limit Layer 2 multicast traffic to specific port groups. As per the Physical Network Design, IGMP snooping is configured with an IGMP snooping querier to limit the physical switch ports that participate in the multicast group to only vSAN VMkernel port uplinks. In some cases, an IGMP snooping querier can be associated with a specific VLAN. However, vendor implementations might differ.

Cluster and Disk Group Design

When considering the cluster and disk group design, you have to decide on the vSAN datastore size, number of hosts per cluster, number of disk groups per host, and the vSAN policy.

VMware vSAN Datastore Size

The size of the VMware vSAN datastore depends on the requirements for the datastore. Consider cost versus availability to provide the appropriate sizing.

Table 7-69. VMware vSAN Datastore Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-SDS-005	Provide the Management cluster with a minimum of 8TB of raw capacity for vSAN.	Management cluster virtual machines that use VMware vSAN require at least 8 TB of raw storage. NFS is used as secondary shared storage of some management components such as backups and log archives.	None
SDDC-VI-Storage-SDS-006	On all VSAN datastores , ensure that at least 30% of free space is always available.	When VSAN reaches 80% usage a re-balance task is started which can be resource intensive.	Increases the amount of available storage needed.

Number of Hosts Per Cluster

The number of hosts in the cluster depends on these factors:

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

For example, if the vSAN cluster has only 3 ESXi hosts, only a single failure is supported. If a higher level of availability is required, additional hosts are required.

Cluster Size Design Background

Table 7-70. Number of Hosts Per Cluster

Design Quality	3 Hosts	32 Hosts	64 Hosts	Comments
Availability	↓	↑	↑↑	The more hosts that are available in the cluster, the more failures the cluster can tolerate.
Manageability	↓	↑	↑	The more hosts in the cluster, the more virtual machines can be in the VMware vSAN environment.
Performance	↑	↓	↓	Having a larger cluster can impact performance if there is an imbalance of resources. Consider performance as you make your decision.
Recoverability	o	o	o	Neither design option impacts recoverability.
Security	o	o	o	Neither design option impacts security.

Legend: ↑ = positive impact on quality; ↓ = negative impact on quality; o = no impact on quality.

Table 7-71. Cluster Size Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-SDS-007	Configure the Management cluster with a minimum of 4 ESXi hosts to support VMware vSAN.	Having 4 hosts addresses the availability and sizing requirements, and allows you to take an ESXi host offline for maintenance or upgrades without impacting the overall VMware vSAN cluster health.	The availability requirements for the management cluster might cause underutilization of the cluster hosts.

Number of Disk Groups Per Host

Disk group sizing is an important factor during volume design.

- If more hosts are available in the cluster, more failures are tolerated in the cluster. This capability adds cost because additional hardware for the disk groups is required.
- More available disk groups can increase the recoverability of VMware vSAN during a failure.

Consider these data points when deciding on the number of disk groups per host:

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

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

Disk Groups Design Background

The number of disk groups can affect availability and performance.

Table 7-72. Number of Disk Groups Per Host

Design Quality	1 Disk Group	3 Disk Groups	5 Disk Groups	Comments
Availability	↓	↑	↑↑	If more hosts are available in the cluster, the cluster tolerates more failures. This capability adds cost because additional hardware for the disk groups is required.
Manageability	o	o	o	If more hosts are in the cluster, more virtual machines can be managed in the vSAN environment.
Performance	o	↑	↑↑	If the flash percentage ratio to storage capacity is large, the vSAN can deliver increased performance and speed.
Recoverability	o	↑	↑↑	More available disk groups can increase the recoverability of vSAN during a failure. Rebuilds complete faster because there are more places to place data and to copy data from.
Security	o	o	o	Neither design option impacts security.

Legend: ↑ = positive impact on quality; ↓ = negative impact on quality; o = no impact on quality.

Table 7-73. Disk Groups Per Host Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-SDS-008	Configure VMware vSAN with a single disk group per ESXi host in the management cluster.	Single disk group provides the required performance and usable space for the datastore.	Losing an SSD in a host takes the disk group offline. Using two or more disk groups can increase availability and performance.

VMware vSAN Policy Design

After you enable and configure VMware vSAN, you can create storage policies that define the virtual machine storage characteristics. Storage characteristics specify different levels of service for different virtual machines. The default storage policy tolerates a single failure and has a single disk stripe. Use the default unless your environment requires policies with non-default behavior. If you configure a custom policy, VMware vSAN will guarantee it; however, if vSAN cannot guarantee a policy, you cannot provision a virtual machine that uses the policy unless you enable force provisioning.

VMware vSAN Policy Options

A storage policy includes several attributes, which can be used alone or combined to provide different service levels. Policies can be configured for availability and performance conservatively to balance space consumed and recoverability properties. In many cases, the default system policy is adequate and no additional policies are required. Policies allow any configuration to become as customized as needed for the application's business requirements.

Policy Design Background

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

Table 7-74. VMware vSAN Policy Options

Capability	Use Case	Value	Comments
Number of failures to tolerate	Redundancy	Default 1 Max 3	A standard RAID 1 mirrored configuration that provides redundancy for a virtual machine disk. The higher the value, the more failures can be tolerated. For n failures tolerated, n+1 copies of the disk are created, and 2n+1 hosts contributing storage are required. A higher n value indicates that more replicas of virtual machines are made, which can consume more disk space than expected.
Number of disk stripes per object	Performance	Default 1 Max 12	A standard RAID 0 stripe configuration used to increase performance for a virtual machine disk. This setting defines the number of HDDs on which each replica of a storage object is striped. If the value is higher than 1, increased performance can result. However, an increase in system resource usage might also result.
Flash read cache reservation (%)	Performance	Default 0 Max 100%	Flash capacity reserved as read cache for the storage is a percentage of the logical object size that will be reserved for that object. Only use this setting for workloads if you must address read performance issues. The downside of this setting is that other objects cannot use a reserved cache. VMware recommends not using these reservations unless it is absolutely necessary because unreserved flash is shared fairly among all objects.

Table 7-74. VMware vSAN Policy Options (Continued)

Capability	Use Case	Value	Comments
Object space reservation (%)	Thick provisioning	Default 0 Max 100%	The percentage of the storage object that will be thick provisioned upon VM creation. The remainder of the storage will be thin provisioned. This setting is useful if a predictable amount of storage will always be filled by an object, cutting back on repeatable disk growth operations for all but new or non-predictable storage use.
Force provisioning	Override policy	Default: No	Force provisioning allows for provisioning to occur even if the currently available cluster resources cannot satisfy the current policy. Force provisioning is useful in case of a planned expansion of the vSAN cluster, during which provisioning of VMs must continue. VMware vSAN automatically tries to bring the object into compliance as resources become available.

By default, policies are configured based on application requirements. However, they are applied differently depending on the object.

Table 7-75. Object Policy Defaults

Object	Policy	Comments
Virtual machine namespace	Failures-to-Tolerate: 1	Configurable. Changes are not recommended.
Swap	Failures-to-Tolerate: 1	Configurable. Changes are not recommended.
Virtual disk(s)	User-Configured Storage Policy	Can be any storage policy configured on the system.
Virtual disk snapshot(s)	Uses virtual disk policy	Same as virtual disk policy by default. Changes are not recommended.

Note If you do not specify a user-configured policy, the default system policy of 1 failure to tolerate and 1 disk stripe is used for virtual disk(s) and virtual disk snapshot(s). Policy defaults for the VM namespace and swap are set statically and are not configurable to ensure appropriate protection for these critical virtual machine components. Policies must be configured based on the application's business requirements. Policies give VMware vSAN its power because it can adjust how a disk performs on the fly based on the policies configured.

Policy Design Recommendations

Policy design starts with assessment of business needs and application requirements. Use cases for VMware vSAN must be assessed to determine the necessary policies. Start by assessing the following application requirements:

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

After assessment, configure the software-defined storage module policies for availability and performance in a conservative manner so that space consumed and recoverability properties are balanced. In many cases the default system policy is adequate and no additional policies are required unless specific requirements for performance or availability exist.

Table 7-76. Policy Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-SDS-009	Use the default VMware vSAN storage policy.	The default vSAN storage policy provides the level of redundancy that is needed within the management cluster. Additionally, using unique vSAN storage policies are not required as the default policy provides the level of performance that is needed for the individual management components.	Additional policies might be needed if 3rd party VMs are hosted in these clusters because their performance or availability requirements might differ from what the default VMware vSAN policy supports.
SDDC-VI-Storage-SDS-010	Configure the virtual machine swap file as a sparse objects on VMware vSAN	Enabling this setting creates virtual swap files as a sparse object on the vSAN datastore. Sparse virtual swap files will only consume capacity on vSAN as they are accessed. The result can be significantly less space consumed on the vSAN datastore, provided virtual machines do not experience memory over commitment, requiring use of the virtual swap file.	Administative overhead to enable the advanced setting on all ESXi hosts running VMware vSAN.

NFS Storage Design

This NFS design does not give specific vendor or array guidance. Consult your storage vendor for the configuration settings appropriate for your storage array.

NFS Storage Concepts

NFS (Network File System) presents file devices to an ESXi host for mounting over a network. The NFS server or array makes its local file systems available to ESXi hosts. The ESXi hosts access the metadata and files on the NFS array or server using a RPC-based protocol. NFS is implemented using Standard NIC that is accessed using a VMkernel port (vmknic).

NFS Load Balancing

No load balancing is available for NFS/NAS on vSphere because it is based on single session connections. You can configure aggregate bandwidth by creating multiple paths to the NAS array, and by accessing some datastores via one path, and other datastores via another path. You can configure NIC Teaming so that if one interface fails, another can take its place. However these load balancing techniques work only in case of a network failure and might not be able to handle error conditions on the NFS array or on the NFS server. The storage vendor is often the source for correct configuration and configuration maximums.

NFS Versions

vSphere is compatible with both NFS version 3 and version 4.1; however, not all features can be enabled when connecting to storage arrays that use NFS v4.1.

Table 7-77. NFS Version Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-NFS-001	Use NFS v3 for all NFS datastores.	NFS v4.1 datastores are not supported with Storage I/O Control.	NFS v3 does not support Kerberos authentication.

Storage Access

NFS v3 traffic is transmitted in an unencrypted format across the LAN. Therefore, best practice is to use NFS storage on trusted networks only and to isolate the traffic on dedicated VLANs.

Many NFS arrays have some built-in security, which enables them to control the IP addresses that can mount NFS exports. Best practice is to use this feature to determine which ESXi hosts can mount the volumes that are being exported and have read/write access to those volumes. This prevents unapproved hosts from mounting the NFS datastores.

Exports

All NFS exports are shared directories that sit on top of a storage volume. These exports control the access between the endpoints (ESXi hosts) and the underlying storage system. Multiple exports can exist on a single volume, with different access controls on each.

Export Size per Region	Size
vRealize Log Insight Archive	1 TB

Table 7-78. NFS Export Design Decisions

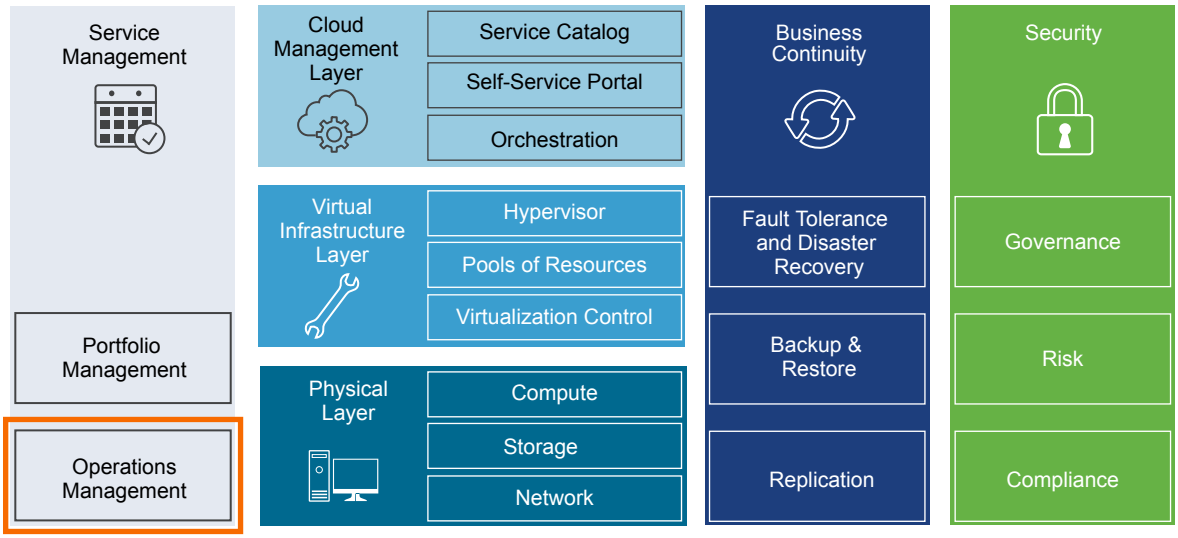
Decision ID	Design Decision	Design Justification	Design Implication
SDDC-VI-Storage-NFS-002	Create 1 export to support the vRealize Log Insight Archive management components.	The storage requirements of these management components are separate from the primary storage.	If you expand the design, you can create additional exports.
SDDC-VI-Storage-NFS-004	For each export, limit access to only the application VMs or hosts requiring the ability to mount the storage.	Limiting access helps ensure the security of the underlying data.	Securing exports individually can introduce operational overhead.

Operations Infrastructure Design

8

Operations Management is a required element of a software-defined data center. Monitoring operations with vRealize Log Insight provide capabilities for performance and capacity management of related infrastructure and cloud management components.

Figure 8-1. Operations Management in the SDDC Layered Architecture



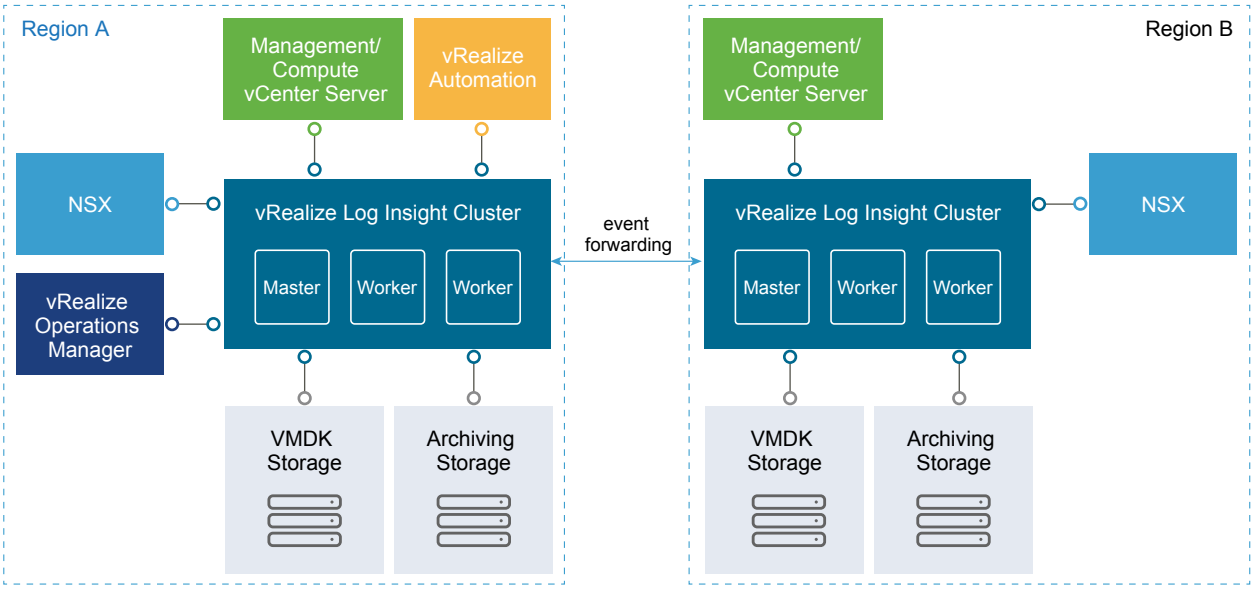
vRealize Log Insight Design

vRealize Log Insight design enables real-time logging for all components that build up the management capabilities of the SDDC in a dual-region setup.

Logical Design

In a multi-region Software Defined Data Center (SDDC) deploy a vRealize Log Insight cluster in each region that consists of three nodes. This configuration allows for continued availability and increased log ingestion rates.

Figure 8-2. Logical Design of vRealize Log Insight



Sources of Log Data

vRealize Log Insight collects logs as to provide monitoring information about the SDDC from a central location.

vRealize Log Insight collects log events from the following virtual infrastructure and cloud management components.

- Management vCenter Server
  - Platform Services Controller
  - vCenter Server
- Compute vCenter Server
  - Platform Services Controller
  - vCenter Server
- Management, shared edge and compute ESXi hosts
- NSX for vSphere for the management cluster and for the shared compute and edge cluster
  - NSX Manager
  - NSX Controller instances
  - NSX Edge instances

Cluster Nodes of vRealize Log Insight

The vRealize Log Insight cluster consists of one master node and two worker nodes. You enable the Integrated Load Balancer (ILB) on the cluster to have vRealize Log Insight to balance incoming traffic fairly among available nodes.

vRealize Log Insight clients, using both the Web user interface, and ingestion through syslog or the Ingestion API, connect to vRealize Log Insight that the ILB addresses.

vRealize Log Insight cluster can scale out to 12 nodes, that is, one master and 11 worker nodes.

Table 8-1. Cluster Node Configuration Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-001	Deploy vRealize Log Insight in a cluster configuration of 3 nodes with an integrated load balancer: one master and two worker nodes.	Provides high availability. Using the integrated load balancer simplifies the Log Insight deployment, and prevents from a single point of failure.	You must size each node identically. If the capacity requirements for your vRealize Log Insight cluster grow, identical capacity must be added to each node.
SDDC-OPS-LOG-002	Apply vSphere Distributed Resource Scheduler (DRS) anti-affinity rules to the vRealize Log Insight cluster components	Using DRS prevents vRealize Log Insight nodes from on the same ESXi host and thereby risking the cluster's high availability capability.	Additional configuration is required to set up anti-affinity rules. Only a single ESXi host in the management cluster, of the four ESXi hosts, will be able to be put into maintenance mode at a time.

Sizing Log Insight Nodes

To accommodate all log data from the products in the SDDC, you must size the compute resources and storage for the Log Insight nodes properly.

By default, the vRealize Log Insight virtual appliance uses the predefined values for small configurations, which have 4 vCPUs, 8 GB of virtual memory, and 510 GB of disk space provisioned. vRealize Log Insight uses 100 GB of the disk space to store raw data, index, metadata, and other information.

Sizing Nodes

Select a size for the vRealize Log Insight nodes to collect and store log data from the SDDC management components and tenant workloads according to the objectives of this design.

Table 8-2. Compute Resources for a vRealize Log Insight Medium-Size Node

Attribute	Specification
Appliance size	Medium
Number of CPUs	8
Memory	16 GB
Disk Capacity	510 GB (490 GB for event storage)
IOPS	1,000 IOPS
Amount of processed log data when using log ingestion	75 GB/day of processing per node
Number of processed log messages	5,000 event/second of processing per node
Environment	Up to 250 syslog connections per node

Sizing Storage

Sizing is based on IT organization requirements, but this design provides calculations according based on a single region implementation, and is implemented on a per-region basis. This sizing is calculated according to the following node configuration per region:

- Management vCenter Server
  - Platform Services Controller
  - vCenter Server
- Compute vCenter Server
  - Platform Services Controller
  - vCenter Server
- Management, shared edge and compute ESXi hosts
- NSX for vSphere for the management cluster and for the shared compute and edge cluster
  - NSX Manager
  - NSX Controller instances
  - NSX Edge instances
- Event forwarding configured between vRealize Log Insight clusters

These components aggregate to approximately 210 syslog and vRealize Log Insight Agent sources. Assuming that you want to retain 7 days of data, use the following calculations:

For 210 syslog sources at a basal rate of 150 MB of logs ingested per-day per-source over 7 days, you need the following storage space:

```
210 sources * 150 MB of log data ≈ 31.5 GB log data per-day

31.5 GB * 7 days ≈ 220.5 GB log data per vRealize Log Insight node

220.5 GB * 1.7 indexing overhead ≈ 375 GB
```

Based on this example, the storage space that is allocated per medium-size vRealize Log Insight virtual appliance is enough to monitor the SDDC.

Consider the following approaches when you must increase the Log Insight capacity:

- If you must maintain a log data retention for more than 7 days in your SDDC, you might add more storage per node by adding a new virtual hard disk. vRealize Log Insight supports virtual hard disks of up to 2 TB. If you must add more than 2 TB to a virtual appliance, add another virtual hard disk.

  When you add storage to increase the retention period, extend the storage for all virtual appliances.

  Note Do not extend existing retention virtual disks. Once provisioned, do not reduce the size or remove virtual disks to avoid data loss.

- If you must monitor more components by using log ingestion and exceed the number of syslog connections or ingestion limits defined in this design, you can deploy more vRealize Log Insight virtual appliances to scale out your environment. vRealize Log Insight can scale up to 12 nodes in an HA cluster.

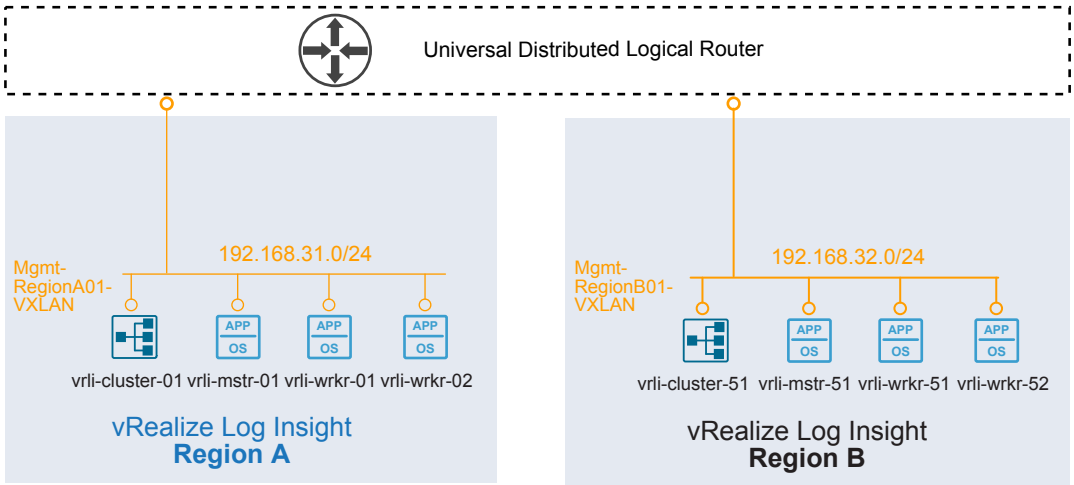
Table 8-3. Compute Resources for the vRealize Log Insight Nodes Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-003	Deploy vRealize Log Insight nodes of medium size.	Accommodates the number of expected Syslog and vRealize Log Insight Agent connections from the following. This is approximately 210 syslog and vRealize Log Insight Agent sources. ■ Management & Compute vCenter Server, Platform Services Controller ■ Management, shared edge and compute ESXi hosts, the ■ Management and compute components for NSX for vSphere ■ Cross-vRealize Log Insight cluster event forwarding. This ensure the storage space for the vRealize Log Insight cluster is sufficient for 7 days of data retention.	You must increase the size of the nodes if you configure Log Insight to monitor additional syslog sources.

vRealize Log Insight Networking Design

A vRealize Log Insight instance is deployed within the shared management application isolated network. When you expand the design to dual region, the vRealize Log Insight instances are connected to the region-specific management VXLANs Mgmt-RegionA01-VXLAN and Mgmt-RegionB01-VXLAN.

Figure 8-3. Networking Design for the vRealize Log Insight Deployment



Application Network Design

This networking design has the following features:

- All nodes have routed access to the vSphere management network through the Management NSX UDLR for the home region.

- Routing to the vSphere management network and the external network is dynamic, and is based on the Border Gateway Protocol (BGP).

Table 8-4. vRealize Log Insight Network Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-004	Deploy vRealize Log Insight on the region-specific application virtual networks.	■ Ensures centralized access to log data per region if a cross-region network outage occurs. ■ Co-located log collection to the region local SDDC applications using the region-specific application virtual networks. ■ Provides a consistent deployment model for management applications.	■ Interruption in the cross-region network can impact event forwarding between the vRealize Log Insight clusters and cause gaps in log data. ■ You must use NSX to support this network configuration.

IP Subnets

You can allocate the following example subnets to the vRealize Log Insight deployment.

Table 8-5. IP Subnets in the Application Isolated Networks

vRealize Log Insight Cluster	IP Subnet
Region A	192.168.31.0/24
Region B	192.168.32.0/24

vRealize Log Insight DNS Names

vRealize Log Insight node name resolution uses a region-specific suffix, such as sfo01.rainpole.local or lax01.rainpole.local, including the load balancer virtual IP addresses (VIPs). The Log Insight components in both regions have the following node names.

Table 8-6. DNS Names of the vRealize Log Insight Nodes

DNS Name	Role	Region
vrli-cluster-01.sfo01.rainpole.local	Log Insight ILB VIP	A
vrli-mstr-01.sfo01.rainpole.local	Master node	A
vrli-wrkr-01.sfo01.rainpole.local	Worker node	A
vrli-wrkr-02.sfo01.rainpole.local	Worker node	A

Table 8-7. DNS Names Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-005	Configure forward and reverse DNS records for all vRealize Log Insight nodes and VIPs.	All nodes are accessible by using fully-qualified domain names instead of by using IP addresses only.	You must manually provide a DNS record for each node and VIP.

vRealize Log Insight Retention and Archiving

Configure archive and retention parameters of vRealize Log Insight according to the company policy for compliance and governance.

vRealize Log Insight virtual appliances contain three default virtual disks and can use addition virtual disks for storage.

Table 8-8. Virtual Disk Configuration in the vRealize Log Insight Virtual Appliance

Hard Disk	Size	Usage
Hard disk 1	20 GB	Root file system
Hard disk 2	510 GB for medium-size deployment	Contains two partitions: ■ `/storage/var` System logs ■ `/storage/core` Storage for Collected logs.
Hard disk 3	512 MB	First boot only

Calculate the storage space that is available for log data using the following equation:

```
/storage/core = hard disk 2 space – system logs space on hard disk 2
```

Based on the size of the default disk, the storage core is equal to 490 GB.

```
/storage/core = 510GB – 20 GB = 490 GB
Retention = /storage/core – 3% * /storage/core
```

If `/storage/core` is 490 GB, vRealize Log Insight can use 475 GB for retention.

```
Retention = 490 GB – 3% * 490 ≈ 475 GB
```

Configure a retention period of 7 days for the medium-size vRealize Log Insight appliance.

Table 8-9. Retention Period Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-007	Configure vRealize Log Insight to retain data for 7 days.	Accommodates logs from 750 syslog sources (250 per node) as per the SDDC design.	None

If you must support a log retention period that is greater than the prescribed 7 days, additional storage may be added to each virtual appliance. When adding storage to increase the retention, the capacity supplied must be identical across all virtual appliances.

Archiving

You configure vRealize Log Insight to archive log data only if your must retain logs for an extended period of time, either for compliance, auditability, and so on.

vRealize Log Insight archives log messages as soon as possible. At the same time, the logs are retained on the virtual appliance until the free local space is almost filled. Data exists on both the vRealize Log Insight appliance and the archive location for most of the retention period. The archiving period must be longer than the retention period.

The archive location must be on an NFS version 3 shared storage. The archive location must be available and must have enough capacity to accommodate the archives.

Apply an archive policy of 90 days for the medium-size vRealize Log Insight appliance. The vRealize Log Insight appliance will use about 1 TB of shared storage. According to the business compliance regulations of your organization, these sizes might change.

Table 8-10. Log Archive Policy Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-008	Provide 1 TB of NFS version 3 shared storage to each vRealize Log Insight cluster.	Archives logs from 750 syslog sources.	■ You must manually maintain the vRealize Log Insight archive blobs stored on the NFS store, selectively pruning as more space is required. ■ You must enforce the archive policy directly on the shared storage. ■ If the NFS mount does not have enough free space or is unavailable for a period greater than the retention period of the virtual appliance, vRealize Log Insight stops ingesting new data until the NFS mount has enough free space, becomes available, or archiving is disabled.

vRealize Log Insight Alerting

vRealize Log Insight supports alerts that trigger notifications about its health.

Alert Types

The following types of alerts exist in vRealize Log Insight:

System Alerts vRealize Log Insight generates notifications when an important system event occurs, for example when the disk space is almost exhausted and vRealize Log Insight must start deleting or archiving old log files.

Content Pack Alerts Content packs contain default alerts that can be configured to send notifications, these alerts are specific to the content pack and are disabled by default.

User-Defined Alerts Administrators and users can define their own alerts based on data ingested by vRealize Log Insight.

vRealize Log Insight handles alerts in two ways:

- Send an e-mail over SMTP.
- Send to vRealize Operations Manager.

SMTP Notification

Enable e-mail notification for alerts in vRealize Log Insight.

Table 8-11. SMTP Alert Notification Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-009	Enable alerting over SMTP.	Enables administrators and operators to receive alerts via email from vRealize Log Insight.	Requires access to an external SMTP server.

Information Security and Access Control in vRealize Log Insight

Protect the vRealize Log Insight deployment by providing centralized role-based authentication and secure communication with the other components in the Software-Defined Data Center (SDDC).

Authentication

Enable role-based access control in vRealize Log Insight by using the existing rainpole.local Active Directory domain.

Table 8-12. Authorization and Authentication Management Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-012	Use Active Directory for authentication.	Provides fine-grained role and privilege-based access for administrator and operator roles.	You must provide access to the Active Directory from all Log Insight nodes.
SDDC-OPS-LOG-013	Configure a service account svc-loginsight on vCenter Server for application-to-application communication from vRealize Log Insight with vSphere.	Provides the following access control features: ■ vRealize Log Insight accesses vSphere with the minimum set of permissions that are required to collect vCenter Server events, tasks and alarms and to configure ESXi hosts for syslog forwarding. ■ In the event of a compromised account, the accessibility in the destination application remains restricted. ■ You can introduce improved accountability in tracking request-response interactions between the components of the SDDC.	You must maintain the service account's life cycle outside of the SDDC stack to ensure its availability.

Table 8-12. Authorization and Authentication Management Design Decisions (Continued)

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-014	Use global permissions when you create the svc-loginsight service account in vCenter Server.	■ Simplifies and standardizes the deployment of the service account across all vCenter Servers in the same vSphere domain. ■ Provides a consistent authorization layer.	All vCenter Server instances must be in the same vSphere domain.
SDDC-OPS-LOG-015	Configure a service account svc-vrli-vrops on vRealize Operations Manager for application-to-application communication from vRealize Log Insight for a two-way launch in context.	Provides the following access control features: ■ vRealize Log Insight and vRealize Operations Manager access each other with the minimum set of required permissions. ■ In the event of a compromised account, the accessibility in the destination application remains restricted. ■ You can introduce improved accountability in tracking request-response interactions between the components of the SDDC.	You must maintain the service account's life cycle outside of the SDDC stack to ensure its availability.

Encryption

Replace default self-signed certificates with a CA-signed certificate to provide secure access to the vRealize Log Insight Web user interface.

Table 8-13. Custom Certificates Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-016	Replace the default self-signed certificates with a CA-signed certificate.	Configuring a CA-signed certificate ensures that all communication to the externally facing Web UI is encrypted.	The administrator must have access to a Public Key Infrastructure (PKI) to acquire certificates.

Configuration for Collecting Logs

As part of vRealize Log Insight configuration, you configure syslog and vRealize Log Insight agents.

Client applications can send logs to vRealize Log Insight in one of the following ways:

- Directly to vRealize Log Insight over the syslog protocol
- By using vRealize Log Insight to directly query the vSphere Web Server APIs
- By using a vRealize Log Insight Agent

Table 8-14. Direct Log Communication to vRealize Log Insight Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-014	Configure syslog sources to send log data directly to vRealize Log Insight.	Simplifies the design implementation for log sources that are syslog capable.	You must configure syslog sources to forward logs to the vRealize Log Insight VIP.
SDDC-OPS-LOG-016	Configure vCenter Server Appliances and Platform Services Controller Appliances as syslog sources to send log data directly to vRealize Log Insight.	Simplifies the design implementation for log sources that are syslog capable.	■ You must manually configure syslog sources to forward logs to the vRealize Log Insight VIP. ■ Certain dashboards within vRealize Log Insight require the use of the vRealize Log Insight Agent for proper ingestion. ■ Not all Operating System-level events are forwarded to vRealize Log Insight.
SDDC-OPS-LOG-017	Configure vRealize Log Insight to ingest events, tasks, and alarms from the Management and Compute vCenter Server instances .	Ensures that all tasks, events and alarms generated across all vCenter Server instances in a specific region of the SDDC are captured and analyzed for the administrator.	You must create a service account on vCenter Server to connect vRealize Log Insight for events, tasks, and alarms pulling. This does not capture Events that occur on the Platform Services Controller.
SDDC-OPS-LOG-018	Do not configure vRealize Log Insight to automatically update all Agents deployed.	Manually update the Log Insight Agents on each of the specified components within the SDDC .	You must maintain manually the vRealize Log Insight agents on each of the SDDC components.

Time Synchronization

Time synchronization is critical for the core functionality of vRealize Log Insight. By default, vRealize Log Insight synchronizes time with a predefined list of public NTP servers.

NTP Configuration

Configure consistent NTP sources on all systems that send log data (vCenter Server, ESXi). See *Time Synchronization* in the *Planning and Preparation* document.

Table 8-15. Time Synchronization Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-019	Configure consistent NTP sources on all virtual infrastructure and cloud management applications for correct log analysis in vRealize Log Insight.	Guarantees accurate log timestamps.	Requires that all applications synchronize time to the same NTP time source.

Cluster Communication

All vRealize Log Insight cluster nodes must be in the same LAN with no firewall or NAT between the nodes.

External Communication

vRealize Log Insight receives log data over the syslog TCP, syslog TLS/SSL, or syslog UDP protocols. Use the default syslog UDP protocol because security is already designed at the level of the management network.

Table 8-16. Syslog Protocol Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
SDDC-OPS-LOG-020	Communicate with the syslog clients, such as ESXi, vCenter Server, NSX for vSphere, on the default UDP syslog port.	Using the default syslog port simplifies configuration for all syslog sources.	■ If the network connection is interrupted, the syslog traffic is lost. ■ UDP syslog traffic is not secure.