

Telco Cloud Infrastructure – Cloud Director Edition Reference Architecture Guide 2.0

VMware Telco Cloud Infrastructure Cloud Director Edition
2.0

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

<https://docs.vmware.com/>

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

Copyright © 2021 VMware, Inc. All rights reserved. [Copyright and trademark information.](#)

Contents

- 1** About the Telco Cloud Infrastructure Cloud Director Edition Reference Architecture Guide 5
- 2** Introduction to Telco Cloud Infrastructure Cloud Director Edition 7
- 3** Overview of the Telco Cloud Infrastructure Cloud Director Edition Reference Architecture 9
 - Physical Tier 9
 - Workload Domain Overview 9
 - Network Overview 11
 - Storage Overview 15
 - Platform Tier 15
 - Virtual Infrastructure Overview 15
 - Cloud Automation Tier 19
 - Cloud Automation Overview 19
 - Operations Management Tier 21
 - Operations Management Overview 21
- 4** Telco Cloud Infrastructure Cloud Director Edition Solution Design 23
 - Physical Design 23
 - Physical ESXi Host Design 23
 - Physical Network Design 25
 - Physical Storage Design 27
 - Platform Design 28
 - vCenter Server Design 29
 - Workload Domains and Clusters Design 30
 - Network Virtualization Design 35
 - NSX Design 40
 - Shared Storage Design 46
 - Cloud Automation Design 47
 - VMware Cloud Director Cells 47
 - VMware Cloud Director Leases 50
 - VMware Cloud Director Tenancy and Resource Isolation 50
 - VMware Cloud Director Networking 56
 - VMware Cloud Director Storage 58
 - VMware Cloud Director Roles and Authentication 59
 - Operations Management Design 61
 - vRealize Log Insight Design 61

[vRealize Operations Manager Design](#) 64

[vRealize Network Insight Design](#) 66

About the Telco Cloud Infrastructure Cloud Director Edition Reference Architecture Guide

1

This reference architecture guide provides guidance for designing and creating a greenfield Network Functions Virtualization (NFV) platform by using VMware Telco Cloud Infrastructure™ – Cloud Director Edition.

Intended Audience

This guide is intended for telecommunications and solution architects, sales engineers, field consultants, advanced services specialists, and customers who are responsible for the Virtualized Network Services (VNFs) and the NFV environment in which the VNFs run.

Acronyms and Definitions

Table 1-1. General Acronyms

Abbreviation	Description
BFD	Bidirectional Forwarding Detection, for failure detection on the transport links.
DPDK	Data Plane Development Kit, an Intel-led packet processing acceleration technology.
MTTR	Mean-Time-To-Repair
MTTU	Mean-Time-To-Understand

Table 1-2. Telco Cloud Infrastructure Acronyms

Abbreviation	Description
CCP	Centralized Control Plane in the VMware NSX-T™ Data Center architecture.
CNF	Cloud Native Network Function, executing within a Kubernetes environment.
LCP	Local Control Plane in the NSX-T Data Center architecture.
MANO	Management and Orchestration components, a term originating from the ETSI NFV architecture framework.
NFV	Network Functions Virtualization

Table 1-2. Telco Cloud Infrastructure Acronyms (continued)

Abbreviation	Description
NFVI	Network Functions Virtualization Infrastructure
NFVO	Network Functions Virtualization Orchestrator
N-VDS (E)	Enhanced mode when using the NSX-T Data Center N-VDS logical switch. This mode enables DPDK for workload acceleration.
N-VDS (S)	Standard mode when using the NSX-T Data Center N-VDS logical switch.
VIM	Virtualized Infrastructure Manager
VNF	Virtual Network Function, executing in a Virtual Machine (VM).
VNFM	Virtual Network Function Manager

Table 1-3. Telco Acronyms

Abbreviation	Description
HSS	Home Subscriber Server in the mobile evolved packet core 4G architecture.
MVNO	Mobile Virtual Network Operator
PCRF	Policy and Charging Rules Function, in the mobile evolved packet core 4G architecture.
PGW	Packet Gateway in the mobile evolved packet core 4G architecture.
SGW	Service Gateway in the mobile evolved packet core 4G architecture.
SBC	Session Border Controller used in voice telephone for control and data plane communications between clients.
VPP	Vector Packet Processing

Introduction to Telco Cloud Infrastructure Cloud Director Edition

2

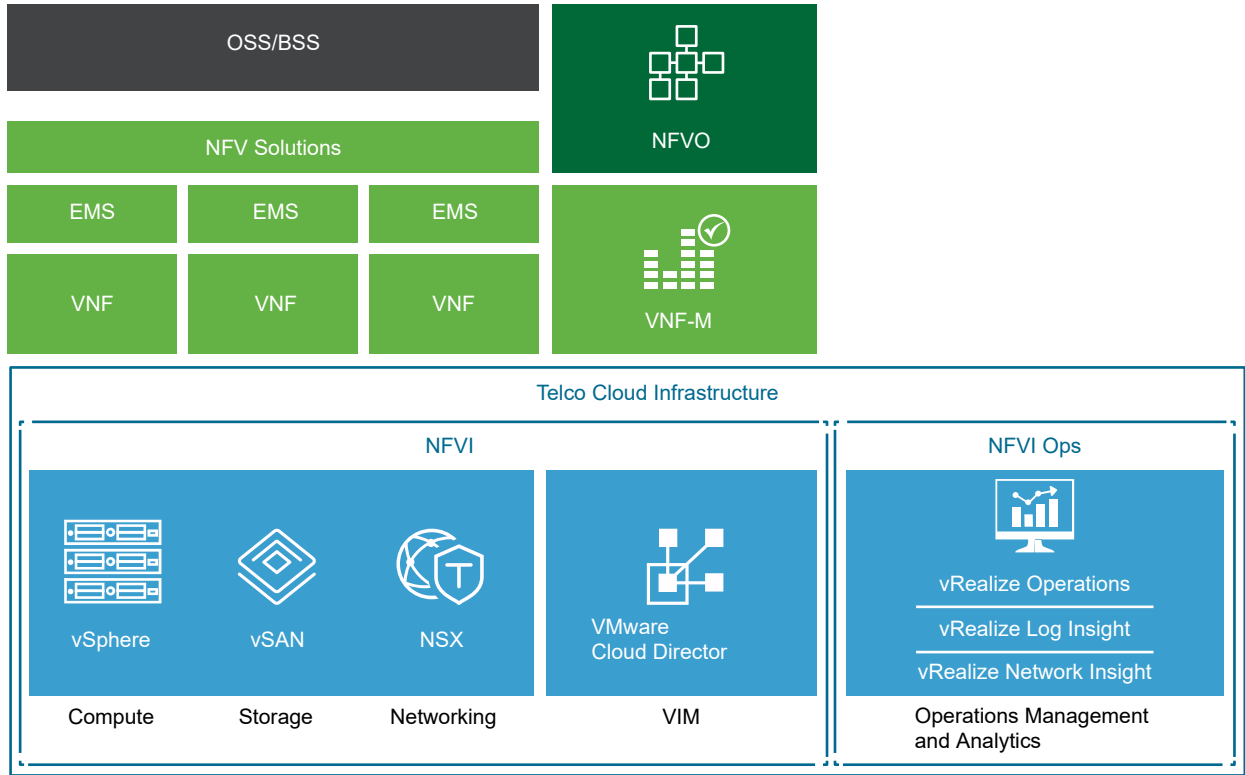
VMware Telco Cloud Infrastructure™ – Cloud Director Edition combines a carrier-grade infrastructure with VMware Cloud Director™ as the Virtualized Infrastructure Manager (VIM).

The Telco Cloud Infrastructure Cloud Director Edition platform combines VMware Cloud Director with stable and supportable infrastructure. This way, Telco Cloud Infrastructure provides a platform to support Communication Service Providers (CSPs) with network modernization and business transformation.

The Telco Cloud Infrastructure platform implements a modular design with abstractions that enable multi-vendor, multi-domain, and hybrid physical, and virtual execution environments. The IaaS layer that is exposed through VMware Cloud Director, provides a CI/CD environment for workload life cycle management. The platform also delivers an automation framework to interoperate with external functions for service orchestration and management.

In addition to the core infrastructure components for compute, storage, networking, and VIM, the Telco Cloud Infrastructure platform includes a fully integrated suite for operational intelligence and monitoring. This suite can be used to further enhance the runtime environments with workflows for dynamic workload optimization and proactive issue avoidance.

Figure 2-1. Telco Cloud Infrastructure Components



Overview of the Telco Cloud Infrastructure Cloud Director Edition Reference Architecture

3

This Architecture Overview section describes the high level physical and virtual infrastructure, networking, and storage elements in this reference architecture.

This chapter includes the following topics:

- [Physical Tier](#)
- [Platform Tier](#)
- [Cloud Automation Tier](#)
- [Operations Management Tier](#)

Physical Tier

The architecture of the physical layers must support modularity of the infrastructure for compute, networking, and storage elements.

Workload Domain Overview

A workload domain consists of VMware ESXi hosts managed by a single VMware vCenter Server instance, storage component for workload data, and network equipment to connect to the data center.

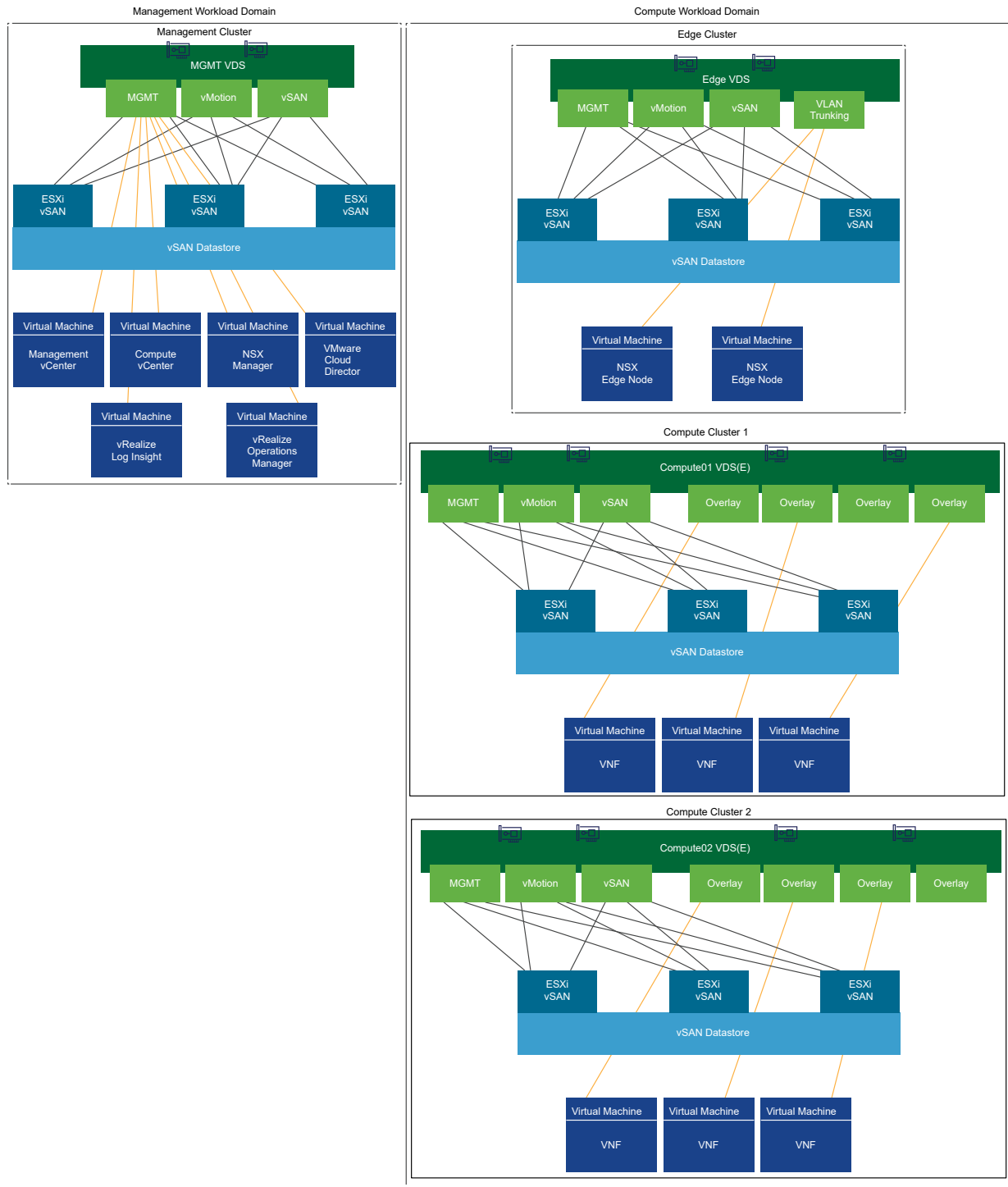
Workload Domain Characteristics

Workload domains can include different combinations of ESXi hosts and network equipment that can be set up with varying levels of hardware redundancy and varying quality of components. Workload domains must be reachable through a Layer 3 network.

A workload domain represents a logical boundary of functionality, managed by a single vCenter Server instance. The workload domain is not defined by any physical properties such as physical location.

For the deployment of any size, consider homogeneity and easy replication. Different workload domains can provide different characteristics for varying requirements. For example, one workload domain can use full hardware redundancy for each component for increased availability. Another workload domain in the same deployment can use hardware without any hardware redundancy. These variations make the architecture suitable for different workload requirements.

Figure 3-1. Workload Domains



Network Overview

You can implement the switch fabric at the physical layer by providing Layer 2 or Layer 3 transport services. For a scalable and vendor-neutral data solution, use Layer 3 transport.

Both Layer 2 and Layer 3 transport have their own sets of benefits and drawbacks. When deciding on an architecture, consider the following benefits and drawbacks of each transport.

Layer 2 Transport Considerations

The following considerations apply for a design that uses Layer 2 transport:

- Top-of-Rack (ToR) switches and upstream Layer 3 devices such as core switches and routers form a switched fabric.
- The upstream Layer 3 devices terminate each VLAN and provide the default gateway functionality.
- Uplinks from the ToR switch to the upstream Layer 3 devices are 802.1Q trunks carrying all required VLANs.

Figure 3-2. Layer 2 Transport

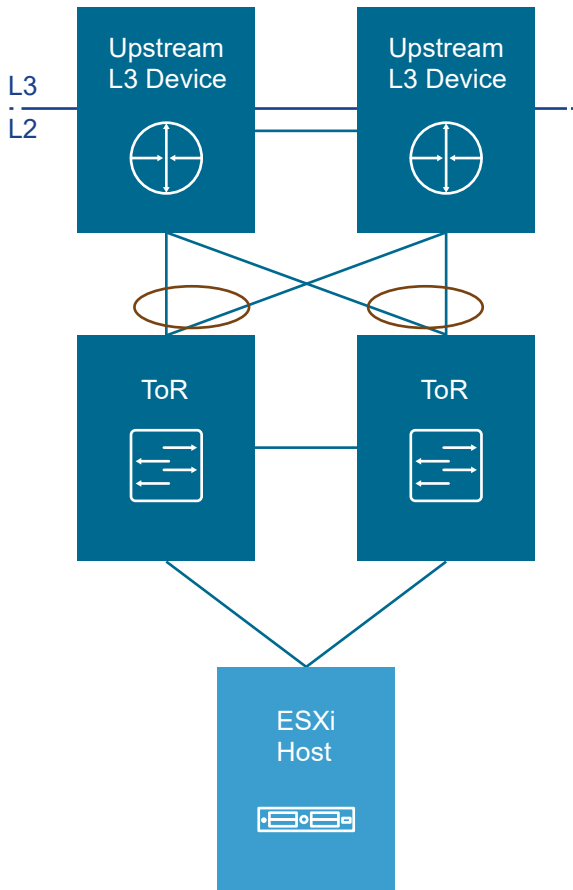


Table 3-1. Benefits and Drawbacks of a Layer 2 Transport

Characteristic	Description
Benefits	<ul style="list-style-type: none">■ Additional design flexibility■ You can span VLANs across racks.
Drawbacks	<ul style="list-style-type: none">■ The size of this deployment is limited because the fabric elements share a limited number of VLANs.■ You might have to rely on a specialized switching fabric product from a single vendor.

Layer 3 Transport Considerations

The following considerations apply for a design using Layer 3 transport:

- Layer 2 connectivity is limited to the ToR switches.
- The ToR switch terminates each VLAN and provides the default gateway functionality. That is, it has a Switch Virtual Interface (SVI) for each VLAN.
- Uplinks from the ToR switch to the upstream layer are routed point-to-point links. You cannot use VLAN trunking on the uplinks.
- A dynamic routing protocol, such as eBGP, connects the ToR switches and upstream switches. Each ToR switch advertises the prefixes, typically one per VLAN or subnet. In turn, the ToR switch calculates equal-cost paths to the prefixes received from the upstream layer it peers with.

Figure 3-3. Layer 3 Transport

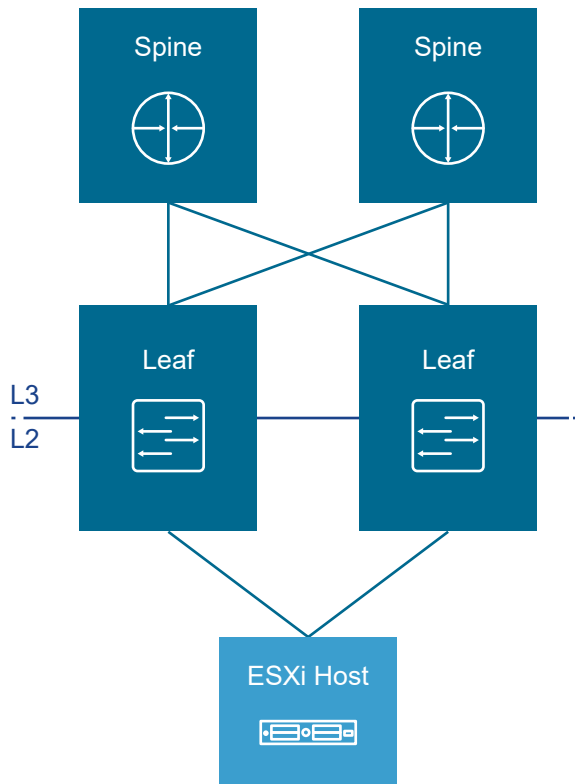


Table 3-2. Benefits and Drawbacks of a Layer 3 Transport

Characteristic	Description
Benefits	<ul style="list-style-type: none"> ■ You can select from many Layer 3 capable switch products for the physical switching fabric. ■ You can mix switches from different vendors because of the general interoperability between the implementation of routing protocols. ■ This approach is cost-effective because it uses only the basic functionality of the physical switches.
Drawbacks	<ul style="list-style-type: none"> ■ VLANs are restricted to a single rack. The restriction can affect IP-based storage networks, such as iSCSI and NFS.

Physical Network Interfaces

The Telco Cloud Infrastructure requires that the ESXi hosts contain four or more physical NICs of the same speed. Switch ports supporting the ESXi hosts must have the required VLANs tagged.

The VMware vSphere Distributed Switch supports several NIC teaming options. Load-based NIC teaming ensures the optimal use of available bandwidth and redundancy in case of a link failure. Use a minimum of four physical network interfaces for compute and two for management and edge clusters. Use physical network interfaces with at least 25 GbE throughput in combination with a pair of ToR switches. Depending on the switch vendor, 802.1Q network trunks can support a maximum of 4095 or 4096 VLANs. Link aggregation, such as LACP, must not be used between the ESXi hosts and the physical switches.

Storage Overview

VMware vSAN is used in both management and compute workload domains to provide highly available shared storage to the clusters.

vSAN uses the local disks, all-flash devices, or flash and standard magnetic hard drives in each ESXi host to create a highly available shared datastore for a vSphere cluster. vSAN reduces storage costs, especially in remote or edge locations where a dedicated array is impractical.

When using vSAN, you can use all-flash or a combination of flash and magnetic hard drives, and compatible I/O controllers including the firmware level of each component from the [VMware vSAN hardware compatibility list](#) or use vSAN Ready Nodes.

vSAN Ready Nodes

VMware vSAN Ready Node is a validated server configuration in a certified hardware form factor for vSAN deployment, jointly recommended by the server OEM and VMware. For examples about standardized configurations from different vSAN partners, see the [vSAN Ready Node documentation](#).

Platform Tier

The virtual infrastructure is the foundation of the platform. It contains the software-defined infrastructure, software-defined networking, and software-defined storage.

In the virtual infrastructure layer, access to the underlying physical infrastructure is controlled and allocated to the management and customer workloads. The platform layer consists of hypervisors on the physical hosts and the hypervisor management components. The management components consist of elements in the virtual management layer, network and storage layers, business continuity, and security areas.

Virtual Infrastructure Overview

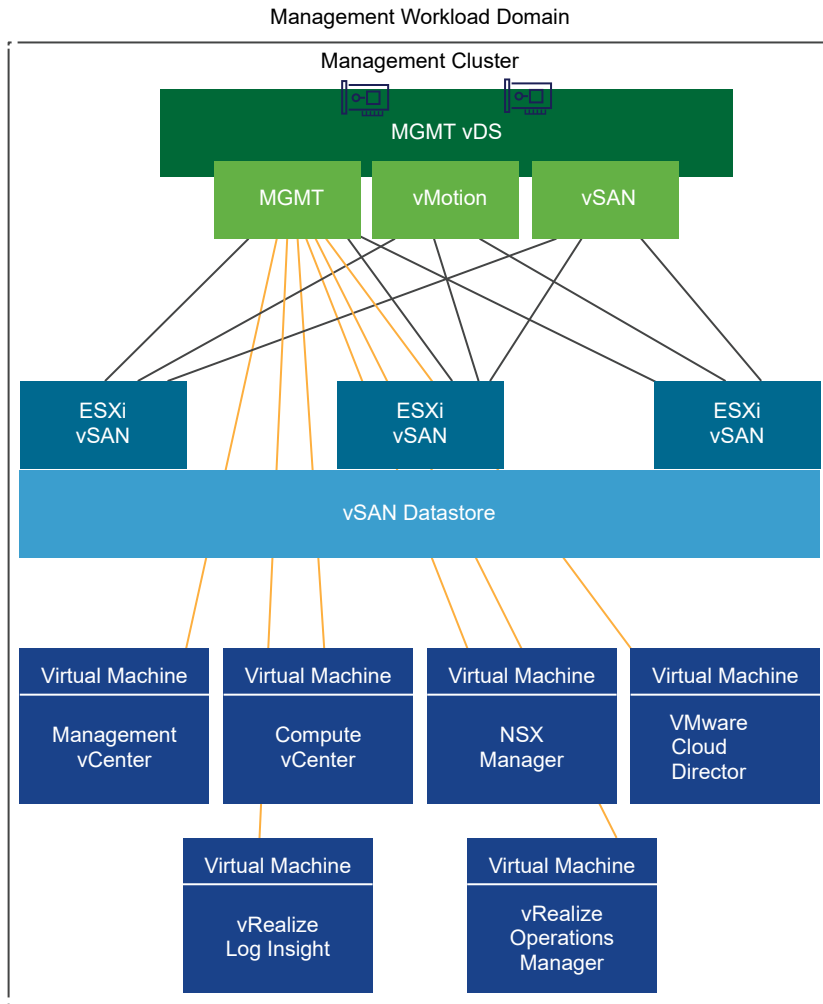
This architecture consists of a centralized management workload domain along with workload domains to support the required workloads.

Management Workload Domain

The management workload domain contains a single vSphere cluster called the management cluster. The management cluster hosts the VMs that manage the solution. This cluster is crucial for the management and monitoring of the solution. Its high availability deployment ensures that the management and monitoring services are always available.

Table 3-3. Management Workload Domain Components

Component	Description
Management vCenter Server	Manages the Management Workload Domain
Compute vCenter Server	Manages the Compute Workload Domain
NSX Manager	Operational and central control plane for software-defined networking, implemented as a cluster of three VMs
VMware Cloud Director	VMware VIM
vRealize Suite Standard	Includes vRealize Log Insight and vRealize Operations Manager
vRealize Network Insight	Communicates with the vCenter Server and NSX Manager instances to collect metrics that are presented through various dashboards and views

Figure 3-4. Management Workload Domain

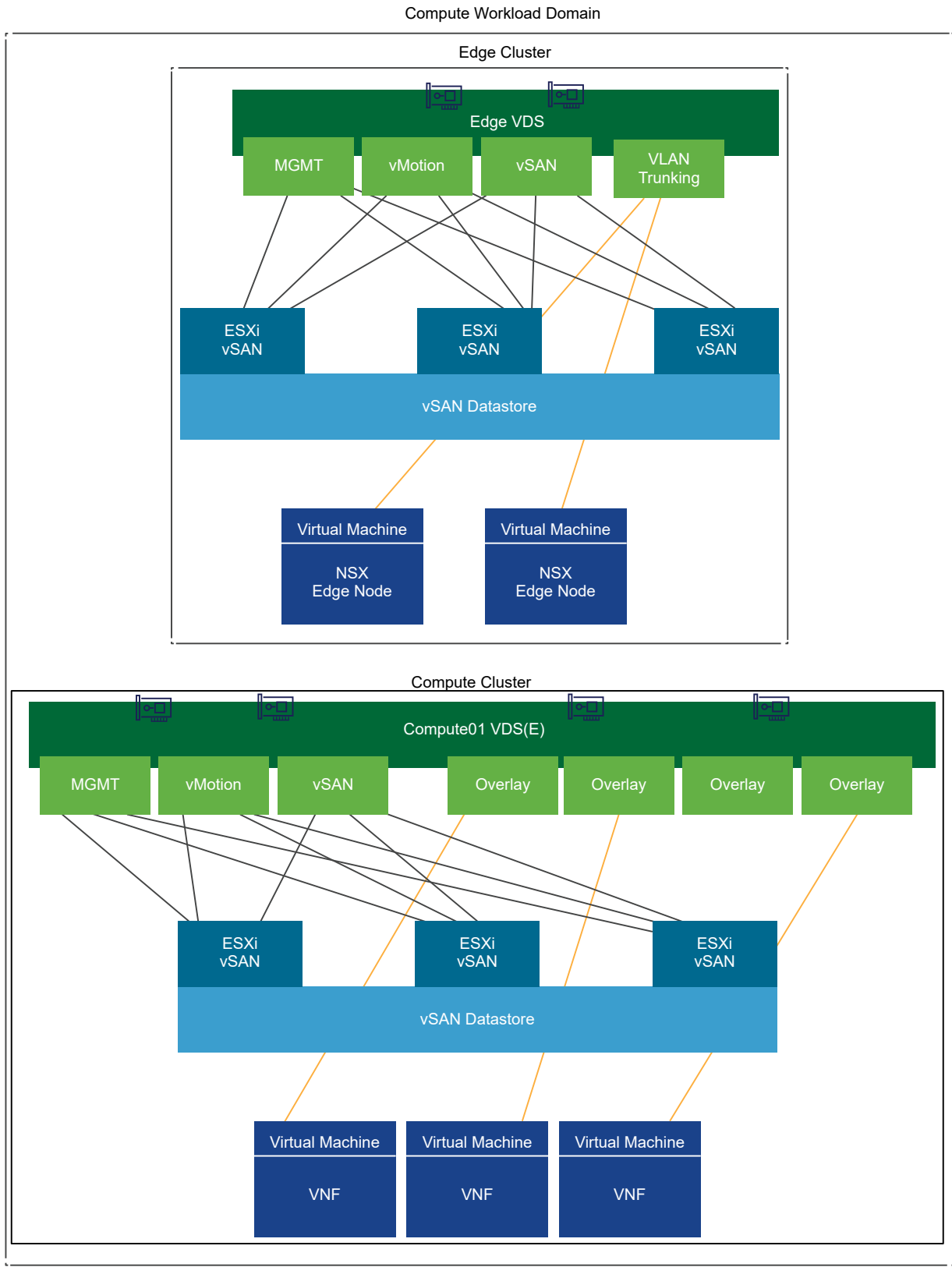
Compute Workload Domains

The compute workload domain can contain multiple vSphere clusters. These clusters can contain a minimum of two ESXi hosts and a maximum of 96 or 64 hosts (when using vSAN), depending on the resource and availability requirements of the solution being deployed.

The Edge cluster, which hosts NSX Edge VMs, is part of the compute workload domain. This cluster provides virtualized network services such as load balancers and the north/south routing infrastructure to support the workloads.

Each compute workload domain can support a maximum of 2000 ESXi hosts and 25,000 VMs. If you use other management and monitoring tools, the vCenter maximums are not applicable and the actual number of ESXi hosts and VMs per workload domain is fewer.

Figure 3-5. Compute Workload Domain



Cloud Automation Tier

The Cloud Automation Tier contains various NFVI abstractions for compute, storage, and networking, and contains the resource orchestration component known as the Virtual Infrastructure Manager (VIM).

The Cloud Automation Tier offers service management and control functions that bridge the virtual resource orchestration and physical functions to deliver services and service chains. It is a centralized control and management function, including embedded automation and optimization capabilities.

Cloud Automation Overview

VMware Cloud Director is a component that Telco Cloud Infrastructure Cloud Director Edition exposes as the interface to the VNF services. It uses vCenter Server and NSX Manager to orchestrate compute, storage, and networking from a single programmable interface.

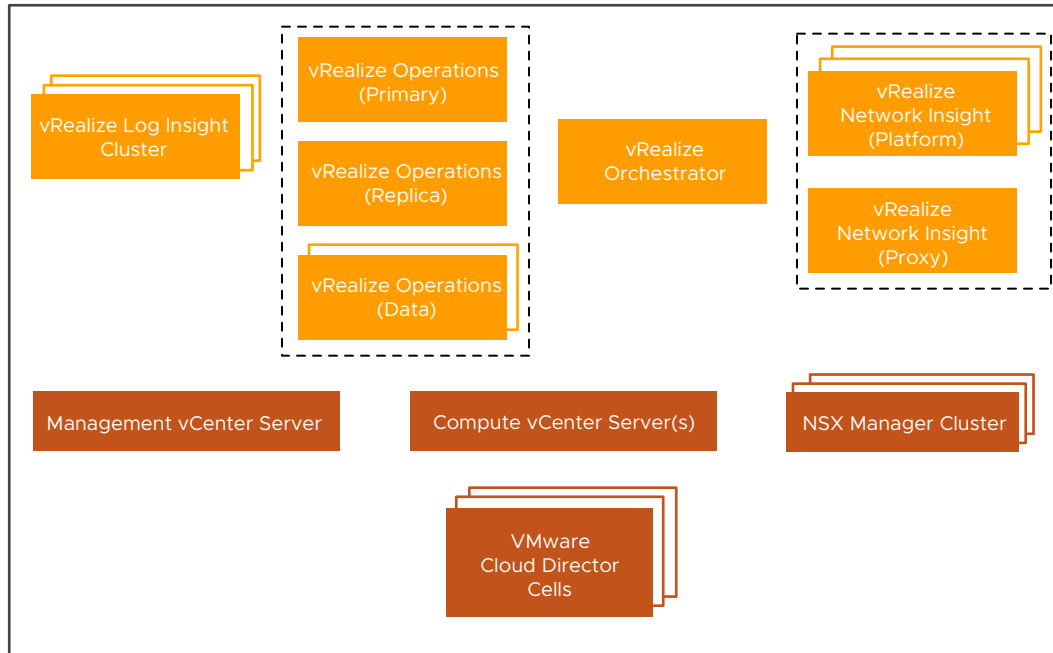
VMware Cloud Director Overview

VMware Cloud Director is used for the cloud automation plane. It natively supports, pools, and further abstracts the virtualization platform in terms of virtual data centers. It provides multitenancy features and self-service access for tenants through a native graphical user interface or API that allows programmable access for both tenant consumption and the provider for cloud management.

The cloud architecture consists of management components that are deployed in the management cluster, and of resource groups for hosting the tenant workloads. Some of the reasons for the separation between the management and compute resources include:

- Different SLAs such as availability, recoverability, and performance for management components and tenant workloads
- Separation of responsibilities. For example, resource groups are managed by VMware Cloud Director.
- Consistent management and scaling of resource groups

The management cluster runs the cloud management components and resource group management components. Resource groups are independent infrastructure domains represented by virtualized compute, networking, and storage, each managed by its vCenter Server.

Figure 3-6. Management Cluster Components

Some of the most commonly-used terms in VMware Cloud Director are as follows:

Allocation Pool

A pool of allocated resources for which a certain percentage of compute resources is guaranteed.

Catalog

A repository of vApp templates and media available to users for deployment. Catalogs can be published and shared between organizations in the same VMware Cloud Director environment.

External Network

External network provides Internet connectivity to organization networks and is configured for Internet accessibility.

Network Pool

A collection of isolated Layer 2 virtual networks available to VMware Cloud Director for the automated deployment of organization and vApp networks.

Organization

The unit of multitenancy representing a single logical security boundary. An organization contains users, virtual data centers, and catalogs.

Organization Administrator

Administrator for a VMware Cloud Director organization responsible for managing provided resources, network services, users, and vApp policies.

Organization VDC Network

Organization VDC networks are instantiated through network pools and bound to a single organization VDC or shared across an Organization. Organization VDC networks can be isolated, routed, or directly connected to an external network.

Organization Virtual Data Center

A subgrouping of compute, network, and storage resources allocated from a provider virtual data center and assigned to a single organization. A virtual data center is a deployment environment where vApps can be instantiated, deployed, and powered on. Organization virtual data centers cannot span multiple organizations.

Provider Virtual Data Center

A grouping of compute and storage resources from a single VMware vCenter Server. A provider virtual data center consists of one or more resource pools and one or more datastores. Provider virtual data center resources can be shared with multiple organizations.

vApp

vApp is a container for software solution in the cloud. A vApp is the standard unit of deployment for VMware Cloud Director. It contains one or more VMs, networks, network services, power-on operations. vApps can be imported or exported.

Cloud Director Cell

A cell is the runtime of VMware Cloud Director services on a physical or virtual machine. Multiple cells within one VMware Cloud Director instance can be grouped and connected to one VMware Cloud Director database for load balancing and high availability.

Operations Management Tier

The components of the operations management layer support centralized data monitoring and logging for the solutions in this design.

The physical infrastructure, virtual infrastructure, and workloads are monitored in real-time in the operations management layer, collecting information for intelligent and dynamic operational management.

Operations Management Overview

Operations management includes vRealize Network Insight, vRealize Log Insight, and vRealize Operations Manager to provide real-time monitoring and logging for the infrastructure and compute workloads.

vRealize Log Insight

vRealize Log Insight collects data from ESXi hosts using the syslog protocol. vRealize Log Insight has the following capabilities:

- Connects to other VMware products such as vCenter Server to collect events, tasks, and alarm data.
- Integrates with vRealize Operations Manager to send notification events and enable launch in context.
- Functions as a collection and analysis point for any system that sends syslog data.

vRealize Operations Manager

vRealize Operations Manager tracks and analyzes the operation of multiple data sources by using specialized analytic algorithms. These algorithms help vRealize Operations Manager learn and predict the behavior of every object it monitors. Users access this information by using views, reports, and dashboards.

vRealize Network Insight

vRealize Network Insight collects and analyzes information about the operation of network data sources such as NSX-T Data Center and vCenter Server. Administrators access this information by using dashboards.

Telco Cloud Infrastructure Cloud Director Edition Solution Design

4

The Telco Cloud Infrastructure Cloud Director Edition solution design considers both physical and virtual infrastructure design elements.

This chapter includes the following topics:

- [Physical Design](#)
- [Platform Design](#)
- [Cloud Automation Design](#)
- [Operations Management Design](#)

Physical Design

The physical design includes the physical ESXi hosts, storage, and network design.

Physical ESXi Host Design

You must ensure the physical specifications of the ESXi hosts allow for successful deployment and operation of the design.

Physical Design Specification Fundamentals

The configuration and assembly process for each system is standardized, with all components installed in the same manner on each ESXi host. Because the standardization of the physical configuration of the ESXi hosts removes variability, you can operate an easily manageable and supportable infrastructure. Deploy ESXi hosts with identical configurations across all cluster members, including storage and networking configurations. For example, consistent PCI card slot placement, especially for network controllers, is essential for accurate alignment of physical to virtual I/O resources. By using identical configurations, you can balance the VM storage components across storage and compute resources.

The sizing of physical servers that run ESXi requires special considerations when you use vSAN storage. The design provides details on using vSAN as the primary storage system for the management cluster and the compute cluster. This design also uses vSAN ReadyNodes.

ESXi Host Memory

The amount of memory required for compute clusters varies according to the workloads running in the cluster. When sizing the memory of hosts in the compute cluster, consider the admission control setting (n+1) that reserves the resources of a host for failover or maintenance. In addition, leave at least 8% of the resources for ESXi host operations.

The number of vSAN disk groups and disks managed by an ESXi host determines the memory requirements. To support the maximum number of disk groups, 32 GB of RAM is required for vSAN.

Table 4-1. Recommended Physical ESXi Host Design

Design Recommendation	Design Justification	Design Implication
Use vSAN ReadyNodes.	vSAN ReadyNodes ensure full compatibility with vSAN.	Hardware choices might be limited.
Ensure that all ESXi hosts have a uniform configuration across a cluster.	A balanced cluster has more predictable performance even during hardware failures. In addition, the impact on performance during re-sync or rebuild is minimal if the cluster is balanced.	As new models of servers become available, the deployed model phases out. So, it becomes difficult to keep a uniform cluster when adding hosts later.
Set up the management cluster with a minimum of four ESXi hosts.	Allocating 4 ESXi hosts provides full redundancy for the management cluster.	Additional ESXi host resources are required for redundancy.
Set up each ESXi host in the management cluster with a minimum of 256 GB RAM.	<ul style="list-style-type: none"> Ensures that the management components have enough memory to run during a single host failure. Provides a buffer for future management or monitoring of components in the management cluster. 	<ul style="list-style-type: none"> In a four-node cluster, only the resources of three ESXi hosts are available as the resources of one host are reserved for vSphere HA. Depending on the products deployed and their configuration, more memory per host might be required.
Set up each edge cluster with a minimum of three ESXi hosts.	Ensures full redundancy for the required 2 NSX Edge nodes.	As Edge Nodes are added, more hosts must be added to the cluster to ensure redundancy.
Set up each ESXi host in the edge cluster with a minimum of 192 GB RAM.	Ensures that the NSX Edge Nodes have the required memory.	In a three-node cluster, only the resources of two ESXi hosts are available as the resources of one host are reserved for vSphere HA.
Set up each compute cluster with a minimum of four ESXi hosts.	Allocating 4 ESXi hosts provides full redundancy for the compute clusters.	Additional ESXi host resources are required for redundancy.
Set up each ESXi host in the compute cluster with a minimum of 384 GB RAM.	<ul style="list-style-type: none"> A good starting point for most workloads. Allows for ESXi and vSAN overhead. Increase the RAM size based on vendor recommendations. 	In a four-node cluster, the resources of only three ESXi hosts are available as the resources of one host are reserved for vSphere HA.

Physical Network Design

The physical network design includes defining the network topology for connecting physical switches and the ESXi hosts, determining switch port settings for VLANs, and designing routing.

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 support 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 on a Cisco physical switch.
- Configure jumbo frames on all switch ports, Inter-Switch Link (ISL), and Switched Virtual Interfaces (SVIs).

Top-of-Rack Connectivity and Network Settings

Each ESXi host is connected redundantly to the network fabric ToR switches by using a minimum of four physical Interfaces. Configure the ToR switches to provide all necessary VLANs through an 802.1Q trunk. These redundant connections use the features of vSphere Distributed Switch to guarantee that a physical interface is not overrun and redundant paths are used if they are available.

- **Spanning Tree Protocol (STP):** Although this design does not use the STP, switches usually include STP configured by default. Designate the ports connected to ESXi hosts as trunk PortFast.
- **Trunking:** Configure the switch ports as members of an 802.1Q trunk.
- **MTU:** Set MTU for all switch ports, VLANs, and SVIs to jumbo frames for consistency.

Jumbo Frames

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

The workload determines whether to configure jumbo frames on a VM. If the workload consistently transfers large amounts of network data, configure jumbo frames, if possible. Also, ensure that both the VM operating system and the VM NICs support jumbo frames. Jumbo frames also improve the performance of vSphere vMotion.

Table 4-2. Recommended Physical Network Design

Design Recommendation	Design Justification	Design Implication
Use a Layer 3 transport	<ul style="list-style-type: none"> ■ You can select Layer 3 switches from different vendors for the physical switching fabric. ■ You can mix switches from different vendors because of the general interoperability between the implementation of routing protocols. ■ This approach is cost-effective because it uses only the basic functionality of the physical switches. 	VLANs are restricted to a single rack.
Implement the following physical network architecture: <ul style="list-style-type: none"> ■ Two physical interfaces, one per NUMA node, on each ToR switch for ESXi host uplinks. ■ No EtherChannel (LAG/vPC/MLAG) configuration for ESXi host uplinks. ■ Layer 3 device with BGP support. 	<ul style="list-style-type: none"> ■ Guarantees availability during a switch failure. ■ Provides compatibility with vSphere host profiles because they do not store link-aggregation settings. ■ Supports BGP as the dynamic routing protocol. ■ BGP is the only dynamic routing protocol supported by NSX-T Data Center. 	Hardware choices might be limited.
Use two ToR switches for each rack.	<ul style="list-style-type: none"> ■ This design uses four physical interfaces, two per NUMA node, on each ESXi host. ■ Provides redundancy and reduces the overall design complexity. 	Two ToR switches per rack can increase costs.
Use VLANs to segment physical network functions.	<ul style="list-style-type: none"> ■ Supports physical network connectivity without requiring many NICs. ■ Isolates different network functions of the Software-Defined Data Center (SDDC) so that you can have differentiated services and prioritized traffic as needed. 	Requires uniform configuration and presentation on all the switch ports made available to the ESXi hosts.
Assign static IP addresses to all management components.	Ensures that interfaces such as management and storage always have the same IP address. In this way, you provide support for continuous management of ESXi hosts using vCenter Server and for provisioning IP storage by storage administrators	Requires precise IP address management.
Create DNS records for all ESXi hosts and management VMs to enable forward, reverse, short, and FQDN resolution.	Ensures consistent resolution of management components using both IP address (reverse lookup) and name resolution.	Adds administrative overhead.

Table 4-2. Recommended Physical Network Design (continued)

Design Recommendation	Design Justification	Design Implication
Use an NTP, or Precision Time Protocol time source for all management components.	It is critical to maintain accurate and synchronized time between management components.	None
Configure the MTU size to at least 9000 bytes (jumbo frames) on the physical switch ports, VLANs, SVIs, vSphere Distributed Switches, and VMkernel ports.	<ul style="list-style-type: none"> ■ Improves traffic throughput. ■ Required for NSX-T Data Center. 	When you adjust the MTU size, you must also configure the entire network path (VMkernel port, distributed switch, physical switches, and routers) to support the same MTU size.

Physical Storage Design

This design uses VMware vSAN to implement software-defined storage as the primary storage type.

vSAN is a hyper-converged storage solution that is integrated with the ESXi hypervisor. vSAN creates disk groups consisting of hard disk drives and flash devices or all-flash devices in the local ESXi host, and presents a highly resilient shared storage datastore to the vSphere Cluster. By using vSAN storage policies, you can control capacity, performance, and availability on a per virtual disk basis.

Table 4-3. vSAN Requirements

Category	Requirements
Number of ESXi hosts	Minimum of 3 ESXi hosts providing storage resources to the cluster. This can be 3 ESXi hosts or 2 ESXi hosts and 1 vSAN witness.
vSAN configuration	vSAN can be configured in all-flash or hybrid mode. <ul style="list-style-type: none"> ■ All-flash vSAN configuration requires flash devices for both the caching and capacity tiers. ■ vSAN hybrid storage configuration requires both the magnetic devices for capacity and the flash devices for caching.
Requirements for individual ESXi hosts that provide storage resources.	<ul style="list-style-type: none"> ■ Minimum of one flash device. The flash cache tier must be at least 10% of the size of the capacity tier. ■ Minimum of two HDDs for hybrid mode, or an additional flash device for an all-flash configuration. ■ RAID controller that is compatible with vSAN with a separate controller for the boot device. ■ Minimum 10 Gbps network for vSAN traffic. ■ Host isolation response of vSphere High Availability is set to power off VMs.

I/O Controllers

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

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

Controller queue depth is an important aspect of performance. All I/O controllers in the [VMware vSAN Hardware Compatibility Guide](#) have a minimum queue depth of 256. If you increase the queue depth to a value higher than 256, ensure that you consider the regular day-to-day operations in your environment. Examples of events that require higher queue depth are as follows:

- VM deployment operations
- Re-sync I/O activity as a result of automatic or manual fault remediation.

Table 4-4. Recommended Physical Storage Design

Design Recommendation	Design Justification	Design Implication
Use all-flash vSAN in all clusters.	<ul style="list-style-type: none"> ■ Provides best performance with low latency. ■ When using all-flash vSAN, you can enable de-duplication and compression that saves space on the datastores. 	Flash storage may cost more than traditional magnetic disks.
For the management cluster, provide a vSAN configuration with at least 6 TB of usable space.	Provides all the required space for this solution while allowing the deployment of additional monitoring and management components in the management cluster.	On day 1, more space is required.
For the edge cluster, provide a vSAN configuration with at least 500 GB of usable space.	Provides required storage to run NSX Edge Nodes.	None
For the compute clusters, size the vSAN datastore according to the current workloads plus 5 years of expected growth.	Ensures that the storage solution is not required to be upgraded that can cause downtime to workloads.	On day 1, more space is required.

Note This design uses vSAN. Any supported storage solution can be used as long as it meets the characteristics of this storage design. For best practices, see the vendor documentation.

Platform Design

The platform design includes software components for providing software-defined storage, networking, and compute. These components include the hypervisor, virtualization management, storage virtualization, and network virtualization.

vCenter Server Design

The vCenter Server design includes the design for all the vCenter Server instances. For this design, determine the number of instances, their sizes, networking configuration, cluster layout, redundancy, and security configuration.

A vCenter Server deployment can consist of two or more vCenter Server instances according to the scale, number of VMs, and continuity requirements for your environment.

Protecting the vCenter Server system is important because it is the central point of management and monitoring. You can protect vCenter Server according to the maximum downtime tolerated. The following methods are available to protect a vCenter Server instance:

- Automated protection using vSphere HA
- Automated protection using vCenter Server HA

vCenter Server Sizing: You can size the resources and storage for the Management vCenter Server Appliance and the Compute vCenter Server Appliance according to the expected number of VMs in the environment.

Table 4-5. Recommended Sizing for Management vCenter Servers

Attribute	Specification
Appliance Size	Small (up to 100 hosts or 1000 VMs)
Number of vCPUs	4
Memory	19 GB
Disk Space	528 GB

Table 4-6. Recommended Sizing for Compute vCenter Servers

Attribute	Specification
Appliance Size	Large (up to 1,000 hosts or 10,000 VMs)
Number of vCPUs	16
Memory	37 GB
Disk Space	1,113 GB

TLS Certificates in vCenter Server

By default, vSphere uses TLS/SSL certificates that are signed by VMware Certificate Authority (VMCA). These certificates are not trusted by end-user devices or browsers.

As a security best practice, replace at least all user-facing certificates with certificates that are signed by a third-party or enterprise Certificate Authority (CA).

Table 4-7. Recommended vCenter Server Design

Design Recommendation	Design Justification	Design Implication
Deploy two vCenter Servers. One to support the management workloads and another to support the compute workloads.	<ul style="list-style-type: none"> ■ Isolates vCenter Server failures to the management or compute workloads. ■ Isolates vCenter Server operations between the management and compute workloads. ■ Supports a scalable cluster design where you might reuse the management components as more compute workload domains are added. ■ Simplifies capacity planning for compute workloads because you do not consider management workloads for the Compute vCenter Server. ■ Improves the ability to upgrade the vSphere environment and related components by the separation of maintenance windows. ■ Supports separation of roles and responsibilities to ensure that only administrators with proper authorization can attend to the management workloads. ■ Facilitates quicker troubleshooting and problem resolution. 	Requires licenses for each vCenter Server instance.
Protect all vCenter Servers by using vSphere HA.	Supports the availability objectives for the vCenter Servers without manual intervention during a failure event.	vCenter Server becomes unavailable during the vSphere HA failover.
Replace the vCenter Server machine certificate with a certificate signed by a third-party Public Key Infrastructure.	<ul style="list-style-type: none"> ■ Infrastructure administrators connect to the vCenter Server instances using a Web browser to perform configuration, management, and troubleshooting. ■ The default certificate results in certificate warning messages. 	Replacing and managing certificates is an operational overhead.
Use an SHA-2 or higher algorithm when signing certificates.	The SHA-1 algorithm is considered less secure and is deprecated.	Not all certificate authorities support SHA-2.

Workload Domains and Clusters Design

The vCenter Server functionality is distributed across a minimum of two workload domains and two clusters. This solution uses a minimum of two vCenter Server instances: one for the management workload domain and another for the first compute workload domain. The compute workload domain can contain multiple vSphere clusters.

The cluster design must consider the workloads that the cluster handles. Different cluster types in this design have different characteristics. When you design the cluster layout in vSphere, consider the following guidelines:

- Use fewer large-sized ESXi hosts or more small-sized ESXi hosts.
 - A scale-up cluster has fewer large-sized ESXi hosts.
 - A scale-out cluster has more small-sized ESXi hosts.
- Compare the capital costs of purchasing fewer, larger ESXi hosts with the costs of purchasing more, smaller ESXi hosts. Costs vary between vendors and models.
- Evaluate the operational costs for managing a few ESXi hosts with the costs of managing more ESXi hosts.
- Consider the purpose of the cluster.
- Consider the total number of ESXi hosts and cluster limits.

vSphere High Availability

VMware vSphere High Availability (vSphere HA) protects your VMs in case of an ESXi host failure by restarting VMs on other hosts in the cluster. During the cluster configuration, the ESXi hosts elect a primary ESXi host. The primary ESXi host communicates with the vCenter Server system and monitors the VMs and secondary ESXi hosts in the cluster.

The primary ESXi host detects different types of failure:

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

The vSphere HA Admission Control Policy allows an administrator to configure how the cluster determines available resources. In a small vSphere HA cluster, a large proportion of the cluster resources is reserved to accommodate ESXi host failures, based on the selected policy.

The following policies are available:

- **Cluster resource percentage:** Reserves a specific percentage of cluster CPU and memory resources for recovery from host failures. With this type of admission control, vSphere HA ensures that a specified percentage of aggregate CPU and memory resources is reserved for failover.
- **Slot policy:** vSphere HA admission control ensures that a specified number of hosts can fail and sufficient resources remain in the cluster to failover all the VMs from those hosts.
 - A slot is a logical representation of memory and CPU resources. By default, the slot is sized to meet the requirements for any powered-on VM in the cluster.
 - vSphere HA determines the current failover capacity in the cluster and leaves enough slots for the powered-on VMs. The failover capacity specifies the number of hosts that can fail.
- **Dedicated failover hosts:** When a host fails, vSphere HA attempts to restart its VMs on any of the specified failover hosts.

vSphere Distributed Resource Scheduler

The distribution and usage of CPU and memory resources for all hosts and VMs in the cluster are monitored continuously. The vSphere Distributed Resource Scheduler (DRS) compares these metrics to an ideal resource usage based on the attributes of the cluster's resource pools and VMs, the current demand, and the imbalance target. DRS then provides recommendations or performs VM migrations accordingly.

DRS supports the following modes of operation:

- Manual
 - Initial placement: Recommended host is displayed.
 - Migration: Recommendation is displayed.
- Partially Automated
 - Initial placement: Automatic
 - Migration: Recommendation is displayed.
- Fully Automated
 - Initial placement: Automatic
 - Migration: Recommendation is run automatically.

Resource Pools

A resource pool is a logical abstraction for flexible management of resources. Resource pools can be grouped into hierarchies and used to partition the available CPU and memory resources hierarchically.

Each DRS cluster has an invisible root resource pool that groups the resources of that cluster. The root resource pool does not appear because the resources of the cluster and the root resource pool are always the same.

Users can create child resource pools of the root resource pool or of any user-created child resource pool. Each child resource pool owns some of the parent's resources and can, in turn, have a hierarchy of child resource pools to represent successively smaller units of computational capability.

A resource pool can contain child resource pools, VMs, or both. You can create a hierarchy of shared resources. The resource pools at a higher level are called parent resource pools. Resource pools and VMs that are at the same level are called siblings. The cluster itself represents the root resource pool. If you do not create child resource pools, only the root resource pools exist.

Scalable Shares allow the resource pool shares to dynamically scale as the VMs are added or removed from the resource pool hierarchy.

vSphere Cluster Services

vSphere Cluster Services (vCLS) is enabled by default and runs in all vSphere clusters. vCLS ensures that if vCenter Server becomes unavailable, cluster services remain available to maintain the resources and health of the workloads that run in the clusters.

vSphere DRS is a critical feature of vSphere to maintain the health of the workloads running in the vSphere Cluster. DRS depends on the availability of vCLS VMs.

vCLS VMs are always powered-on because vSphere DRS depends on the availability of these VMs. These VMs must be treated as system VMs. No operations are blocked on vCLS VMs. However, any disruptive operation can result in failure of vSphere DRS. To avoid the failure of cluster services, avoid performing any configuration or operations on the vCLS VMs.

vSphere Lifecycle Manager

vSphere Lifecycle Manager allows for the management of software and firmware lifecycle of the ESXi hosts in a cluster with a single image. The vSphere Lifecycle Manager image is a new functionality that provides a simplified and unified workflow for patching and upgrading ESXi hosts. You can also use these images for bootstrapping and firmware updates.

An image defines the exact software stack to run on all ESXi hosts in a cluster. When you set up an image, you select an ESXi version and a vendor add-on from the vSphere Lifecycle Manager depot. If no ESXi base images and vendor add-ons are available in the depot, you must populate the depot with software updates by synchronizing the depot or uploading updates to the depot manually.

Table 4-8. Recommended vSphere Cluster Design

Design Recommendation	Design Justification	Design Implication
Create a single management cluster that contains all the management ESXi hosts.	<ul style="list-style-type: none"> ■ Simplifies configuration by isolating management workloads from compute workloads. ■ Ensures that the 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.
Create a single edge cluster per compute workload domain.	Supports running NSX Edge nodes in a dedicated cluster.	Requires an additional vSphere cluster.
Create at least one compute cluster. This cluster contains compute workloads.	<ul style="list-style-type: none"> ■ The clusters can be placed close to end-users where the workloads run. ■ 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 the operational overhead.
Create a management cluster with a minimum of four ESXi hosts.	Allocating 4 ESXi hosts provides full redundancy for the cluster.	Additional ESXi host resources are required for redundancy.
Create an edge cluster with a minimum of three ESXi hosts.	Supports availability for a minimum of two NSX Edge Nodes.	As Edge Nodes are added, additional ESXi hosts must be added to the cluster to maintain availability.
Create a compute cluster with a minimum of four ESXi hosts.	Allocating four ESXi hosts provides full redundancy for the cluster.	Additional ESXi host resources are required for redundancy.
Use vSphere HA to protect all VMs against failures.	Provides a robust level of protection for VM availability.	You must provide sufficient resources on the remaining hosts so that VMs can be migrated to those hosts in the event of a host outage.
Set the Host Isolation Response of vSphere HA to Power Off and Restart VMs.	vSAN requires that the HA Isolation Response is set to Power Off so that the VMs can be restarted on the available ESXi hosts.	VMs are powered off in case of a false positive and an ESXi host is declared isolated incorrectly.
Enable vSphere DRS in the management cluster and set it to Fully Automated, with the default setting (medium).	Provides the best trade-off between load balancing and excessive migration with vSphere vMotion events.	If a vCenter Server outage occurs, mapping from VMs to ESXi hosts might be more difficult to determine.
Enable vSphere DRS in the edge and compute clusters and set it to Partially Automated mode.	<ul style="list-style-type: none"> ■ Enables automatic initial placement. ■ Ensures that the latency-sensitive VMs do not move between ESXi hosts automatically. 	Increases the administrative overhead in ensuring that the cluster is properly balanced.

Table 4-8. Recommended vSphere Cluster Design (continued)

Design Recommendation	Design Justification	Design Implication
When creating resource pools, enable Scalable Shares.	Scalable shares ensure that the shares available in each resource pool is dynamic based on the number and priority of VMs in the resource pool instead of a static value.	None
Use vSphere Lifecycle Manager images to ensure that all hosts in a cluster contain the same software versions.	Images allow for a single ESXi image plus vendor add-on to be assigned to the cluster, ensuring that each ESXi host is running the same ESXi version and vendor add-ons.	Workload Management is not compatible with vSphere Lifecycle Manager Images.

Network Virtualization Design

This network virtualization design uses the vSphere Distributed Switch (VDS) and associated features.

Design Goals

The following high-level design goals apply regardless of your environment:

- **Meet diverse needs:** The network must meet the diverse requirements of 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 required. For example, configuring two 25 GbE NICs might be more cost-effective than configuring four 10 GbE NICs.
- **Improve performance:** You can achieve performance improvement and decrease the maintenance time by providing sufficient bandwidth, which in turn reduces the contention and latency.
- **Improve availability:** A well-designed network improves availability, by providing network redundancy.
- **Support security:** A well-designed network supports an acceptable level of security through controlled access and isolation, where required.
- **Enhance infrastructure functionality:** You can configure the network to support vSphere features such as vSphere vMotion, vSphere High Availability, and vSphere Fault Tolerance.

Network Best Practices

Follow these networking best practices throughout your environment:

- Separate the network services to achieve high security and better performance.

- Use Network I/O Control and traffic shaping to guarantee bandwidth to critical VMs. During the network contention, these critical VMs receive a high percentage of the bandwidth.
- Separate the network services on a vSphere Distributed Switch by attaching them to port groups with different VLAN IDs.
- Keep vSphere vMotion traffic on a separate network. When a migration using vSphere vMotion occurs, the contents of the memory of the guest operating system is transmitted over the network. You can place vSphere vMotion on a separate network by using a dedicated vSphere vMotion VLAN.
- Ensure that physical network adapters that are connected to the same vSphere Standard or Distributed Switch are also connected to the same physical network.

Network Segmentation and VLANs

Separate the different types of traffic for access security and to reduce the contention and latency.

High latency on any network can negatively affect the performance. Some components are more sensitive to high latency than others. For example, high latency IP storage and the vSphere Fault Tolerance logging network can negatively affect the performance of multiple VMs.

According to the application or service, high latency on specific VM networks can also negatively affect performance. Determine which workloads and networks are sensitive to high latency by using the information gathered from the current state analysis and by interviewing key stakeholders and SMEs.

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

vSphere Distributed Switch

Create a single virtual switch per cluster. For each type of network traffic, configure a port group to simplify the configuration and monitoring.

When using NSX-T, allocate four physical NICs to the distributed switch. Use two physical NICs, one per NUMA Node, for vSphere Distributed Switch port groups, and the other two physical NICs to NSX-T segments.

Figure 4-1. Management and Edge vSphere Distributed Switch

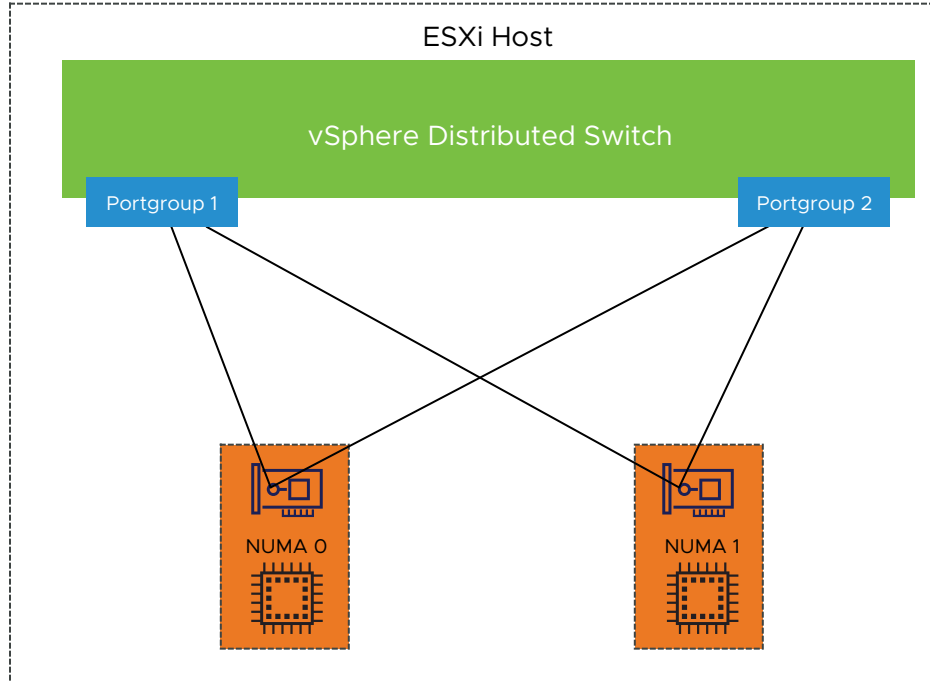
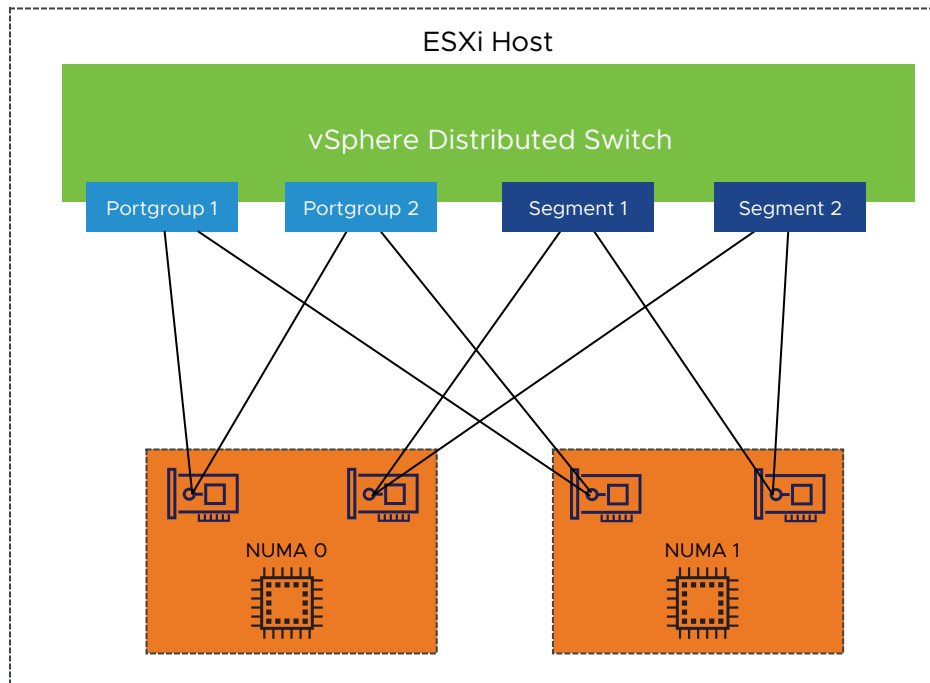


Figure 4-2. Compute vSphere Distributed Switch



NIC Teaming

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

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

Health Check

Health Check helps identify and troubleshoot configuration errors in vSphere distributed switches. The common configuration errors are as follows:

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

Health Check also monitors VLAN, MTU, and teaming policies.

Network I/O Control

When Network I/O Control is enabled, the distributed switch allocates bandwidth for the traffic that is related to the main vSphere features.

When network contention occurs, Network I/O Control enforces the share value specified for different traffic types. As a result, less important traffic, as defined by the share percentage, is throttled, granting access to more network resources to more important traffic types.

Network I/O Control supports bandwidth reservation for system traffic based on the capacity of physical adapters on an ESXi host. It also enables fine-grained resource control at the VM network adapter. Resource control is similar to the CPU and memory reservation model in vSphere DRS.

TCP/IP Stack

Use the vMotion TCP/IP stack to isolate the traffic for vSphere vMotion and to assign a dedicated default gateway for the vSphere vMotion traffic.

By using a separate TCP/IP stack, you can manage vSphere vMotion and cold migration traffic according to the network topology and as required by your organization.

- Route the traffic for the migration of VMs (powered on or off) by using a default gateway. The default gateway is different from the gateway assigned to the default stack on the ESXi host.
- Assign a separate set of buffers and sockets.
- Avoid the routing table conflicts that might otherwise appear when many features are using a common TCP/IP stack.
- Isolate the traffic to improve security.

SR-IOV

SR-IOV is a specification that allows a single Peripheral Component Interconnect Express (PCIe) physical device under a single root port to appear as multiple separate physical devices to the hypervisor or the guest operating system.

SR-IOV uses Physical Functions (PFs) and Virtual Functions (VFs) to manage global functions for the SR-IOV devices. PFs are full PCIe functions that can configure and manage the SR-IOV functionality. VFs are lightweight PCIe functions that support data flow but have a restricted set of configuration resources. The number of VFs provided to the hypervisor or the guest operating system depends on the device. SR-IOV enabled PCIe devices require appropriate BIOS, hardware, and SR-IOV support in the guest operating system driver or hypervisor instance.

In vSphere, a VM can use an SR-IOV virtual function for networking. The VM and the physical adapter exchange data directly without using the VMkernel stack as an intermediary. Bypassing the VMkernel for networking reduces the latency and improves the CPU efficiency for high data transfer performance.

Table 4-9. Recommended Network Virtualization Design

Design Recommendation	Design Justification	Design Implication
Use two physical NICs in the management and edge clusters.	Provides redundancy to all portgroups.	None
Use a minimum of four physical NICs in the compute clusters.	Provides redundancy to all port groups and segments.	None
Use vSphere Distributed Switches.	Simplifies the management of the virtual network.	Migration from a standard switch to a distributed switch requires a minimum of two physical NICs to maintain redundancy.
Use a single vSphere Distributed Switch per cluster.	Reduces the complexity of the network design.	Increases the number of vSphere Distributed Switches that must be managed.
Use ephemeral port binding for the management port group.	Provides the recovery option for the vCenter Server instance that manages the distributed switch.	Port-level permissions and controls are lost across power cycles, and no historical context is saved.
Use static port binding for all non-management port groups.	Ensures that a VM connects to the same port on the vSphere Distributed Switch. This allows for historical data and port-level monitoring.	None
Enable health check on all vSphere distributed switches.	Verifies that all VLANs are trunked to all ESXi hosts attached to the vSphere Distributed Switch and the MTU sizes match the physical network.	You must have a minimum of two physical uplinks to use this feature.
Use the Route based on the physical NIC load teaming algorithm for all port groups.	<ul style="list-style-type: none"> ■ Reduces the complexity of the network design. ■ Increases the resiliency and performance. 	None

Table 4-9. Recommended Network Virtualization Design (continued)

Design Recommendation	Design Justification	Design Implication
Enable Network I/O Control on all distributed switches.	Increases the resiliency and performance of the network.	If configured incorrectly, Network I/O Control might impact the network performance for critical traffic types.
Set the share value to Low for non-critical traffic types such as vMotion and any unused IP storage traffic types such as NFS and iSCSI.	During the network contention, these traffic types are not as important as the VM or vSAN traffic.	During the network contention, vMotion takes longer than usual to complete.
Set the share value for management traffic to Normal .	<ul style="list-style-type: none"> ■ By keeping the default setting to Normal, management traffic is prioritized higher than vSphere vMotion but lower than vSAN traffic. ■ Management traffic ensures that the hosts can be managed during the network contention. 	None
Set the share value to High for the VM and vSAN traffic.	<ul style="list-style-type: none"> ■ VMs are the most important asset in the environment. Leaving the default setting as High ensures that the VMs always have access to the network resources they need. ■ During the network contention, vSAN traffic needs a guaranteed bandwidth to support VM performance. 	None
Use the vMotion TCP/IP stack for vSphere vMotion traffic.	By using the vMotion TCP/IP stack, vSphere vMotion traffic can be assigned a default gateway on its own subnet and can go over Layer 3 networks.	None

NSX Design

By using NSX-T Data Center, virtualization delivers for networking what it has already delivered for compute and storage.

The server virtualization programmatically creates, takes snapshots of, deletes, and restores software-based VMs. Similarly, the NSX-based network virtualization programmatically creates, takes snapshots of, deletes, and restores software-based virtual networks. As a result, you follow a simplified operational model for the underlying physical network.

NSX-T Data Center is a non-disruptive solution. You can deploy it on any IP network such as traditional networking models and next-generation fabric architectures, regardless of the vendor. This is accomplished by decoupling the virtual networks from their physical counterparts.

NSX Manager

NSX Manager is the centralized network management component of NSX-T Data Center. It implements the management and control plane for the NSX infrastructure.

NSX Manager provides the following:

- The Graphical User Interface (GUI) and the RESTful API for creating, configuring, and monitoring NSX-T components, such as segments and gateways.
- An aggregated system view.
- A method for monitoring and troubleshooting workloads attached to virtual networks.
- Configuration and orchestration of the following services:
 - Logical networking components, such as logical switching and routing
 - Networking and edge services
 - Security services and distributed firewall
- A RESTful API endpoint to automate consumption. Because of this architecture, you can automate all configuration and monitoring operations using any cloud management platform, security vendor platform, or automation framework.

Some of the components of the NSX Manager are as follows:

- **NSX Management Plane Agent (MPA):** Available on each ESXi host. The MPA persists the desired state of the system and communicates Non-Flow-Controlling (NFC) messages such as configuration, statistics, status, and real-time data between transport nodes and the management plane.
- **NSX Controller:** Controls the virtual networks and overlay transport tunnels. The controllers are responsible for the programmatic deployment of virtual networks across the entire NSX-T architecture.
- **Central Control Plane (CCP):** Logically separated from all data plane traffic. A failure in the control plane does not affect existing data plane operations. The controller provides configuration to other NSX Controller components such as the segments, gateways, and edge VM configuration.

Virtual Distributed Switch

NSX in vSphere 7 and newer environments can use the vSphere Distributed Switch which makes the management of switches simpler. When an ESXi host is prepared for NSX-T, new vSphere Installable Bundles (VIBs) are installed on the host to enable this functionality. The vSphere Distributed Switch provides the underlying forwarding service that each segment relies on. To implement network virtualization, a network controller must configure the ESXi host virtual switch with network flow tables. The network flow tables form the logical broadcast domains the tenant administrators define when they create and configure segments.

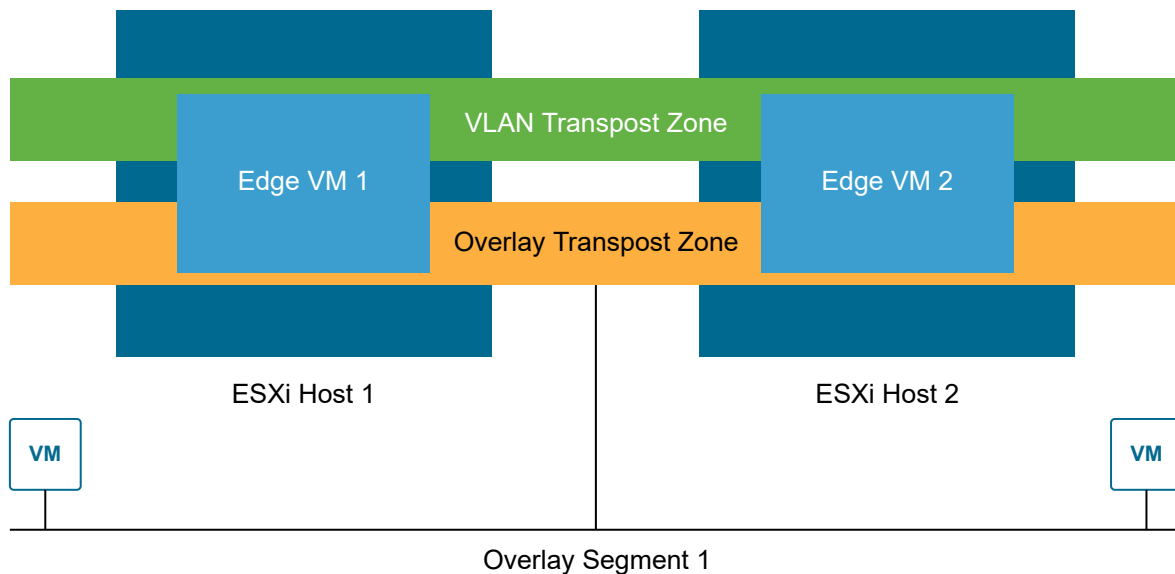
NSX-T Data Center implements each logical broadcast domain by tunneling VM-to-VM traffic and VM-to-gateway traffic using the Geneve tunnel encapsulation mechanism. The network controller has a global view of the data center and ensures that the virtual switch flow tables in the ESXi host are updated as the VMs are created, moved, or removed.

NSX-T Data Center implements virtual switching in Standard and Enhanced modes. The enhanced data path provides superior network performance for telco workloads. ENS mode supports both overlay and VLAN traffic.

Transport Zones

Transport zones determine which hosts can use a particular network. A transport zone identifies the type of traffic, such as VLAN or overlay. You can configure one or more transport zones. A transport zone does not represent a security boundary.

Figure 4-3. Transport Zones



Logical Switching

NSX Segments create logically abstracted segments to which the workloads can be connected. A single NSX Segment is mapped to a unique Geneve segment ID that is distributed across the ESXi hosts in a transport zone. NSX Segments support switching in the ESXi host without the constraints of VLAN sprawl or spanning tree issues.

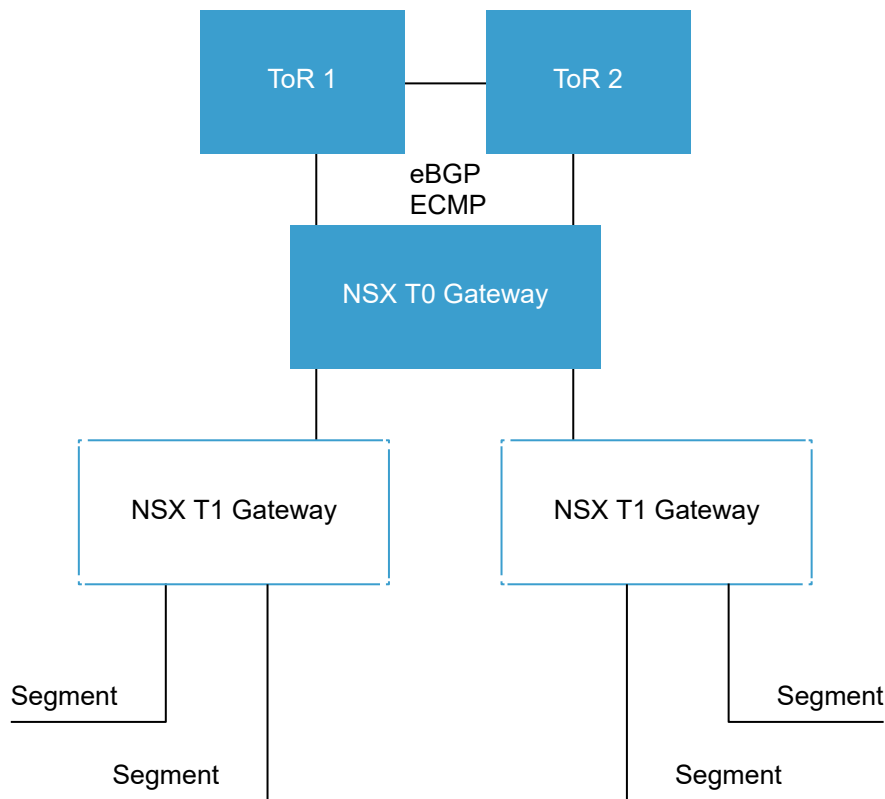
Gateways

NSX Gateways provide the North-South connectivity for the workloads to access external networks and East-West connectivity between different logical networks.

A gateway is a configured partition of a traditional network hardware router. It replicates the functionality of the hardware, creating multiple routing domains in a single router. Gateways perform a subset of the tasks that are handled by the physical router. Each gateway can contain multiple routing instances and routing tables. Using gateways can be an effective way to maximize router use.

- **Distributed Router:** A Distributed Router (DR) spans ESXi hosts whose VMs are connected to this gateway, and edge nodes the gateway is bound to. Functionally, the DR is responsible for one-hop distributed routing between segments and other gateways connected to this Gateway.
- **Service Router:** A Service Router (SR) delivers services such as stateful Network Address Translation (NAT) that are not currently implemented in a distributed fashion. A Gateway always has a DR. A Gateway has SRs when it is a Tier-0 Gateway or a Tier-1 Gateway and has services configured such as load balancing, NAT, or Dynamic Host Configuration Protocol (DHCP).

Figure 4-4. Traditional NSX Routing



Virtual Routing and Forwarding

A Virtual Routing and Forwarding (VRF) gateway ensures that multiple instances of a routing table exist within the same gateway at the same time. VRFs are the Layer 3 equivalent of a VLAN. A VRF gateway must be linked to a Tier-0 gateway. From the Tier-0 gateway, the VRF gateway inherits the failover mode, Edge cluster, internal transit subnet, T0-T1 transit subnets, and BGP routing configuration.

In a multi-tenant solution, such as this architecture, VRFs allow a single Tier-0 gateway to be deployed and managed while isolating the routing tables between tenants. Each VRF can peer to a different eBGP neighbor and Autonomous System (AS).

Figure 4-5. VRF Routing

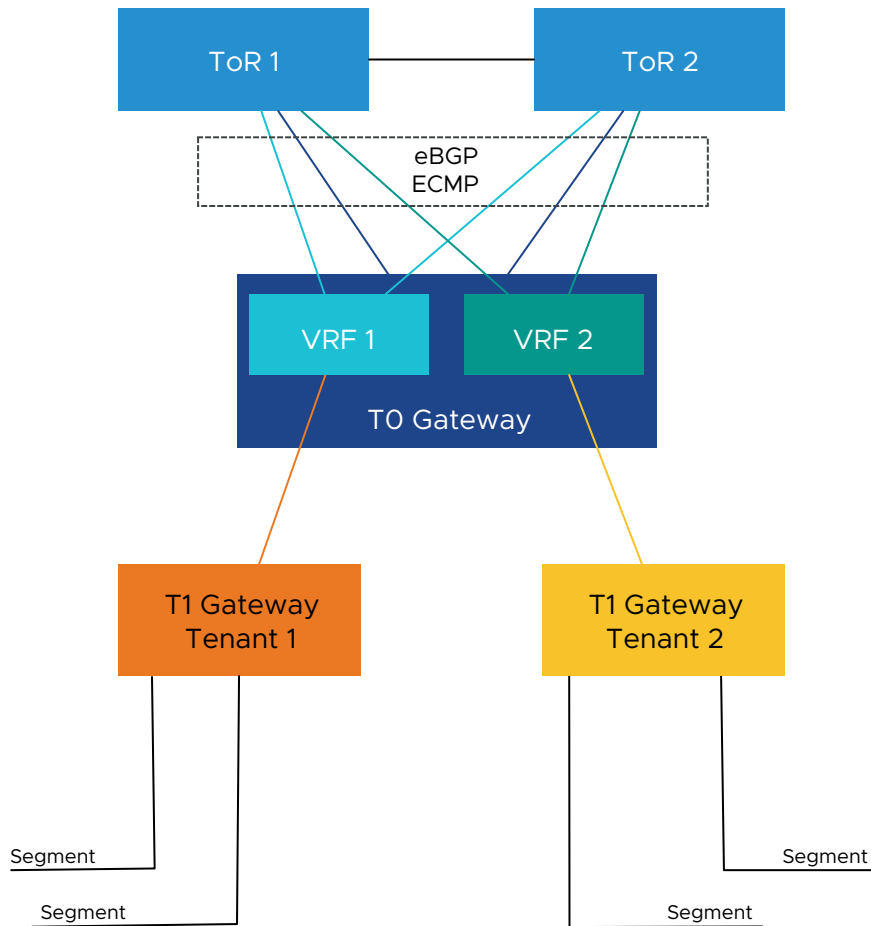


Table 4-10. Recommended NSX Design

Design Recommendation	Design Justification	Design Implication
Deploy a three-node NSX Manager cluster using the large-sized appliance to configure and manage all NSX-based compute clusters.	The large-sized appliance supports more than 64 ESXi hosts. The small-sized appliance is for proof of concept and the medium size supports up to 64 ESXi hosts only.	The large-size deployment requires more resources in the management cluster.
Create a VLAN and Overlay Transport zone.	Ensures that all Segments are available to all ESXi hosts and edge VMs configured as Transport Nodes.	None
Configure ESXi hosts to use the vSphere Distributed Switch with enhanced data path mode in each compute cluster.	Provides a high-performance network stack for NFV workloads.	Enhanced data path mode requires more CPU resources compared to standard or ENS interrupt mode.
Use large-sized NSX Edge VMs.	The large-sized appliance provides the required performance characteristics, if a failure occurs.	Large-sized Edges consume more CPU and memory resources.
Deploy at least two large-sized NSX Edge VMs in the vSphere Edge Cluster.	Creates the NSX Edge cluster to meet availability requirements.	None
Create an uplink profile with the load balance source teaming policy with two active uplinks for ESXi hosts.	For increased resiliency and performance, supports the concurrent use of two physical NICs on the ESXi hosts by creating two TEPs.	None
Create a second uplink profile with the load balance source teaming policy with two active uplinks for Edge VMs.	For increased resiliency and performance, supports the concurrent use of two virtual NICs on the Edge VMs by creating two TEPs.	None
Create a Transport Node Policy with the VLAN and Overlay Transport Zones, VDS settings, and Physical NICs per vSphere Cluster.	Allows the profile to be assigned directly to the vSphere cluster and ensures consistent configuration across all ESXi hosts in the cluster.	You must create all required Transport Zones before creating the Transport Node Policy.
Create two VLANs to enable ECMP between the Tier-0 Gateway and the Layer 3 device (ToR or upstream device).	Supports multiple equal-cost routes on the Tier-0 Gateway and provides more resiliency and better bandwidth use in the network.	Extra VLANs are required.
Note The ToR switches or the upstream Layer 3 devices have an SVI on one of the two VLANs. Each edge VM has an interface on each VLAN.		
Deploy an Active-Active Tier-0 Gateway.	Supports ECMP North-South routing on all edge VMs in the NSX Edge cluster.	Active-Active Tier-0 Gateways cannot provide services such as NAT. If you deploy a specific solution that requires stateful services on the Tier-0 Gateway, you must deploy a Tier-0 Gateway in Active-Standby mode.

Table 4-10. Recommended NSX Design (continued)

Design Recommendation	Design Justification	Design Implication
Create two VLANs per VRF to enable ECMP between the Tenant VRFs and the Layer 3 device (ToR or upstream device).	Supports multiple equal-cost routes on the VRFs and provides more resiliency and better bandwidth use in the network.	Extra VLANs are required.
Note The ToR switches or the upstream Layer 3 devices have an SVI on one of the two VLANs. Each edge VM has an interface on each VLAN.		
Deploy a VRF per tenant.	Allows each tenant to maintain their own isolated routing table. By using VRFs a single NSX Edge Cluster and TO Gateway can be used for all tenants.	Similar to TO gateways, VRFs must be created manually before VMware Cloud Director can use them.
Create Tier-1 Gateways for each tenant and connect them to the tenants VRF.	Creates a two-tier routing architecture that supports load balancers and NAT. Because the Tier-1 is always Active/Standby, creation of services such as load balancers or NAT is possible.	None
Deploy Tier-1 Gateways with the Non-Preemptive setting.	Ensures that when the failed Edge Transport Node comes back online it does not move services back to itself resulting in a small service outage.	None
Replace the certificate of the NSX Manager instances with a certificate that is signed by a third-party Public Key Infrastructure.	Ensures that the communication between NSX administrators and the NSX Manager instance is encrypted by using a trusted certificate.	Replacing and managing certificates is an operational overhead.
Replace the NSX Manager cluster certificate with a certificate that is signed by a third-party Public Key Infrastructure.	Ensures that the communication between the virtual IP address of the NSX Manager cluster and NSX administrators is encrypted by using a trusted certificate.	Replacing and managing certificates is an operational overhead.

Note This design describes VM-based NSX Edges. If required, Bare Metal Edges can be used instead of the VM-based NSX Edges.

Shared Storage Design

VMware vSAN Storage design includes conceptual design, logical design, network design, cluster and disk group design, and policy design.

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

vSAN Disk Groups

Disk group sizing is an important factor during the volume design. If more ESXi hosts are available in a cluster, more failures are tolerated in the cluster. This capability adds cost because additional hardware is required for the disk groups.

More available disk groups can increase the recoverability of vSAN during a failure.

When deciding on the number of disk groups per ESXi host, consider these data points:

- Amount of available space on the vSAN datastore.
- Number of failures that can be tolerated in the cluster.

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

Table 4-11. Recommended Shared Storage Design

Design Recommendation	Design Justification	Design Implication
Configure vSAN with a minimum of one disk group per ESXi host.	Single disk group provides the required performance and usable space for the datastore.	<ul style="list-style-type: none"> ■ Losing the caching tier disk in an ESXi host takes the disk group offline. ■ Two or more disk groups must be used to increase the availability and performance.

Note This shared storage design uses vSAN. Any supported storage solution can be used as long as it meets the characteristics of this storage design. For best practices, see the vendor documentation.

Cloud Automation Design

The Cloud Automation design includes VMware Cloud Director components to consume software-defined storage, networking, and compute.

VMware Cloud Director Cells

A VMware Cloud Director server group consists of one or more VMware Cloud Director servers.

Each server in the group runs a collection of services called a VMware Cloud Director cell. All cells share a single VMware Cloud Director database, transfer server storage, and connect to the vSphere and network resources. The installation and configuration of VMware Cloud Director process creates the cells, connects them to the shared database, and transfers server storage.

The cells communicate with each other through an ActiveMQ message bus on the primary interface. They also share a common VMware Cloud Director database where the cells persist configuration and state data. The transfer service requires that all cells have access to a common NFS mount.

Figure 4-6. VMware Cloud Director Deployment Architecture

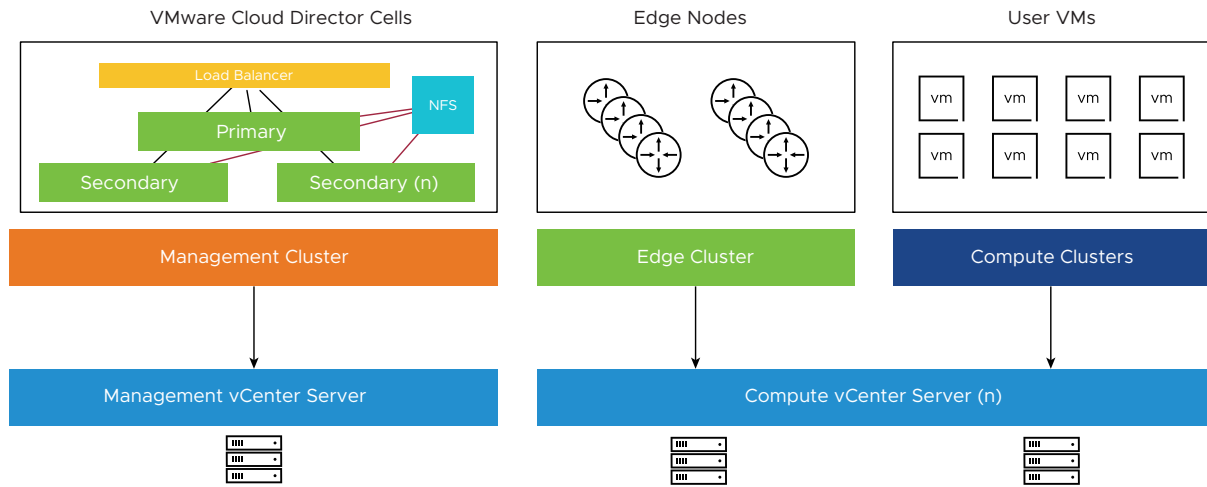


Figure 4-7. VMware Cloud Director Appliance Database HA Cluster

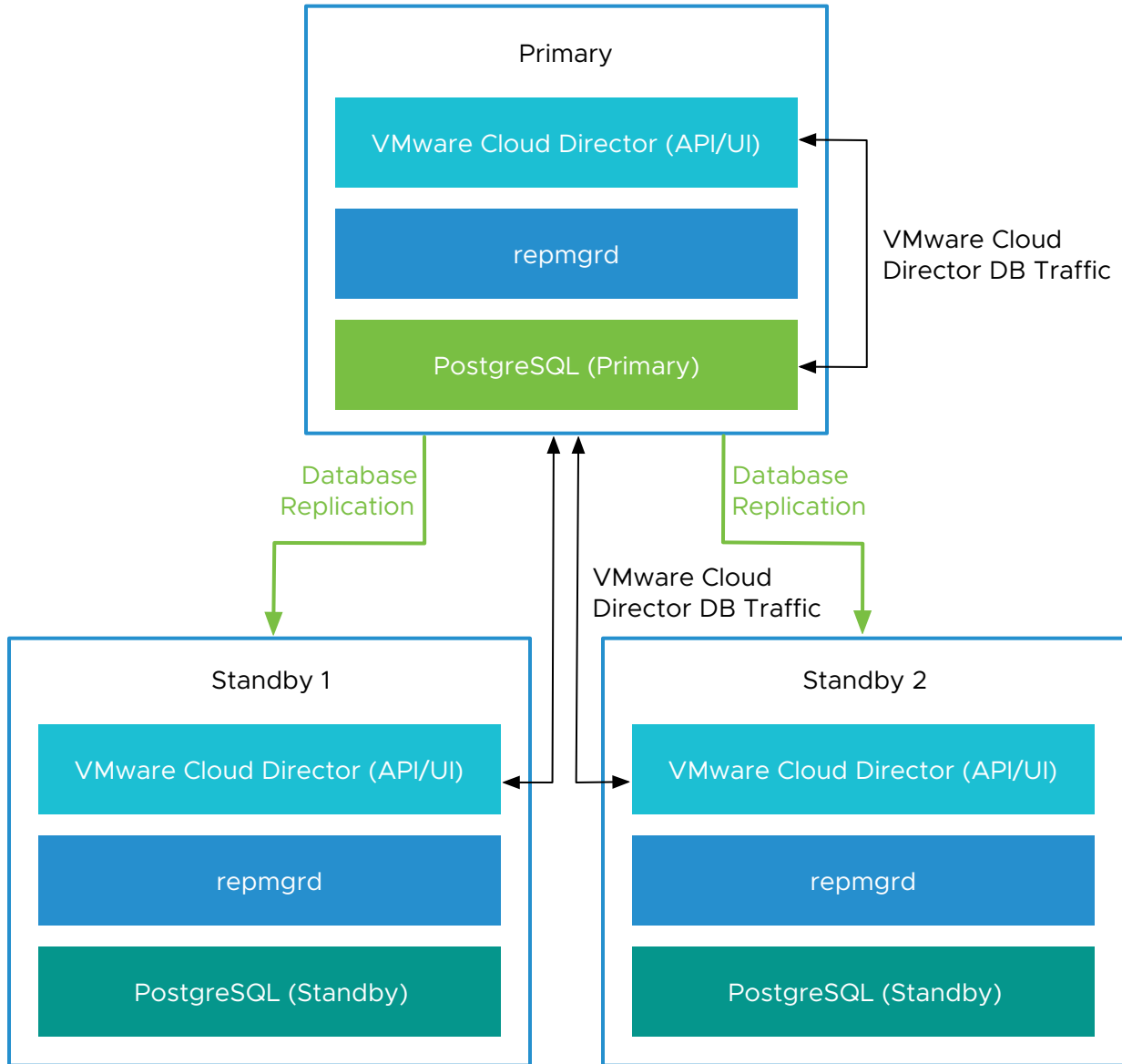


Table 4-12. Recommended VMware Cloud Director Cell Design

Design Recommendation	Design Justification	Design Implication
Deploy cells using the VMware Cloud Director Appliance.	Ensures a consistent deployment across cells.	None
Deploy at least 3 cells, one Primary and two Secondary.	Provides resiliency across cells and tolerates a cell failure.	Consumes more resources in the management cluster.
Configure VMware Cloud Director for automatic Primary Cell failover.	Reduces the downtime of VMware Cloud Director by eliminating manual failover of the Primary Cell.	Each time a new cell is added, you must change the failover setting back to automatic.
Ensure that cells can communicate with each other through the message bus on the primary network interface.	Ensures that cells do not enter a split-brain state.	None

Table 4-12. Recommended VMware Cloud Director Cell Design (continued)

Design Recommendation	Design Justification	Design Implication
Use the same consoleproxy certificate on all cells.	Ensures that users can connect to the consoleproxy service regardless of the cell they are connected to.	You must manually install the certificate on each cell.
Verify that the cell transfer share is accessible for all cells.	Required for proper functioning of VMware Cloud Director.	None

VMware Cloud Director Leases

VMware Cloud Director Leases provide control over an organization's storage and compute resources by specifying the maximum amount of time the vApps can run and the maximum amount of time the vApps and vApp templates can be stored.

- **Runtime Lease:** A runtime lease prevents inactive vApps from consuming compute resources. For example, if a user starts a vApp and goes on vacation without stopping it, the vApp continues to consume resources. A runtime lease begins when a user starts a vApp. When a runtime lease expires, VMware Cloud Director stops the vApp.
- **Storage Lease:** A storage lease prevents unused vApps and vApp templates from consuming storage resources. A vApp storage lease begins when a user stops the vApp. Storage leases do not affect the running vApps. A vApp template storage lease begins when a user adds the vApp template to a vApp, adds the vApp template to a workspace, and downloads, copies, or moves the vApp template.

When a storage lease expires, VMware Cloud Director marks the vApp or vApp template as expired or it deletes the vApp or vApp template, depending on the organization policy you set.

In a production Telco deployment, the provisioned applications are meant to run until their end of life. Hence, set the runtime and storage leases to not expire.

VMware Cloud Director Tenancy and Resource Isolation

To provide resources to an organization, one or more organization virtual data centers are created for the organization.

Provider Virtual Data Center

To make the vSphere compute, memory, and storage resources available to VMware Cloud Director, a provider Virtual Data Center (VDC) is required.

Before an organization deploys VMs or creates catalogs, the system administrator must create a provider VDC and the organization VDCs that consume its resources. The relationship of provider VDCs to the organization VDCs they support is an administrative decision. The decision can be based on the scope of your service offerings, the capacity and geographical distribution of your vSphere infrastructure, and similar considerations.

Because a provider VDC constrains the vSphere capacity and services available to tenants, system administrators commonly create provider VDCs that furnish different classes of service, as measured by performance, capacity, and features.

Tenants can then be provisioned with organization VDCs that deliver specific classes of service defined by the configuration of the backing provider VDC. Before you create a provider VDC, consider the set of vSphere capabilities that you plan to offer your tenants. Some of these capabilities can be implemented in the primary resource pool of the provider VDC.

Allocation Models

To allocate resources to an organization, you must create an organization VDC. An organization VDC obtains its resources from a provider VDC. One organization can have multiple organization VDCs.

An allocation model determines how and when the allocated compute and memory resources of provider VDC are committed to the organization VDC.

The following table shows the vSphere resource distribution settings at the VM or resource pool level based on the organization VDC allocation model:

	Flex Allocation Model	Elastic Allocation Pool Model	Non-Elastic Allocation Pool Model	Pay-As-You-Go Model	Reservation Pool Model
Elastic	The Elastic setting is based on the organization VDC configuration.	Yes	No	Yes	No
vCPU Speed	If a VM CPU limit is not defined in a VM sizing policy, vCPU speed might impact the VM CPU limit within the VDC.	Impacts the number of running vCPUs in the Organization VDC.	Not Applicable	Impacts the VM CPU limit.	Not Applicable
Resource Pool CPU Limit	The CPU limit of an Organization VDC is apportioned based on the number of VMs in the resource pool.	Organization VDC CPU allocation	Organization VDC CPU allocation	Unlimited	Organization VDC CPU allocation

	Flex Allocation Model	Elastic Allocation Pool Model	Non-Elastic Allocation Pool Model	Pay-As-You-Go Model	Reservation Pool Model
Resource Pool CPU Reservation	The CPU reservation of an Organization VDC is apportioned based on the number of vCPUs in the resource pool. Organization VDC CPU reservation equals the organization VDC CPU allocation times the CPU guarantee.	Sum of powered on VMs and equals the CPU guarantee times the vCPU speed, times the number of vCPUs.	Organization VDC CPU allocation times the CPU guarantee	None, expandable	Organization VDC CPU allocation
Resource Pool Memory Limit	The memory limit of an Organization VDC is apportioned based on the number of VMs in the resource pool.	Unlimited	Organization VDC RAM allocation	Unlimited	Organization VDC RAM allocation
Resource Pool Memory Reservation	The RAM reservation of an Organization VDC is apportioned based on the number of VMs in the resource pool. The organization VDC RAM reservation equals the organization VDC RAM allocation times the RAM guarantee.	Sum of RAM guarantee times vRAM of all powered-on VMs in the resource pool. The resource pool RAM reservation is expandable.	Organization VDC RAM allocation times the RAM guarantee	None, expandable	Organization VDC RAM allocation
VM CPU Limit	Based on the VM sizing policy of the VM.	Unlimited	Unlimited	vCPU speed times the number of vCPUs	Custom

	Flex Allocation Model	Elastic Allocation Pool Model	Non-Elastic Allocation Pool Model	Pay-As-You-Go Model	Reservation Pool Model
VM CPU Reservation	Based on the VM sizing policy of the VM.	0	0	Equals the CPU speed times the vCPU speed, times the number of vCPUs.	Custom
VM RAM Limit	Based on the VM sizing policy of the VM.	Unlimited	Unlimited	vRAM	Custom
VM RAM Reservation	Based on the VM sizing policy of the VM.	0	Equals vRAM times RAM guarantee plus RAM overhead.	Equals vRAM times RAM guarantee plus RAM overhead.	Custom

Each allocation model can be used for different levels of performance control and management. The suggested uses of each allocation model are as follows:

■ **Flex Allocation Model:**

- With the flex allocation model, you can achieve a fine-grained performance control at the workload level. VMware Cloud Director system administrators can manage the elasticity of individual organization VDCs. Cloud providers can have better control over memory overhead in an organization VDC and can enforce a strict burst capacity use for tenants.

Note The flex allocation model uses policy-based management of workloads.

■ **Allocation Pool Allocation Model:**

- Use the allocation pool allocation model for long-lived, stable workloads, where tenants subscribe to a fixed compute resource consumption and cloud providers can predict and manage the compute resource capacity. This model is optimal for workloads with diverse performance requirements.
- With this model, all workloads share the allocated resources from the resource pools of vCenter Server.
- Regardless of whether you enable or disable elasticity, tenants receive a limited amount of compute resources. Cloud providers can enable or disable the elasticity at the system level and the setting applies to all allocation pool organization VDCs. If you use the non-elastic allocation pool allocation, the organization VDC pre-reserves the VDC resource pool and tenants can overcommit vCPUs but cannot overcommit any memory. If you use the elastic pool allocation, the organization VDC does not pre-reserve any compute resources, and capacity can span through multiple clusters.

Note Cloud providers manage the overcommitment of physical compute resources and tenants cannot overcommit vCPUs and memory.

■ **Pay-as-You-Go Model:**

- Use the pay-as-you-go model when you do not have to allocate compute resources in vCenter Server upfront. Reservation, limit, and shares are applied on every workload that tenants deploy in the VDC.
- With this model, every workload in the organization VDC receives the same percentage of the configured compute resources reserved. In VMware Cloud Director, the CPU speed of every vCPU for every workload is the same and you can only define the CPU speed at the organization VDC level. From a performance perspective, because the reservation settings of individual workloads cannot be changed, every workload receives the same preference.
- This model is optimal for tenants that need workloads with different performance requirements to run within the same organization VDC.
- Because of the elasticity, this model is suitable for generic, short-lived workloads that are part of autoscaling applications.
- With this model, tenants can match spikes in compute resources demand within an organization VDC.

■ **Reservation Pool Allocation Model:**

- Use the reservation pool allocation model when you need a fine-grained control over the performance of workloads that are running in the organization VDC.
- From a cloud provider perspective, this model requires an upfront allocation of all compute resources in vCenter Server.

Note This model is not elastic.

- This model is optimal for workloads that run on hardware that is dedicated to a specific tenant. In such cases, tenant users can manage use and overcommitment of compute resources.

Figure 4-8. VMware Cloud Director with Multiple Organizations

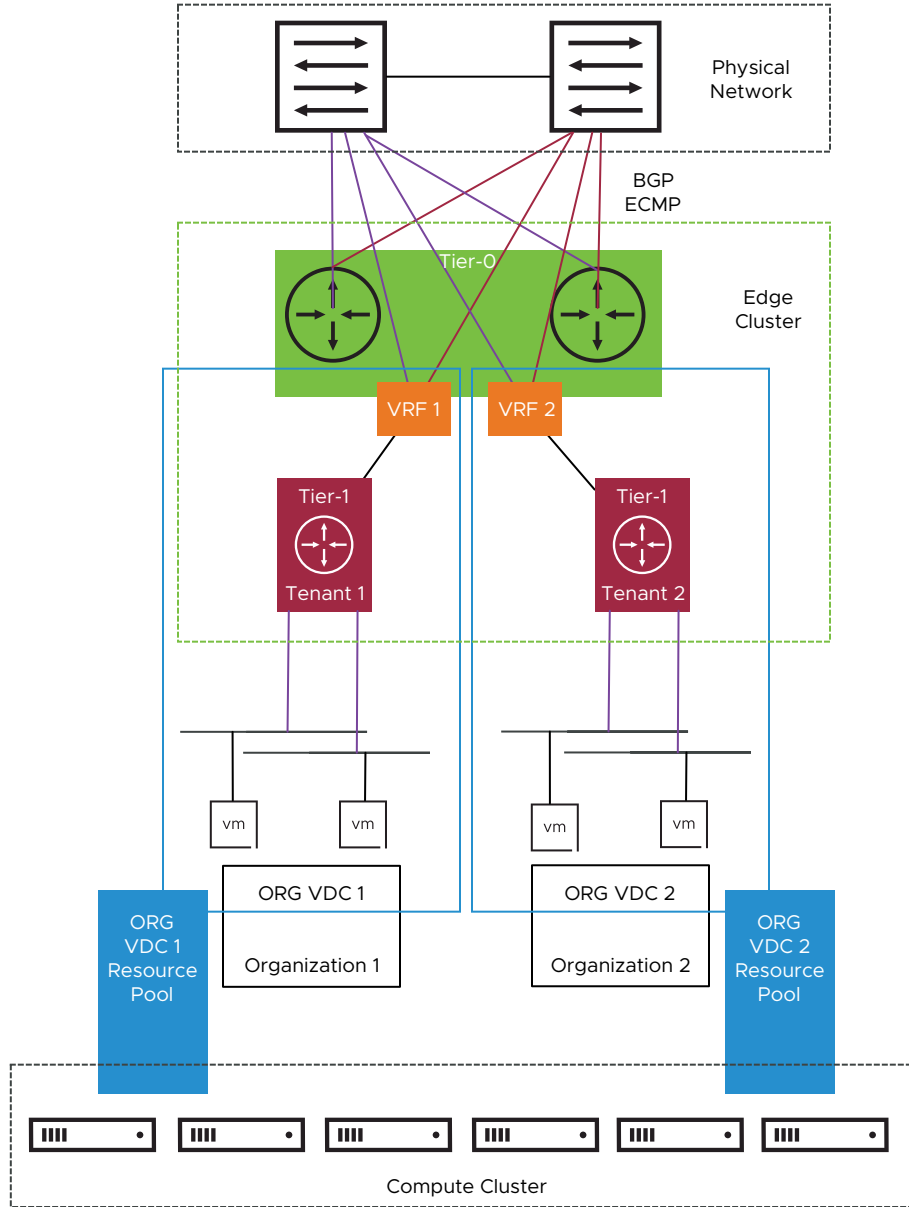


Table 4-13. Recommended Tenancy and Resource Isolation Design

Design Recommendation	Design Justification	Design Implication
Create at least a single Provider VDC.	Required to create Organizations and corresponding Organization VDCs.	None
Create an Organization per vendor.	Ensures isolation between different vendors in the environment.	None
Create at least a single Organization VDC per Organization.	Allows the Organization to deploy workloads.	If not sized properly, an Organization VDC can have unused or overcommitted resources.

Table 4-13. Recommended Tenancy and Resource Isolation Design (continued)

Design Recommendation	Design Justification	Design Implication
Use the Flex allocation model.	Allows for fine-grained control of resource allocations to each organization VDC.	None
Configure Storage and runtime leases for production VDCs to not expire.	Production workloads are meant to run until their end of life and then decommissioned manually.	By not setting any leases, workloads that are no longer being used could continue to run, resulting in wasted resources.

VMware Cloud Director Networking

VMware Cloud Director integrates with NSX-T to build rich networking topologies and configure advanced network policies in the cloud.

Network Pools

A network pool is a group of undifferentiated networks that is available for use in an organization VDC to create vApp networks and certain types of organization VDC networks. VMware Cloud Director uses network pools to create NAT-routed and internal organization VDC networks and all vApp networks. Network traffic on each network in a pool is isolated at Layer 2 from all other networks.

Each organization VDC in VMware Cloud Director can have one network pool. Multiple organization VDCs can share a network pool. The network pool for an organization VDC provides the networks created to satisfy the network quota for an organization VDC.

Every provider VDC that is backed by NSX-T Data Center includes a Geneve network pool. When you create a provider VDC that is backed by NSX-T Data Center, you can associate that provider VDC with an existing Geneve network pool, or you can create a Geneve network pool for the provider VDC.

VMware Cloud Director Geneve networks provide a number of benefits:

- Logical networks spanning Layer 3 boundaries
- Logical networks spanning multiple racks on a single Layer 2
- Broadcast containment
- Higher performance
- Greater scale (up to 16 million network addresses)

External Networks

A VMware Cloud Director external network provides an uplink interface that connects networks and VMs in the system to a network outside of the system, such as a VPN, a corporate intranet, or the public Internet.

Note The range of IP addresses defined for the external network are allocated either to an edge gateway or to the VMs that are directly connected to the network. Hence, the IP addresses must not be used outside of VMware Cloud Director.

An external network can be backed by an NSX-T Data Center tier-0 logical router. You can also create an external network that is backed by a VRF-lite tier-0 gateway in NSX-T Data Center. A VRF gateway is created from a parent tier-0 gateway. It has its own routing tables. Multiple VRF's can exist within the parent tier-0 gateway. This allows VDC's to have their own external network without deploying multiple tier-0 gateways.

NSX Edge Gateways

An NSX-T Data Center edge gateway provides a routed organization VDC network or a data center group network with connectivity to external networks and IP management properties. It can also provide services such as firewall, NAT, IPsec VPN, DNS forwarding, and DHCP, which is enabled by default.

Routing

By using route advertisement, you can create a fully routed network environment in an organization virtual data center (VDC). You can decide which of the network subnets that are attached to the NSX-T Data Center edge gateway to advertise to the dedicated external network. If a subnet is not added to the advertisement filter, the route to it is not advertised to the external network and the subnet remains private.

Route advertisement is automatically configured on the NSX-T Data Center edge gateway. VMware Cloud Director supports automatic route redistribution when you use route advertisement on an NSX-T edge gateway. Route redistribution is automatically configured on the tier-0 logical router that represents the dedicated external network.

You can configure an external or internal Border Gateway Protocol (eBGP or iBGP) connection between an NSX-T Data Center edge gateway that has a dedicated external network and a router in your physical infrastructure.

- **BGP speaker:** A networking device that is running BGP. Two BGP speakers establish a connection before any routing information is exchanged.
- **BGP neighbor:** A BGP speaker that established a connection. After establishing the connection, the devices exchange routes and synchronize their tables. Each device sends keep-alive messages to keep this connection alive.

In an edge gateway that is connected to an external network backed by a VRF gateway, the local AS number and graceful restart settings are inherited from the parent Tier-0 gateway and they cannot be changed at the VRF level.

Table 4-14. Recommended VMware Cloud Director Network Design

Design Recommendation	Design Justification	Design Implication
Create a Geneve Network pool.	Required to create NSX-T backed network resources.	None
Create a dedicated external network per Organization.	Allows the use of a fully routed network topology.	None
Create a VRF per Organization.	Allows all Organizations to share a single parent Tier-0 gateway, while maintaining isolation between Organizations. Sharing a single parent gateway reduces the number of NSX Edges required in the deployment.	VRFs inherit the local AS and graceful restart configuration of the parent Tier-0, such that each VRF has the same local AS.
Create one or more Edge Gateways (Tier-1 gateways) per Organization VDC.	Enables networking services for the Organization VDC.	None

VMware Cloud Director Storage

Storage Policy Based Management (SPBM): VMware Cloud Director uses SPBM to define storage characteristics. Within a software-defined data center, SPBM plays a major role by helping to align storage with application demands of your virtual machines. It provides a storage policy framework that serves as a single unified control panel across a broad range of data services and storage solutions.

As an abstraction layer, SPBM abstracts storage services delivered by vVols, vSAN, I/O filters, or other storage entities. Instead of integrating with each individual type of storage and data services, SPBM provides a universal framework for different types of storage entities.

SPBM provides the following mechanisms:

- Advertisement of storage capabilities and data services that storage arrays and other entities such as I/O filters offer.
- Bidirectional communications between ESXi and vCenter Server on one side and between storage arrays and entities on the other side.
- VM provisioning based on VM storage policies.

VMware Cloud Director uses the SPBM policies defined in vCenter Server. These policies are assigned to a Provider VDC and are available to Organization VDCs managed by the Provider VDC.

IOPS: You can enable the I/O operations per second (IOPS) setting for a storage policy so that tenants can set per-disk IOPS limits.

Managed read/write performance in physical storage devices and virtual disks is defined in units called IOPS, which measure read/write operations per second. To limit I/O performance, a provider VDC storage policy that includes storage devices with enabled IOPS allocation must back an organization VDC storage policy. Afterwards, a tenant can configure disks to request a

specified level of I/O performance. A storage profile configured with IOPS support delivers its default IOPS value to all disks that use it. That includes disks that are not configured to request a specific IOPS value. A hard disk configured to request a specific IOPS value cannot use a storage policy whose maximum IOPS value is lower than the requested value, or a storage policy that is not configured with IOPS support.

You can edit the default IOPS settings. For example, you can set limits on IOPS per disk or IOPS per storage policy. You can set IOPS limits per disk based on the disk size in GB so that you grant more IOPS to larger disks. Tenants can set custom IOPS on a disk within these limits. You can use IOPS limiting with or without IOPS capacity considerations for placement.

You cannot enable IOPS on a storage policy backed by a Storage DRS cluster.

The storage policies are dependent on the workload being deployed, so generic design recommendations would be irrelevant. Instead, create storage policies based on the storage type and workload demands such that the vendor requirements are met.

VMware Cloud Director Roles and Authentication

A right is the fundamental unit of access control in VMware Cloud Director. A role associates a role name with a set of rights. Each organization can have different rights and roles.

Roles

VMware Cloud Director uses roles and their associated rights to determine whether a user or group is authorized to perform an operation. Most of the procedures documented in the VMware Cloud Director guides include a prerequisite role. These prerequisites assume that the named role is the unmodified predefined role or a role that includes an equivalent set of rights.

System administrators can use rights bundles and global tenant roles to manage the rights and roles that are available to each organization.

Predefined Provider Roles:

- **System Administrator:** The System Administrator role exists only in the provider organization. This role includes all rights in the system. A System Administrator can create additional system administrator and user accounts in the provider organization.
- **Multisite System:** Used for running the heartbeat process for multisite deployments. This role has only a single right, Multisite: System Operations, which gives the permission to make a VMware Cloud Director OpenAPI request that retrieves the status of the remote member of a site association.

Predefined Global Tenant Roles:

- **Organization Administrator:** After creating an organization, a System Administrator can assign the role of Organization Administrator to any user in the organization. A user with the predefined Organization Administrator role can manage users and groups in their organization and assign them roles, including the predefined Organization Administrator role. Roles created or modified by an Organization Administrator are not visible to other organizations.
- **Catalog Author:** The predefined Catalog Author role allows a user to create and publish catalogs.
- **vApp Author:** The predefined vApp Author role allows a user to use catalogs and create vApps.
- **vApp User:** The predefined vApp User role allows a user to use existing vApps.
- **Console Access Only:** The predefined Console Access Only role allows a user to view VM state and properties and to use the guest OS.
- **Defer to Identity Provider:** The rights associated with the predefined Defer to Identity Provider role are determined based on the information received from the user's OAuth or SAML Identity Provider. When a user or group is assigned the Defer to Identity Provider role, the user or group name provided by the Identity Provider must exactly match the role or group name defined in your organization. Otherwise, the user or group is not qualified for inclusion.
 - If an OAuth Identity Provider defines a user, the user is assigned the roles named in the roles array of the user's OAuth token.
 - If a SAML Identity Provider defines a user, the user is assigned the roles named in the SAML attribute. The SAML attribute name appears in the RoleAttributeName element, which is in the SamlAttributeMapping element in the organization's OrgFederationSettings.

Note

- If a user is assigned the Defer to Identity Provider role but no matching role or group name is available in your organization, the user can log in to the organization but has no rights. If an Identity Provider associates a user with a system-level role such as System Administrator, the user can log in to the organization but has no rights. You must manually assign a role to such users.
- Except the Defer to Identity Provider role, each predefined role includes a set of default rights. Only a System Administrator can modify the rights in a predefined role. If a System administrator modifies a predefined role, the modifications propagate to all instances of the role in the system.

Authentication

You can integrate VMware Cloud Director with an external identity provider and import users and groups to your organizations. You can configure an LDAP server connection at a system or organization level. You can configure a SAML integration at an organization level.

You can configure an organization to use the system LDAP connection as a shared source of users and groups, or an organization can use a separate LDAP connection as a private source of users and groups.

If you want to import users and groups from a SAML identity provider to your system organization, you must configure your system organization with this SAML identity provider. Imported users can log in to the system organization with the credentials established in the SAML identity provider.

To configure VMware Cloud Director with a SAML identity provider, you establish a mutual trust by exchanging SAML service provider and identity provider metadata.

Table 4-15. Recommended VMware Cloud Director Roles and Authentication Design

Design Recommendation	Design Justification	Design Implication
Use the default VMware Cloud Director Roles, unless necessary.	Simplifies the user rights management and configuration.	Custom roles might be required for some cases where the built-in roles do not work.
Configured a system LDAP connection.	<ul style="list-style-type: none"> ■ Allows for centralized account management by leveraging the existing LDAP infrastructure. ■ Provides high security, as you do not need to create local accounts that could be left unused. 	Requires manual user import and role assignment.
Use the System LDAP connection for Organizations.	<p>Allows for centralized account management by leveraging the existing LDAP infrastructure.</p> <p>Provides a higher level of security as local accounts, that could be left when a user leaves, do not need to be created.</p>	Requires manual user import and role assignment.

Operations Management Design

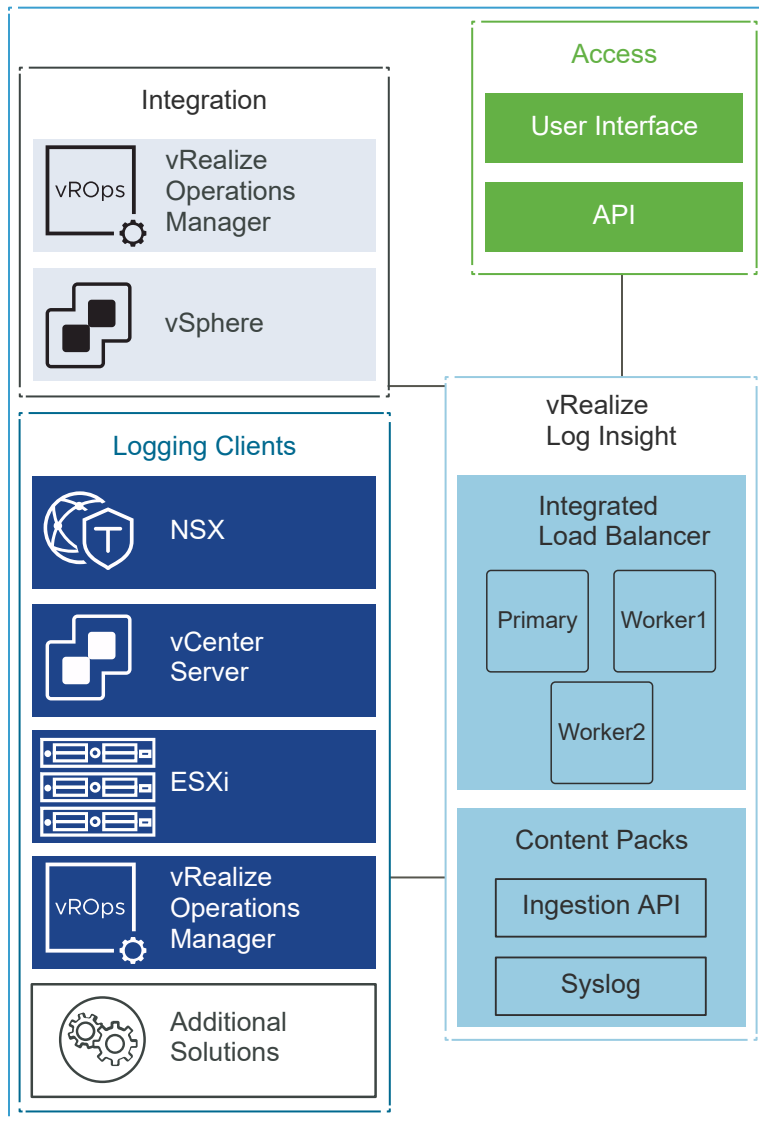
The operations management design includes software components that form the operations management layer. This design provides guidance on the main elements such as sizing, networking, and diagnostics of a product design.

vRealize Log Insight Design

The vRealize Log Insight design enables real-time logging for all components in the solution.

The vRealize Log Insight cluster consists of one primary node and two secondary nodes behind a load balancer.

Figure 4-9. Logical vRealize Log Insight Design



Enable the Integrated Load Balancer (ILB) on the three-node cluster so that all log sources can address the cluster by its ILB. By using the ILB, you do not need to reconfigure log sources with a new destination address in case of a scale-out or node failure. The ILB also guarantees that vRealize Log Insight accepts all incoming ingestion traffic.

The ILB address is required for users to connect to vRealize Log Insight through the Web UI or API and for clients to ingest logs using syslog or the Ingestion API. A vRealize Log Insight cluster can scale out to 12 nodes: 1 primary and 11 worker nodes.

To accommodate all log data in the solution, size the compute resources and storage for the Log Insight nodes correctly.

By default, the vRealize Log Insight appliance uses the predefined values for small configurations: 4 vCPUs, 8 GB virtual memory, and 530.5 GB disk space. vRealize Log Insight uses 100 GB disk space to store raw data, index, metadata, and other information.

vRealize Log Insight supports the following alerts that trigger notifications about its health and the monitored solutions:

- **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 alerts based on the data ingested by vRealize Log Insight.

Table 4-16. Recommended vRealize Log Insight Design

Design Recommendation	Design Justification	Design Implication
Deploy vRealize Log Insight in a cluster configuration of three nodes with an integrated load balancer: <ul style="list-style-type: none"> ■ one primary node ■ two worker nodes 	<ul style="list-style-type: none"> ■ Provides high availability. ■ The integrated load balancer: <ul style="list-style-type: none"> ■ Prevents a single point of failure. ■ Simplifies the vRealize Log Insight deployment and subsequent integration. ■ Simplifies the vRealize Log Insight scale-out operations reducing the need to reconfigure existing logging sources. 	<ul style="list-style-type: none"> ■ You must deploy a minimum of three medium nodes. ■ You must size each node identically. ■ If the capacity of your vRealize Log Insight cluster must expand, identical capacity must be added to each node.
Deploy vRealize Log Insight nodes of medium size.	Accommodates the number of expected syslog and vRealize Log Insight Agent connections from the following sources: <ul style="list-style-type: none"> ■ Management and Compute vCenter Servers ■ Management and Compute ESXi hosts ■ NSX-T Components ■ vRealize Operations Manager components ■ VMware Cloud Director Using medium-size appliances ensures that the storage space for the vRealize Log Insight cluster is sufficient for 7 days of data retention.	If you configure vRealize Log Insight to monitor additional syslog sources, increase the size of the nodes.

Table 4-16. Recommended vRealize Log Insight Design (continued)

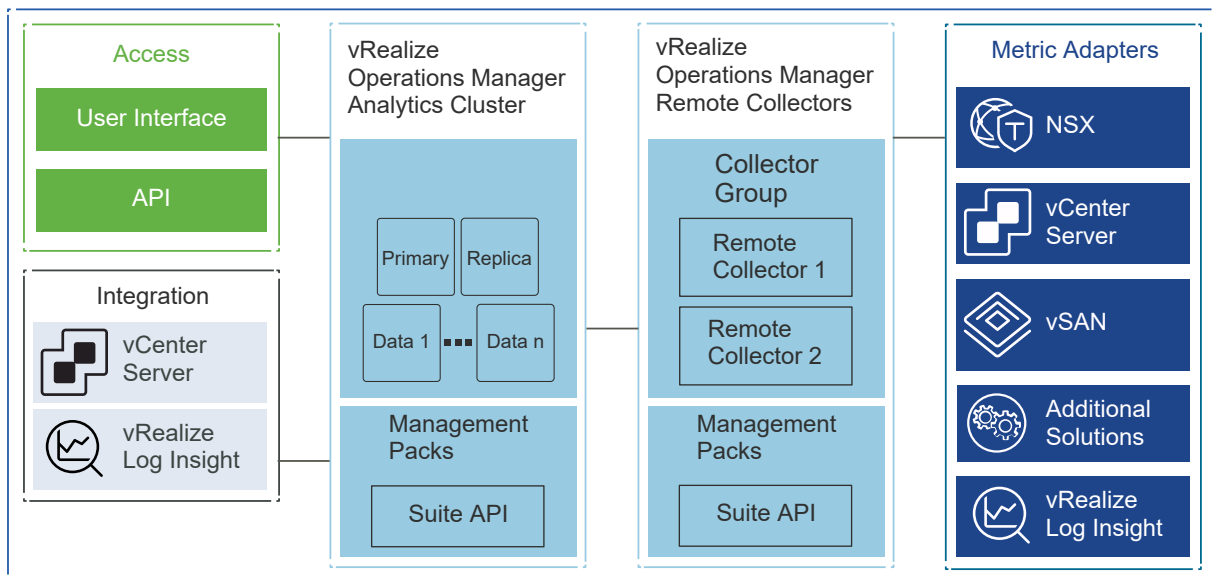
Design Recommendation	Design Justification	Design Implication
Enable alerting over SMTP.	Administrators and operators can receive email alerts from vRealize Log Insight.	Requires access to an external SMTP server.
Forward alerts to vRealize Operations Manager.	Provides monitoring and alerting information that is pushed from vRealize Log Insight to vRealize Operations Manager for centralized administration.	None

vRealize Operations Manager Design

vRealize Operations Manager communicates with all management components to collect metrics that are presented through various dashboards and views.

The deployment of vRealize Operations Manager is a single instance of a 3-node analytics cluster that is deployed in the management cluster along with a two-node remote collector group.

Figure 4-10. Logical vRealize Operations Manager Design



The analytics cluster of the vRealize Operations Manager deployment contains the nodes that analyze and store data from the monitored components. You deploy a configuration of the analytics cluster that meets the requirements for monitoring the number of VMs.

Deploy a three-node vRealize Operations Manager analytics cluster that consists of one primary node, one replica node, and one data node to enable scale-out and high availability.

This design uses medium-size nodes for the analytics cluster and standard-size nodes for the remote collector group. To collect the required number of metrics, add a virtual disk of 1 TB on each analytics cluster node.

You can use the self-monitoring capability of vRealize Operations Manager to receive alerts about issues that are related to its operational state.

vRealize Operations Manager displays the following administrative alerts:

- **System alert:** Indicates a failed component of the vRealize Operations Manager application.
- **Environment alert:** Indicates that vRealize Operations Manager stopped receiving data from one or more resources. This alert might indicate a problem with system resources or network infrastructure.
- **Log Insight log event:** Indicates that the infrastructure on which vRealize Operations Manager is running has low-level issues. You can also use the log events for root cause analysis.
- **Custom dashboard:** vRealize Operations Manager shows super metrics for data center monitoring, capacity trends, and single pane of glass overview.

Table 4-17. Recommended vRealize Operations Manager Design

Design Recommendation	Design Justification	Design Implication
Deploy vRealize Operations Manager as a cluster of three nodes: <ul style="list-style-type: none"> ■ one primary ■ one replica ■ one data node 	<ul style="list-style-type: none"> ■ Provides the scale capacity required for monitoring up to 10,000 VMs. ■ Supports scale-up with additional data nodes. 	All the nodes must be sized identically.
Deploy two remote collector nodes.	Removes the load from the analytics cluster from collecting application metrics.	When configuring the monitoring of a solution, you must assign a collector group.
Deploy each node in the analytics cluster as a medium-size appliance.	Provides the scale required to monitor the solution.	ESXi hosts in the management cluster must have physical CPUs with a minimum of 8 cores per socket. In total, vRealize Operations Manager uses 24 vCPUs and 96 GB of memory in the management cluster.
Add more medium-size nodes to the analytics cluster if the number of VMs exceeds 10,000.	Ensures that the analytics cluster has enough capacity to meet the VM object and metric growth.	<ul style="list-style-type: none"> ■ The capacity of the physical ESXi hosts must be sufficient to accommodate VMs that require 32 GB RAM without bridging NUMA node boundaries. ■ The number of nodes must not exceed the number of ESXi hosts in the management cluster minus 1. For example, if the management cluster contains six ESXi hosts, you can deploy up to five vRealize Operations Manager nodes in the analytics cluster.
Deploy the standard-size remote collector virtual appliances.	Enables metric collection for the expected number of objects.	You must provide 4 vCPUs and 8 GB memory in the management cluster.

Table 4-17. Recommended vRealize Operations Manager Design (continued)

Design Recommendation	Design Justification	Design Implication
Add a virtual disk of 1 TB for each analytics cluster node.	Provides enough storage for the expected number of objects.	You must add the 1 TB disk manually while the VM for the analytics node is powered off.
Configure vRealize Operations Manager for SMTP outbound alerts.	Enables administrators and operators to receive email alerts from vRealize Operations Manager.	Requires access to an external SMTP server.

vRealize Network Insight Design

vRealize Network Insight communicates with the vCenter Server and the NSX Manager instances to collect metrics that are presented through various dashboards and views.

vRealize Network Insight is deployed as a cluster called the vRealize Network Insight Platform Cluster. This cluster processes the collected data and presents it using a dashboard. vRealize Network Insight also uses a Proxy node to collect data from the data sources, such as vCenter Server and NSX Manager, and send the data to the Platform Cluster for processing.

Figure 4-11. Logical vRealize Network Insight Design

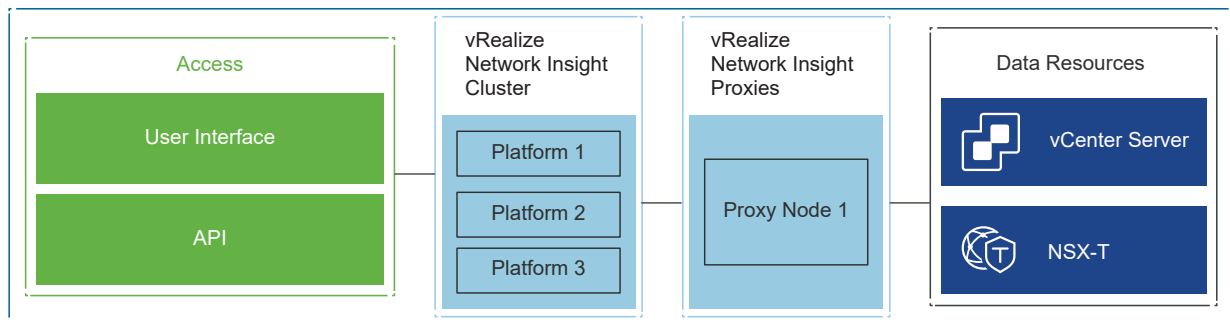


Table 4-18. Recommended vRealize Network Insight Design

Design Recommendation	Design Justification	Design Implication
Deploy a three-node vRealize Network Insight Platform cluster.	Meets the scalability requirements of up to 10,000 VMs and 2 million flows per day.	The Management cluster must be sized properly as each vRealize Network Insight VM requires a 100% CPU reservation.
Deploy vRealize Network Insight Platform nodes of large size.	Large size is the minimum size supported to form a cluster.	None
Deploy at least a single large-sized vRealize Network Insight Proxy node.	A single Proxy node meets the requirements in a new deployment. As the solution grows, additional Proxy nodes might be required.	vRealize Network Insight Proxy nodes cannot be highly available.