Architecture and Design of VMware NSX-T for Workload Domains

19 MAR 2019
VMware Validated Design 5.0.1
VMware NSX-T 2.4

vmware®

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

https://docs.vmware.com/

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

docfeedback@vmware.com

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

Contents

About Architecture and Design of VMware NSX-T for Workload Domains

Architecture and Design of VMware NSX-T for Workload Domains provides detailed information about the requirements for software, tools, and external services to implement VMware NSX-T® in a shared edge and compute cluster in an SDDC that is compliant with VMware Validated Design for Software-Defined Data Center.

Prerequisites

Deploy the management cluster according to VMware Validated Design for Software-Defined Data Center at least in a single region. See the VMware Validated Design documentation page.

Intended Audience

This design is intended for architects and administrators who want to deploy NSX-T in a virtual infrastructure workload domain for tenant workloads.

Required VMware Software

In addition to the VMware Validated Design for Software-Defined Data Center 5.0.1 deployment, you must download NSX-T 2.4. You then deploy and configure NSX-T in the shared edge and compute cluster according to this guide. See *VMware Validated Design Release Notes* for more information about supported product versions

Applying the Guidance for NSX-T for Workload Domains

1

The content in *Architecture and Design of VMware NSX-T for Workload Domains* replaces certain parts of *Architecture and Design* in VMware Validated Design for Software-Defined Data Center, also referred to as the Standard SDDC.

Before You Design the Virtual Infrastructure Workload Domain with NSX-T

Before you follow this documentation, you must deploy the components for the SDDC management cluster according to VMware Validated Design for Software-Defined Data Center at least in a single region. See Architecture and Design, Planning and Preparation and Deployment for Region A in the VMware Validated Design documentation.

- ESXi
- Platform Services Controller pair and Management vCenter Server
- NSX for vSphere
- vRealize Lifecycle Manager
- vSphere Update Manager
- vRealize Operations Manager
- vRealize Log Insight
- vRealize Automation with embedded vRealize Orchestrator
- vRealize Business

Designing a Virtual Infrastructure Workload Domain with NSX-T

Next, follow the guidance to design a virtual infrastructure (VI) workload domain with NSX-T deployed in this way:

- In general, use the guidelines about the VI workload domain and shared edge and compute cluster in the following sections of Architecture and Design in VMware Validated Design for Software-Defined Data Center:
  - **Architecture Overview > Physical Infrastructure Architecture**
  - **Architecture Overview > Virtual Infrastructure Architecture**
  - **Detailed Design > Physical Infrastructure Design**
  - **Detailed Design > Virtual Infrastructure Design**
- For the sections that are available in both *Architecture and Design of VMware NSX-T for Workload Domains* and *Architecture and Design*, follow the design guidelines in *Architecture and Design of VMware NSX-T for Workload Domains*.

First-Level Chapter	Places to Use the Guidance for NSX-T
Architecture Overview	■ Physical Infrastructure Architecture ■ Workload Domain Architecture ■ Cluster Types ■ Physical Network Architecture ■ **Network Transport** ■ Infrastructure Network Architecture ■ Physical Network Interfaces ■ **Virtual Infrastructure Architecture**
Detailed Design	■ Physical Infrastructure Design ■ Physical Design Fundamentals ■ **Physical Networking Design** ■ Physical Storage Design ■ Virtual Infrastructure Design ■ vCenter Server Design ■ vCenter Server Deployment ■ vCenter Server Networking ■ vCenter Server Redundancy ■ vCenter Server Appliance Sizing ■ **vSphere Cluster Design** ■ vCenter Server Customization ■ Use of TLS Certificates in vCenter Server ■ **Virtualization Network Design** ■ **NSX-T Design**

Architecture Overview 2

VMware Validated Design for NSX-T enables IT organizations that have deployed VMware Validated Design for Software-Defined Data Center 5.0 to create a shared edge and compute cluster that uses NSX-T capabilities.

This chapter includes the following topics:

Physical Network Architecture

VMware Validated Designs can use most physical network architectures.

Network Transport

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

VMware Validated Design supports both Layer 2 and Layer 3 transports. To decide whether to use Layer 2 or Layer 3, consider the following factors:

- NSX-T service routers establish Layer 3 routing adjacency with the first upstream Layer 3 device to provide equal cost routing for workloads.
- The investment you have today in your current physical network infrastructure.
- The benefits and drawbacks for both layer 2 and layer 3 designs.

Benefits and Drawbacks of Layer 2 Transport

A design using Layer 2 transport has these considerations:

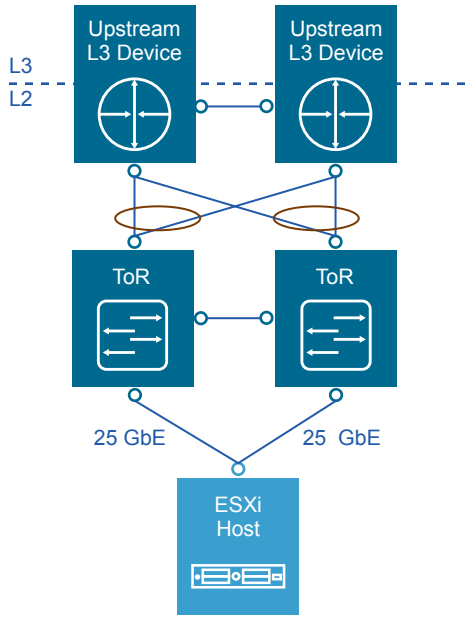
- In a design that uses Layer 2 transport, top of rack switches and upstream Layer 3 devices, such as core switches or routers, form a switched fabric.
- The upstream Layer 3 device terminates each VLAN and provides default gateway functionality.
- Uplinks from the top of rack switch to the upstream Layer 3 devices are 802.1Q trunks carrying all required VLANs.

Using a Layer 2 transport has the following benefits and drawbacks:

Table 2-1. Benefits and Drawbacks for Layer 2 Transport

Characteristic	Description
Benefits	■ More design freedom. ■ You can span VLANs across racks.
Drawbacks	■ The size of such a deployment is limited because the fabric elements have to share a limited number of VLANs. ■ You might have to rely on a specialized data center switching fabric product from a single vendor. ■ Traffic between VLANs must traverse to upstream Layer 3 device to be routed.

Figure 2-1. Example Layer 2 Transport



Benefits and Drawbacks of Layer 3 Transport

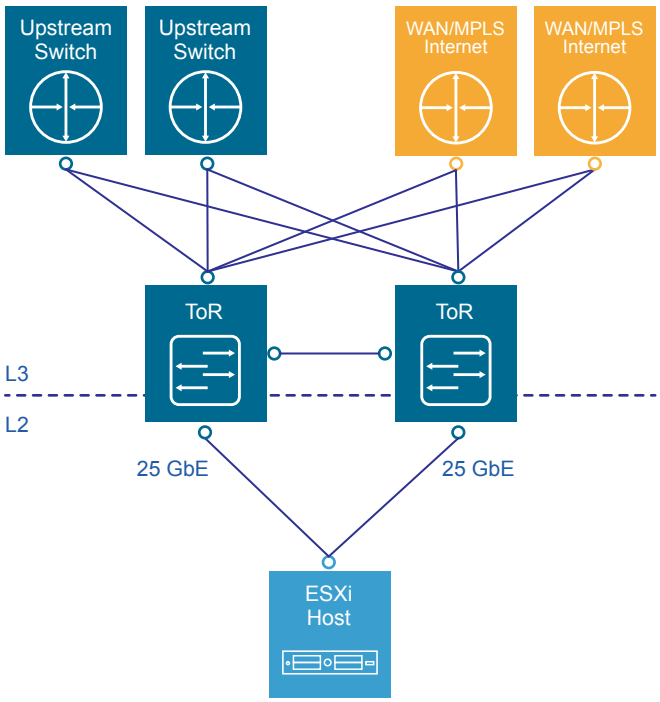
A design using Layer 3 transport requires these considerations:

- Layer 2 connectivity is limited within the data center rack up to the top of rack switches.
- The top of rack switch terminates each VLAN and provides default gateway functionality. The top of rack switch has a switch virtual interface (SVI) for each VLAN.
- Uplinks from the top of rack 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 BGP, connects the top of rack switches and upstream switches. Each top of rack switch in the rack advertises a small set of prefixes, typically one per VLAN or subnet. In turn, the top of rack switch calculates equal cost paths to the prefixes it receives from other top of rack switches.

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

Characteristic	Description
Benefits	■ You can select from many Layer 3 capable switch products for the physical switching fabric. ■ You can mix switches from different vendors because of general interoperability between their implementation of BGP. ■ This approach is typically more cost effective because it uses only the basic functionality of the physical switches.
Drawbacks	■ VLANs are restricted to a single rack. The restriction can affect vSphere Fault Tolerance, and storage networks.

Figure 2-2. Example Layer 3 Transport

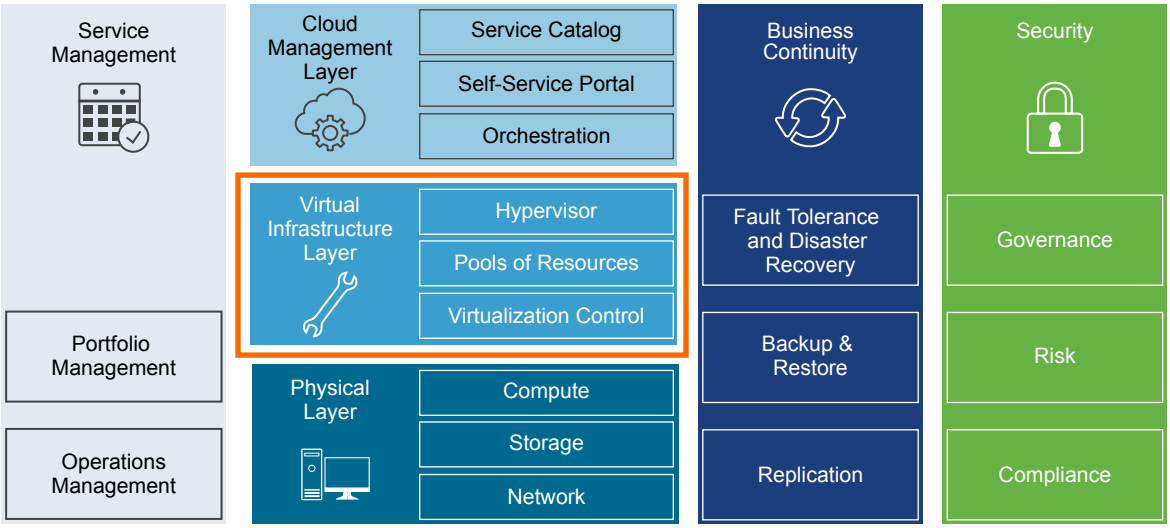


Virtual Infrastructure Architecture

The virtual infrastructure is the foundation of an operational SDDC. 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 compute workloads. The virtual infrastructure layer consists of the hypervisors on the physical hosts and the control of these hypervisors. The management components of the SDDC consist of elements in the virtual management layer itself.

Figure 2-3. Virtual Infrastructure Layer in the SDDC



Virtual Infrastructure Overview

The SDDC virtual infrastructure consists of workload domains. The SDDC virtual infrastructure includes a management workload domain that contains the management cluster and a virtual infrastructure workload domain that contains the shared edge and compute cluster.

Management Cluster

The management cluster runs the virtual machines that manage the SDDC. These virtual machines host vCenter Server, vSphere Update Manager, NSX Manager, and other management components. All management, monitoring, and infrastructure services are provisioned to a vSphere cluster which provides high availability for these critical services. Permissions on the management cluster limit access only to administrators. This limitation protects the virtual machines that are running the management, monitoring, and infrastructure services from unauthorized access. The management cluster leverages software-defined networking capabilities in NSX for vSphere.

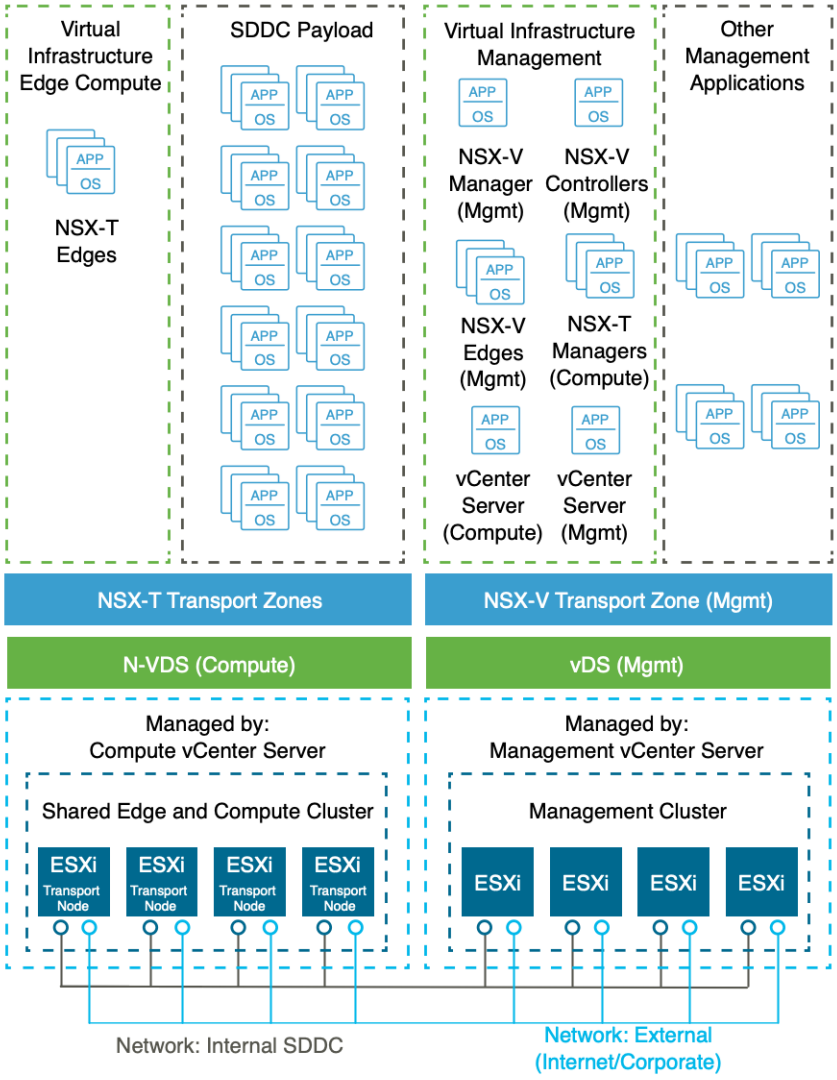
The management cluster architecture and design is covered in the VMware Validated Design for Software-Defined Data Center. The NSX-T validated design does not include the design of the management cluster.

Shared Edge and Compute Cluster

The shared edge and compute cluster runs the NSX-T edge virtual machines and all tenant workloads. The edge virtual machines are responsible for North-South routing between compute workloads and the external network. This is often referred to as the on-off ramp of the SDDC.

The hosts in this cluster provide services such as high availability to the NSX-T edge virtual machines and tenant workloads.

Figure 2-4. SDDC Logical Design



Network Virtualization Components

The NSX-T platform consists of several components that are relevant to the network virtualization design.

NSX-T Platform

NSX-T creates a network virtualization layer, which is an abstraction between the physical and virtual networks. You create all virtual networks on top of this layer.

Several components are required to create this network virtualization layer:

- NSX-T Managers
- NSX-T Edge Nodes
- NSX-T Distributed Routers (DR)
- NSX-T Service Routers (SR)
- NSX-T Segments (Logical Switches)

These components are distributed in different planes to create communication boundaries and provide isolation of workload data from system control messages.

Data plane

Performs stateless forwarding or transformation of packets based on tables populated by the control plane, reports topology information to the control plane, and maintains packet level statistics.

The following traffic runs in the data plane:

- Workload data
- N-VDS virtual switch, distributed routing, and the distributed firewall in NSX-T

  The data is carried over designated transport networks in the physical network.

Control plane

Contains messages for network virtualization control. You place the control plane communication on secure physical networks (VLANs) that are isolated from the transport networks for the data plane.

The control plane computes the runtime state based on configuration from the management plane. Control plane propagates topology information reported by the data plane elements, and pushes stateless configuration to forwarding engines.

Control plane in NSX-T has two parts:

- Central Control Plane (CCP). The CCP is implemented as a cluster of virtual machines called CCP nodes. The cluster form factor provides both redundancy and scalability of resources.

  The CCP is logically separated from all data plane traffic, that is, a failure in the control plane does not affect existing data plane operations.
- Local Control Plane (LCP). The LCP runs on transport nodes. It is near to the data plane it controls and is connected to the CCP. The LCP is responsible for programming the forwarding entries of the data plane.

Management plane

Provides a single API entry point to the system, persists user configuration, handles user queries, and performs operational tasks on all management, control, and data plane nodes in the system.

For NSX-T, all querying, modifying, and persisting user configuration is in the management plane. Propagation of that configuration down to the correct subset of data plane elements is in the control plane. As a result, some data belongs to multiple planes. Each plane uses this data according to stage of existence. The management plane also queries recent status and statistics from the control plane, and under certain conditions directly from the data plane.

The management plane is the only source of truth for the logical system because it is the only entry point for user configuration. You make changes using either a RESTful API or the NSX-T user interface.

For example, responding to a vSphere vMotion operation of a virtual machine is responsibility of the control plane, but connecting the virtual machine to the logical network is responsibility of the management plane.

Network Virtualization Services

Network virtualization services include segments, gateways, firewalls, and other components of NSX-T.

Segments (Logical Switch)

Reproduces switching functionality, broadcast, unknown unicast, and multicast (BUM) traffic in a virtual environment that is decoupled from the underlying hardware.

Segments are similar to VLANs because they provide network connections to which you can attach virtual machines. The virtual machines can then communicate with each other over tunnels between ESXi hosts. Each Segment has a virtual network identifier (VNI), like a VLAN ID. Unlike VLANs, VNIs scale beyond the limits of VLAN IDs.

Gateway (Logical Router)

Provides North-South connectivity so that workloads can access external networks, and East-West connectivity between logical networks.

A Logical Router 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. Logical Routers perform a subset of the tasks that are handled by the physical router, and each can contain multiple

	routing instances and routing tables. Using logical routers can be an effective way to maximize router use, because a set of logical routers within a single physical router can perform the operations previously performed by several pieces of equipment. ■ Distributed router (DR) A DR spans ESXi hosts whose virtual machines 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 gateways connected to this gateway. ■ One or more (optional) service routers (SR). An SR is responsible for delivering services that are not currently implemented in a distributed fashion, such as stateful NAT. A gateway always has a DR. A gateway has SRs when it is a Tier-0 gateway, or when it is a Tier-1 gateway and has routing services configured such as NAT or DHCP.
NSX-T Edge Node	Provides routing services and connectivity to networks that are external to the NSX-T domain through a Tier-0 gateway over BGP or static routing. You must deploy an NSX-T Edge for stateful services at either the Tier-0 or Tier-1 gateways.
NSX-T Edge Cluster	Represents a collection of NSX-T Edge nodes that host multiple service routers in highly available configurations. At a minimum, deploy a single Tier-0 SR to provide external connectivity. An NSX-T Edge cluster does not have a one-to-one relationship with a vSphere cluster. A vSphere cluster can run multiple NSX-T Edge clusters.
Transport Node	Participates in NSX-T overlay or NSX-T VLAN networking. If a node contains an NSX-T Virtual Distributed Switch (N-VDS) such as ESXi hosts and NSX-T Edge nodes, it can be a transport node. If an ESXi host contains at least one N-VDS, it can be a transport node.
Transport Zone	A transport zone can span one or more vSphere clusters. Transport zones dictate which ESXi hosts and which virtual machines can participate in the use of a particular network. A transport zone defines a collection of ESXi hosts that can communicate with each other across a physical network infrastructure. This communication happens over one or more interfaces defined as Tunnel Endpoints (TEPs).

	When you create an ESXi host transport node and then add the node to a transport zone, NSX-T installs an N-VDS on the host. For each transport zone that the host belongs to, a separate N-VDS is installed. The N-VDS is used for attaching virtual machines to NSX-T Segments and for creating NSX-T gateway uplinks and downlinks.
NSX-T Controller	As a component of the control plane, the controllers control virtual networks and overlay transport tunnels. For stability and reliability of data transport, NSX-T deploys the NSX-T Controller as a role in the Manager cluster which consists of three highly available virtual appliances. They are responsible for the programmatic deployment of virtual networks across the entire NSX-T architecture.
Logical Firewall	Responsible for traffic handling in and out the network according to firewall rules. A logical firewall offers multiple sets of configurable Layer 3 and Layer 2 rules. Layer 2 firewall rules are processed before Layer 3 rules. You can configure an exclusion list to exclude segments, logical ports, or groups from firewall enforcement. The default rule, located at the bottom of the rule table, is a catch-all rule. The logical firewall enforces the default rule on packets that do not match other rules. After the host preparation operation, the default rule is set to the allow action. Change this default rule to a block action and enforce access control through a positive control model, that is, only traffic defined in a firewall rule can flow on the network.
Logical Load Balancer	Provides high-availability service for applications and distributes the network traffic load among multiple servers. The load balancer accepts TCP, UDP, HTTP, or HTTPS requests on the virtual IP address and determines which pool server to use. Logical load balancer is supported only in an SR on the Tier-1 gateway.

Detailed Design

3

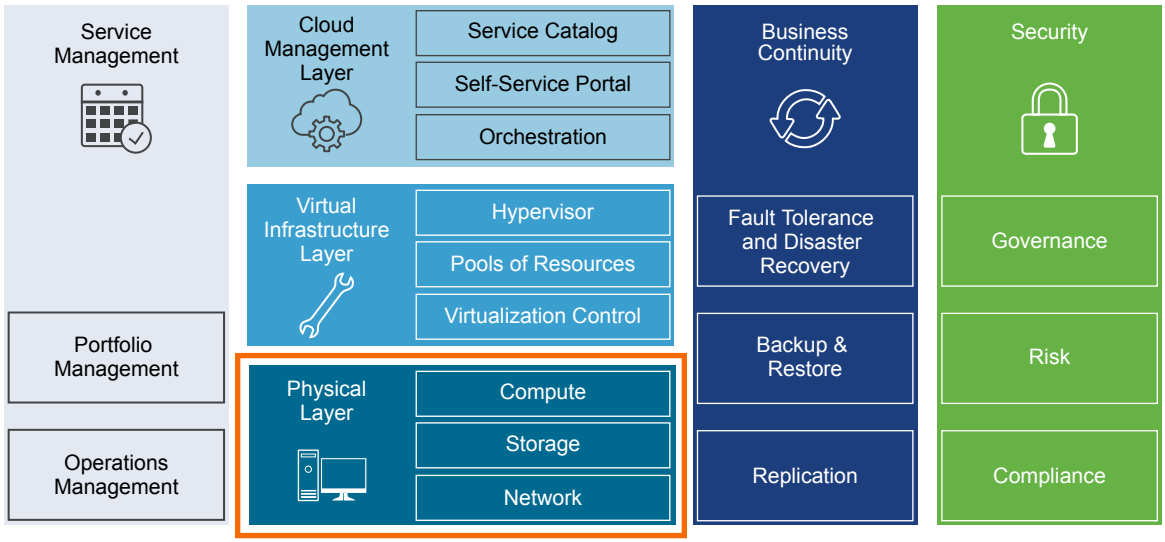
The NSX-T detailed design considers both physical and virtual infrastructure design. It includes numbered design decisions and the justification and implications of each decision.

This chapter includes the following topics:

Physical Infrastructure Design

The physical infrastructure design includes design decision details for the physical network.

Figure 3-1. Physical Infrastructure Design



Physical Networking Design

Design of the physical SDDC network includes defining the network topology for connecting the physical switches and the ESXi hosts, determining switch port settings for VLANs and link aggregation, and designing routing. You can use the VMware Validated Design guidance for design and deployment with most enterprise-grade physical network architectures.

Switch Types and Network Connectivity

Follow the best practices for physical switches, switch connectivity, VLANs and subnets, and access port settings.

Top of Rack Physical Switches

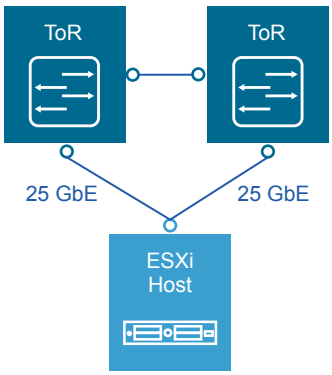
When configuring top of rack (ToR) switches, consider the following best practices:

- Configure redundant physical switches to enhance availability.
- Configure switch ports that connect to ESXi hosts manually as trunk ports. Virtual switches are passive devices and do not 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 to transition ports over to the forwarding state, for example using the Trunk PortFast feature found in a Cisco physical switch.
- Provide DHCP or DHCP Helper capabilities on all VLANs used by TEP VMkernel ports. This setup simplifies the configuration by using DHCP to assign IP address based on the IP subnet in use.
- Configure jumbo frames on all switch ports, inter-switch link (ISL), and switched virtual interfaces (SVIs).

Top of Rack Connectivity and Network Settings

Each ESXi host is connected redundantly to the ToR switches SDDC network fabric by two 25 GbE ports. Configure the ToR switches to provide all necessary VLANs using an 802.1Q trunk. These redundant connections use features in vSphere Distributed Switch and NSX-T to guarantee that no physical interface is overrun and available redundant paths are used.

Figure 3-2. Host to ToR Connectivity



VLANs and Subnets

Each ESXi host uses VLANs and corresponding subnets.

Follow these guidelines:

- Use only /24 subnets to reduce confusion and mistakes when handling IPv4 subnetting.

- Use the IP address .254 as the (floating) interface with .252 and .253 for Virtual Router Redundancy Protocol (VRPP) or Hot Standby Routing Protocol (HSRP).
- Use the RFC1918 IPv4 address space for these subnets and allocate one octet by region and another octet by function.

Access Port Network Settings

Configure additional network settings on the access ports that connect the ToR switches to the corresponding servers.

Spanning Tree Protocol (STP)	Although this design does not use the Spanning Tree Protocol, switches usually include STP configured by default. Designate the access ports as trunk PortFast.
Trunking	Configure the VLANs as members of a 802.1Q trunk with the management VLAN acting as the native VLAN.
MTU	Set MTU for all VLANs and SVIs (Management, vMotion, VXLAN, and Storage) to jumbo frames for consistency purposes.
DHCP Helper	Configure a DHCP helper (sometimes called a DHCP relay) on all TEP VLANs.

Physical Network Design Decisions

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

Physical Network Design Decisions

Routing protocols	NSX-T supports only the BGP routing protocol.
DHCP Helper	Set the DHCP helper (relay) to point to a DHCP server by IPv4 address.

Table 3-1. Physical Network Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-PHY-NET-001	Implement the following physical network architecture: ■ One 25 GbE (10 GbE minimum) port on each ToR switch for ESXi host uplinks. ■ No EtherChannel (LAG/LACP/vPC) configuration for ESXi host uplinks ■ Layer 3 device that supports BGP.	■ Guarantees availability during a switch failure. ■ Uses BGP as the only dynamic routing protocol that is supported by NSX-T.	■ Might limit the hardware choice. ■ Requires dynamic routing protocol configuration in the physical network.
NSXT-PHY-NET-002	Use a physical network that is configured for BGP routing adjacency.	■ Supports flexibility in network design for routing multi-site and multi-tenancy workloads. ■ Uses BGP as the only dynamic routing protocol that is supported by NSX-T.	Requires BGP configuration in the physical network.
NSXT-PHY-NET-003	Use two ToR switches for each rack.	Supports the use of two 10 GbE (25 GbE recommended) links to each server and provides redundancy and reduces the overall design complexity.	Requires two ToR switches per rack which can increase costs.
NSXT-PHY-NET-004	Use VLANs to segment physical network functions.	■ Supports physical network connectivity without requiring many NICs. ■ Isolates the different network functions of the SDDC so that you can have differentiated services and prioritized traffic as needed.	Requires uniform configuration and presentation on all the trunks made available to the ESXi hosts.

Additional Design Decisions

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

Table 3-2. IP Assignment, DNS, and NTP Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-PHY-NET-005	Assign static IP addresses to all management components in the SDDC infrastructure except for NSX-T TEPs. NSX-T TEPs are assigned by using a DHCP server. Set the lease duration for the TEP DHCP scope to at least 7 days.	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. NSX-T TEPs do not have an administrative endpoint. As a result, they can use DHCP for automatic IP address assignment. You are also unable to assign directly a static IP address to the VMkernel port of an NSX-T TEP. IP pools are an option but the NSX-T administrator must create them. If you must change or expand the subnet, changing the DHCP scope is simpler than creating an IP pool and assigning it to the ESXi hosts.	Requires accurate IP address management.
NSXT-PHY-NET-006	Create DNS records for all management nodes to enable forward, reverse, short, and FQDN resolution.	Ensures consistent resolution of management nodes using both IP address (reverse lookup) and name resolution.	None.
NSXT-PHY-NET-007	Use an NTP time source for all management nodes.	Maintains accurate and synchronized time between management nodes.	None.

Jumbo Frames Design Decisions

IP storage throughput can benefit from the configuration of jumbo frames. Increasing the per-frame payload from 1500 bytes to the jumbo frame setting improves the efficiency of data transfer. You must configure jumbo frames end-to-end. Select an MTU that matches the MTU of the physical switch ports.

According to the purpose of the workload, determine whether to configure jumbo frames on a virtual machine. If the workload consistently transfers large amounts of network data, configure jumbo frames, if possible. In that case, confirm that both the virtual machine operating system and the virtual machine NICs support jumbo frames.

Using jumbo frames also improves the performance of vSphere vMotion.

Note The Geneve overlay requires an MTU value of 1600 bytes or greater.

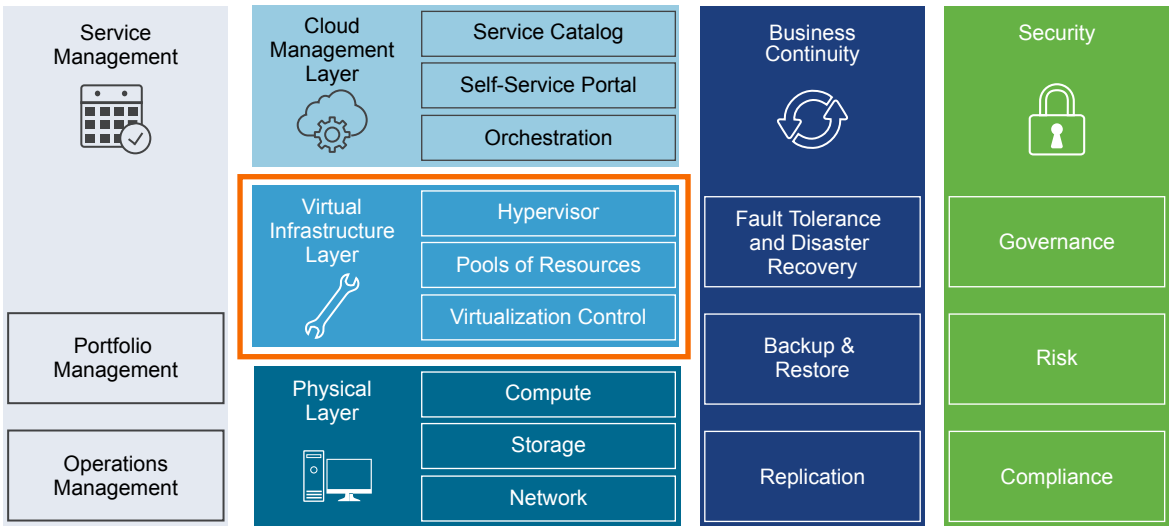
Table 3-3. Jumbo Frames Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-PHY-NET-008	Configure the MTU size to at least 9000 bytes (jumbo frames) on the physical switch ports, vSphere Distributed Switches, vSphere Distributed Switch port groups, and N-VDS switches that support the following traffic types. ■ Geneve (overlay) ■ vSAN ■ vMotion ■ NFS ■ vSphere Replication	Improves traffic throughput. To support Geneve, increase the MTU setting to a minimum of 1600 bytes.	When adjusting the MTU packet size, you must also configure the entire network path (VMkernel ports, virtual switches, physical switches, and routers) to support the same MTU packet size.

Virtual Infrastructure Design

The virtual infrastructure design includes the NSX-T components that make up the virtual infrastructure layer.

Figure 3-3. Virtual Infrastructure Layer in the SDDC



vSphere Cluster Design

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

vSphere Cluster Design Decision Background

When you design the cluster layout in vSphere, consider the following guidelines:

- Use fewer, larger ESXi hosts, or more, smaller ESXi hosts.
  - A scale-up cluster has fewer, larger ESXi hosts.
  - A scale-out cluster has more, smaller 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 of 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 Design

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

vSphere HA Design Basics

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

The master 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 virtual machine OS availability

Table 3-4. vSphere HA Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-VC-001	Use vSphere HA to protect all clusters against failures.	vSphere HA supports a robust level of protection for both ESXi host and virtual machine availability.	You must provide sufficient resources on the remaining hosts so that virtual machines can be migrated to those hosts in the event of a host outage.
NSXT-VI-VC-002	Set vSphere HA Host Isolation Response to Power Off.	vSAN requires that the HA Isolation Response be set to Power Off and to restart VMs on available ESXi hosts.	VMs are powered off in case of a false positive and an ESXi host is declared isolated incorrectly.

vSphere HA Admission Control Policy Configuration

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

The following policies are available:

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

Percentage of cluster resources reserved — vSphere HA reserves a specified percentage of aggregate CPU and memory resources for failover.

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

Shared Edge and Compute Cluster Design

Tenant workloads run on the ESXi hosts in the shared edge and compute cluster. Because of the shared nature of the cluster, NSX-T Edge appliances also run in this cluster. To support these workloads, you must determine the number of ESXi hosts and vSphere HA settings and several other characteristics of the shared edge and compute cluster.

Table 3-5. Shared Edge and Compute Cluster Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-VC-003	Create a shared edge and compute cluster that contains tenant workloads and NSX-T Edge appliances.	Limits the footprint of the design by saving the use of a vSphere cluster specifically for the NSX-T Edge nodes.	In a shared cluster, the VLANs and subnets between the VMkernel ports for ESXi host overlay and the overlay ports of the edge appliances must be separate.
NSXT-VI-VC-004	Configure admission control for a failure of one ESXi host and percentage-based failover capacity.	vSphere HA protects the tenant workloads and NSX-T Edge appliances in the event of an ESXi host failure. vSphere HA powers on the virtual machines from the non-responding ESXi hosts on the remaining ESXi hosts.	Only a single ESXi host failure is tolerated before a resource contention occurs.
NSXT-VI-VC-005	Create a shared edge and compute cluster that consists of a minimum of four ESXi hosts.	Allocating four ESXi hosts provides a full redundancy within the cluster.	Four ESXi hosts is the smallest starting point for the shared edge and compute cluster for redundancy and performance as a result increasing cost.

Table 3-5. Shared Edge and Compute Cluster Design Decisions (Continued)

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-VC-006	Create a resource pool for the two large sized edge virtual machines with a CPU share level of High, a memory share of normal, and a 32-GB memory reservation.	The NSX-T Edge appliances control all network traffic in and out of the SDDC. In a contention situation, these appliances must receive all the resources required.	During contention, the NSX-T components receive more resources than the other workloads. As a result, monitoring and capacity management must be a proactive activity. The resource pool memory reservation must be expanded if you plan to deploy more NSX-T Edge appliances.
NSXT-VI-VC-007	Create a resource pool for all tenant workloads with a CPU share value of Normal and a memory share value of Normal.	Running virtual machines at the cluster level has a negative impact on all other virtual machines during contention. To avoid an impact on network connectivity, in a shared edge and compute cluster, the NSX-T Edge appliances must receive resources with priority to the other workloads. Setting the share values to Normal increases the resource shares of the NSX-T Edge appliances in the cluster.	During contention, tenant workloads might have insufficient resources and have poor performance. Proactively perform monitoring and capacity management, add capacity or dedicate an edge cluster before contention occurs.
NSXT-VI-VC-008	Create a host profile for the shared edge and compute cluster.	Using host profiles simplifies the configuration of ESXi hosts and ensures that settings are uniform across the cluster.	The host profile is only useful for initial cluster deployment and configuration. After you add the ESXi hosts as transport nodes to the NSX-T deployment, the host profile is no longer usable and you must remove them from the cluster.

Compute Cluster Design

As the SDDC expands, you can add compute-only clusters. Tenant workloads run on the ESXi hosts in the compute cluster instances. One Compute vCenter Server instance manages multiple compute clusters. The design determines vSphere HA settings for the compute cluster.

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-VC-009	Configure vSphere HA to use percentage-based failover capacity to ensure n+1 availability.	Using explicit host failover limits the total available resources in a cluster.	The resources of one ESXi host in the cluster is reserved which can cause provisioning to fail if resources are exhausted.

Virtualization Network Design

Design the virtualization network according to the business goals of your organization. Prevent also unauthorized access, and provide timely access to business data.

This network virtualization design uses vSphere and NSX-T to implement virtual networking.

Virtual Network Design Guidelines

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

Design Goals

You can apply the following high-level design goals to your environment:

- Meet diverse needs. The network must meet the diverse needs of many different entities in an organization. These entities include applications, services, storage, administrators, and users.
- Reduce costs. Server consolidation alone reduces network costs by reducing the number of required network ports and NICs, but you should determine a more efficient network design. For example, configuring two 25 GbE NICs with VLANs might be more cost effective than configuring a dozen 1-GbE NICs on separate physical networks.
- Boost performance. You can achieve performance improvements and decrease the time required to perform maintenance by providing sufficient bandwidth, which reduces contention and latency.
- Improve availability. You usually improve availability by providing network redundancy.
- Support security. You can support an acceptable level of security through controlled access where required and isolation where necessary.
- Improve infrastructure functionality. You can configure the network to support vSphere features such as vSphere vMotion, vSphere High Availability, and vSphere Fault Tolerance.

Best Practices

Follow the networking best practices throughout your environment.

- Separate network services from one another for greater security and better performance.
- Use Network I/O Control and traffic shaping to guarantee bandwidth to critical virtual machines. During network contention, these critical virtual machines receive a higher percentage of the bandwidth.
- Separate network services on an NSX-T Virtual Distributed Switch (N-VDS) by attaching them to segments with different VLAN IDs.
- Keep vSphere vMotion traffic on a separate network. When migration with 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.
- When using pass-through devices with Linux kernel version 2.6.20 or an earlier guest OS, avoid MSI and MSI-X modes. These modes have significant performance impact.
- For best performance, use VMXNET3 virtual machine NICs.
- Ensure that physical network adapters connected to the same virtual switch are also connected to the same physical network.

Network Segmentation and VLANs

You separate different types of traffic for access security and to reduce contention and latency.

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

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

Virtual Networks

Determine the number of networks or VLANs that are required according to the type of traffic.

- vSphere operational traffic.
  - Management
  - Geneve (overlay)
  - vMotion
  - vSAN
  - NFS Storage
  - vSphere Replication
- Traffic that supports the services and applications of the organization.

Virtual Switches

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

Virtual Switch Design Background

vSphere Distributed Switch and NSX-T Virtual Distributed Switch (N-VDS) provide several advantages over vSphere Standard Switch.

Centralized management	■ A distributed switch is created and centrally managed on a vCenter Server system. The switch configuration is consistent across ESXi hosts.

- An N-VDS is created and centrally managed in NSX-T Manager. The switch configuration is consistent across ESXi and edge transport nodes.

Centralized management saves time and reduces mistakes and operational costs.

Additional features

Some of the features of distributed switches can be useful to the applications and services running in the organization's infrastructure. For example, NetFlow and port mirroring provide monitoring and troubleshooting capabilities to the virtual infrastructure.

Consider the following caveats for distributed switches:

- Distributed switches are manageable only when the vCenter Server instance is available. As a result, vCenter Server becomes a Tier-1 application.
- N-VDS instances are manageable only when the NSX-T Manager cluster is available. As a result, the NSX-T Manager cluster becomes a Tier-1 application.

Virtual Switch Design Decisions

The virtual switch design decisions determine the use and placement of specific switch types.

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-NET-001	Use N-VDS for the NSX-T based shared edge and compute cluster, and for additional NSX-T based compute clusters.	You need N-VDS for overlay traffic.	N-VDS is not compatible with vSphere host profiles.

Shared Edge and Compute Cluster Switches

The shared edge and compute cluster uses a single N-VDS with a certain configuration for handled traffic types, NIC teaming, and MTU size.

Figure 3-4. Virtual Switch Design for ESXi Hosts in the Shared Edge and Compute Cluster

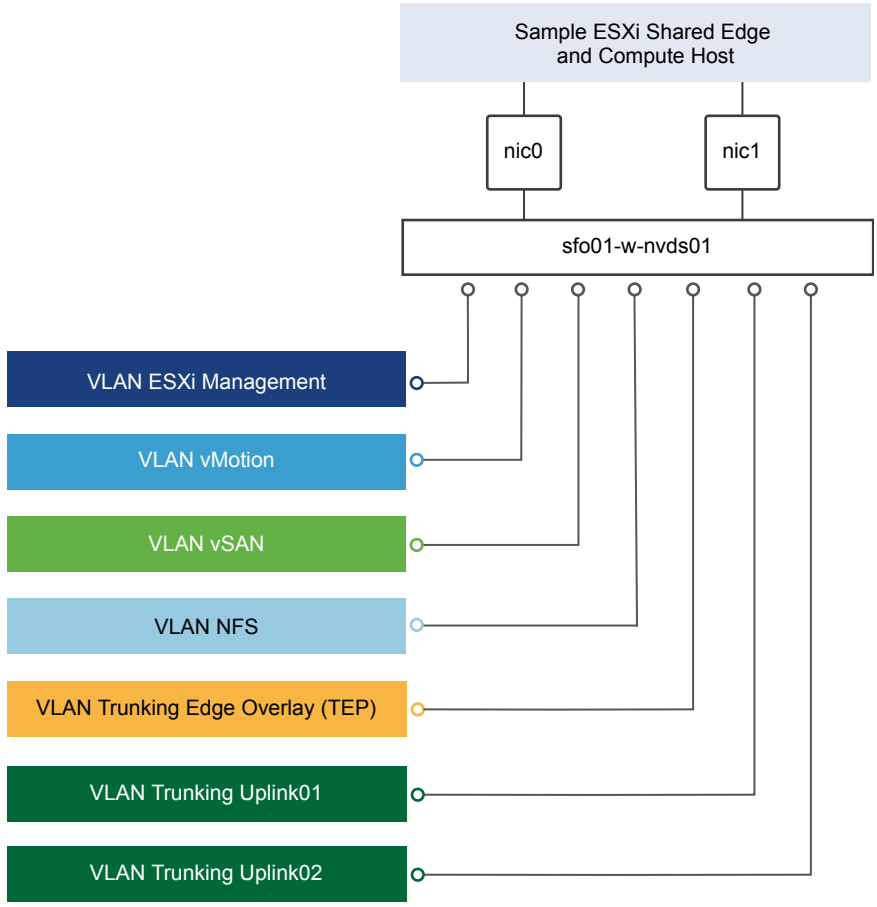


Table 3-6. Virtual Switches for the Shared Edge and Compute Cluster

N-VDS Switch Name	Function	Number of Physical NIC Ports	Teaming Policy	MTU
sfo01-w-nvds01	■ ESXi Management ■ vSphere vMotion ■ vSAN ■ NFS ■ Geneve Overlay (TEP) ■ Uplink trunking (2) for the NSX-T Edge instances	2	■ Load balance source for ESXi traffic ■ Failover order for edge VM traffic	9000
sfo01-w-uplink01	■ Uplink to enable ECMP	1	Failover order	9000
sfo01-w-uplink02	■ Uplink to enable ECMP	1	Failover order	9000

Table 3-7. Virtual Switches in the Shared Edge and Compute Cluster by Physical NIC

N-VDS Switch	vmnic	Function
sfo01-w-nvds01	0	Uplink
sfo01-w-nvds01	1	Uplink

Figure 3-5. Segment Configuration on an ESXi Host That Runs an NSX-T Edge Node

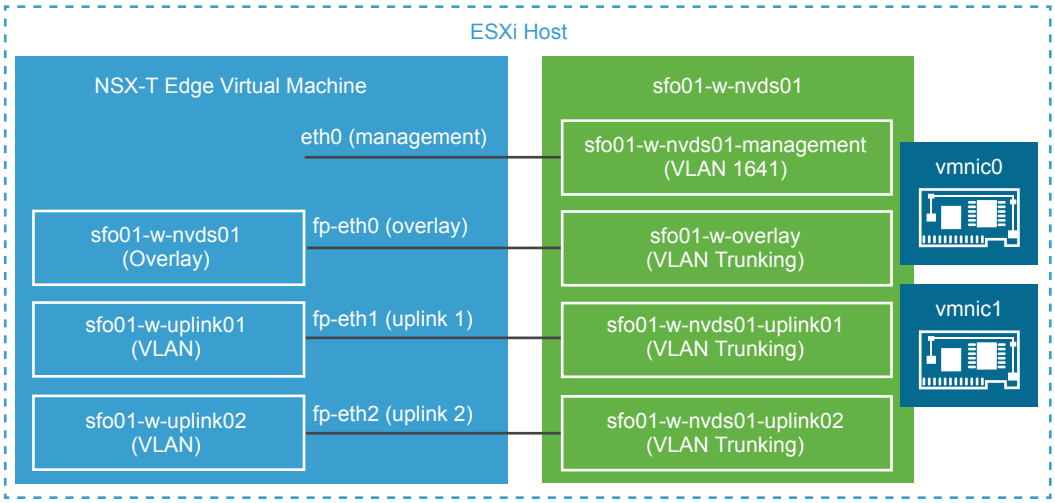


Table 3-8. Segments in the Shared Edge and Compute Cluster

N-VDS Switch	Segment Name
sfo01-w-nvds01	sfo01-w-nvds01-management
sfo01-w-nvds01	sfo01-w-nvds01-vmotion
sfo01-w-nvds01	sfo01-w-nvds01-vsan
sfo01-w-nvds01	sfo01-w-nvds01-overlay
sfo01-w-nvds01	sfo01-w-nvds01-uplink01
sfo01-w-nvds01	sfo01-w-nvds01-uplink02
sfo01-w-nvds01	sfo01-w-nvds01-nfs
sfo01-w02-uplink01	sfo01-w02-uplink01
sfo01-w02-uplink02	sfo01-w02-uplink02

Table 3-9. VMkernel Adapters for the Shared Edge and Compute Cluster

N-VDS Switch	Segment Name	Enabled Services
sfo01-w-nvds01	sfo01-w-nvds01-management	Management Traffic
sfo01-w-nvds01	sfo01-w-nvds01-vmotion	vMotion Traffic
sfo01-w-nvds01	sfo01-w-nvds01-vsan	vSAN
sfo01-w-nvds01	sfo01-w-nvds01-nfs	--

Note

When the NSX-T Edge appliance is on an N-VDS, it must use a different VLAN ID and subnet from the ESXi hosts overlay (TEP) VLAN ID and subnet.

ESXi host TEP VMkernel ports are automatically created when you configure an ESXi host as a transport node.

NIC Teaming

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

Benefits and Overview

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

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

NIC Teaming Design Background

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

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

This validated design uses a non-LAG active-active configuration using the route based on physical NIC load algorithm for vSphere Distributed Switch and load balance source algorithm for N-VDS. By using this configuration, network cards remain active instead of remaining idle until a failure occurs.

Table 3-10. NIC Teaming and Policy

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

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

Table 3-11. NIC Teaming Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-NET-002	In the shared edge and compute cluster, use the Load balance source teaming policy on N-VDS.	NSX-T Virtual Distributed Switch(N-VDS) supports Load balance source and Failover teaming policies. When you use the Load balance source policy, both physical NICs can be active and carry traffic.	None.

Geneve Overlay

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

The first step in creating these logical networks is to isolate and pool the networking resources. By using the Geneve overlay, NSX-T isolates the network into a pool of capacity and separates the consumption of these services from the underlying physical infrastructure. This model is similar to the model vSphere uses to abstract compute capacity from the server hardware to create virtual pools of resources that can be consumed as a service. You can then organize the pool of network capacity in logical networks that are directly attached to specific applications.

Geneve is a tunneling mechanism which provides extensibility while still using the offload capabilities of NICs for performance improvement.

Geneve works by creating Layer 2 logical networks that are encapsulated in UDP packets. A Segment ID in every frame identifies the Geneve logical networks without the need for VLAN tags. As a result, many isolated Layer 2 networks can coexist on a common Layer 3 infrastructure using the same VLAN ID.

In the vSphere architecture, the encapsulation is performed between the virtual NIC of the guest VM and the logical port on the virtual switch, making the Geneve overlay transparent to both the guest virtual machines and the underlying Layer 3 network. The Tier-0 Gateway performs gateway services between overlay and non-overlay hosts, for example, a physical server or the Internet router. The NSX-T Edge virtual machine translates overlay segment IDs to VLAN IDs, so that non-overlay hosts can communicate with virtual machines on an overlay network.

The edge cluster hosts all NSX-T Edge virtual machine instances that connect to the corporate network for secure and centralized network administration.

Table 3-12. Geneve Overlay Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-NET-003	Use NSX-T to introduce overlay networks for workloads.	Simplifies the network configuration by using centralized virtual network management.	■ Requires additional compute and storage resources to deploy NSX-T components. ■ Might require more training in NSX-T.
NSXT-VI-NET-004	To provide virtualized network capabilities to workloads, use overlay networks with NSX-T Edge virtual machines and distributed routing.	Creates isolated, multi-tenant broadcast domains across data center fabrics to deploy elastic, logical networks that span physical network boundaries.	Requires configuring transport networks with an MTU size of at least 1600 bytes.

vMotion TCP/IP Stack

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

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

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

Table 3-13. vMotion TCP/IP Stack Design Decision

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-NET-005	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.	The vMotion TCP/IP stack is not available in the VMkernel adapter creation wizard of vSphere Distributed Switch. You must create the VMkernel adapter directly on the ESXi host.

NSX Design

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

In much the same way that server virtualization programmatically creates, takes snapshots of, deletes, and restores software-based virtual machines (VMs), NSX 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 is a nondisruptive solution. You can deploy it on any IP network, including existing traditional networking models and next-generation fabric architectures, regardless of the vendor.

When administrators provision workloads, network management is a time-consuming task. You spend most time configuring individual components in the physical infrastructure and verifying that network changes do not affect other devices that are using the same physical network infrastructure.

The need to pre-provision and configure networks is a constraint to cloud deployments where speed, agility, and flexibility are critical requirements. Pre-provisioned physical networks enable fast creation of virtual networks and faster deployment times of workloads using the virtual network. If the physical network that you need is already available on the ESXi host to run a workload, pre-provisioning physical networks works well. However, if the network is not available on an ESXi host, you must find an ESXi host with the available network and allocate capacity to run workloads in your environment.

Decouple virtual networks from their physical counterparts. In the virtualized environment, you must recreate all physical networking attributes that are required by the workloads. Because network virtualization supports the creation of virtual networks without modification of the physical network infrastructure, you can provision the workload networks faster.

NSX-T Design

NSX-T components are not dedicated to a specific vCenter Server instance or vSphere construct. You can share them across vSphere environments in the same physical location.

Although an NSX-T deployment is not associated with a vCenter Server instance, it supports only single-region deployments.

Table 3-14. NSX-T Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-001	Deploy a three node NSX-T Manager cluster to configure and manage the shared edge and compute cluster and all additional compute clusters in a single region.	Provides support for software-defined networking (SDN) capabilities, such as load balancing, firewalls, and logical switching.	You must install and configure NSX-T Manager in a highly available cluster.
NSXT-VI-SDN-002	In the management cluster, add the NSX-T Manager nodes to the distributed firewall exclusion list of NSX for vSphere.	Ensures that the management and control plane is still available if a mistake in the configuration of the distributed firewall of NSX for vSphere occurs.	None.

NSX-T Components

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

NSX-T Manager

NSX-T Manager provides the graphical user interface (GUI) and the RESTful API for creating, configuring, and monitoring NSX-T components, such as segments and gateways.

NSX-T Manager implements the management and control plane for the NSX-T infrastructure. NSX-T Manager provides an aggregated system view and is the centralized network management component of NSX-T. It provides a method for monitoring and troubleshooting workloads attached to virtual networks. It provides 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

NSX-T Manager also provides 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.

The NSX-T Management Plane Agent (MPA) is an NSX-T Manager component that is available on each ESXi host. The MPA is in charge of persisting the desired state of the system and for communicating non-flow-controlling (NFC) messages such as configuration, statistics, status, and real-time data between transport nodes and the management plane.

NSX-T Manager also contains the NSX-T Controller component. NSX-T Controllers control the virtual networks and overlay transport tunnels. The controllers are responsible for the programmatic deployment of virtual networks across the entire NSX-T architecture.

The Central Control Plane (CCP) is logically separated from all data plane traffic, that is, a failure in the control plane does not affect existing data plane operations. The controller provides configuration to other NSX-T Controller components such as the segments, gateways, and edge virtual machine configuration.

Table 3-15. NSX-T Manager Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-003	Deploy a three node NSX-T Manager cluster using the large-size appliance.	The large-size appliance supports more than 64 ESXi hosts. The small-size appliance is for proof of concept and the medium size only supports up to 64 ESXi hosts.	The large size requires more resources in the management cluster.
NSXT-VI-SDN-004	Create a virtual IP (VIP) for the NSX-T Manager cluster.	Provides HA for the NSX-T Manager UI and API.	The VIP provides HA only, it does not load balance requests across the manager cluster.
NSXT-VI-SDN-005	■ Grant administrators access to both the NSX-T Manager UI and its RESTful API endpoint. ■ Restrict end-user access to the RESTful API endpoint configured for end-user provisioning, such as vRealize Automation or Pivotal Container Service (PKS).	Ensures that tenants or non-provider staff cannot modify infrastructure components. End-users typically interact only indirectly with NSX-T from their provisioning portal. Administrators interact with NSX-T using its UI and API.	End users have access only to end-point components.

NSX-T Virtual Distributed Switch

An NSX-T Virtual Distributed Switch (N-VDS) runs on ESXi hosts and provides physical traffic forwarding. It transparently 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 that form the logical broadcast domains the tenant administrators define when they create and configure segments.

NSX-T 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 ESXi host virtual switch flow tables are updated as VMs are created, moved, or removed.

Table 3-16. NSX-T N-VDS Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-006	Deploy an N-VDS instance to each ESXi host in the shared edge and compute cluster.	ESXi hosts in the shared edge and compute cluster provide tunnel endpoints for Geneve overlay encapsulation.	None.

Logical Switching

NSX-T Segments create logically abstracted segments to which you can connect tenant workloads. A single Segment is mapped to a unique Geneve segment that is distributed across the ESXi hosts in a transport zone. The Segment supports line-rate switching in the ESXi host without the constraints of VLAN sprawl or spanning tree issues.

Table 3-17. NSX-T Logical Switching Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-007	Deploy all workloads on NSX-T Segments (logical switches).	To take advantage of features such as distributed routing, tenant workloads must be connected to NSX-T Segments.	You must perform all network monitoring in the NSX-T Manager UI or vRealize Network Insight.

Gateways (Logical Routers)

NSX-T Gateways provide North-South connectivity so that workloads can access external networks, and East-West connectivity between different logical networks.

A Logical Router 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. Logical routers perform a subset of the tasks that are handled by the physical router, and each can contain multiple routing instances and routing tables. Using logical routers can be an effective way to maximize router use, because a set of logical routers within a single physical router can perform the operations previously performed by several pieces of equipment.

- Distributed router (DR)

  A DR spans ESXi hosts whose virtual machines 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 Gateways connected to this Gateway.

- One or more (optional) service routers (SR).

  An SR is responsible for delivering services that are not currently implemented in a distributed fashion, such as stateful NAT.

A Gateway always has a DR. A Gateway has SRs when it is a Tier-0 Gateway, or when it is a Tier-1 Gateway and has routing services configured such as NAT or DHCP.

Tunnel Endpoint

Tunnel endpoints enable ESXi hosts to participate in an NSX-T overlay. The NSX-T overlay deploys a Layer 2 network on top of an existing Layer 3 network fabric by encapsulating frames inside packets and transferring the packets over an underlying transport network. The underlying transport network can be another Layer 2 networks or it can cross Layer 3 boundaries. The Tunnel Endpoint (TEP) is the connection point at which the encapsulation and decapsulation take place.

NSX-T Edges

NSX-T Edges provide routing services and connectivity to networks that are external to the NSX-T deployment. You use an NSX-T Edge for establishing external connectivity from the NSX-T domain by using a Tier-0 Gateway using BGP or static routing. Additionally, you deploy an NSX-T Edge to support network address translation (NAT) services at either the Tier-0 or Tier-1 Gateway.

The NSX-T Edge connects isolated, stub networks to shared uplink networks by providing common gateway services such as NAT, and dynamic routing.

Logical Firewall

NSX-T uses handles traffic in and out the network according to firewall rules.

A logical firewall offers multiple sets of configurable Layer 3 and Layer 2 rules. Layer 2 firewall rules are processed before Layer 3 rules. You can configure an exclusion list to exclude segments, logical ports, or groups from firewall enforcement.

The default rule, that is at the bottom of the rule table, is a catchall rule. The logical firewall enforces the default rule on packets that do not match other rules. After the host preparation operation, the default rule is set to the allow action. Change this default rule to a block action and apply access control through a positive control model, that is, only traffic defined in a firewall rule can flow on the network.

Logical Load Balancer

The NSX-T logical load balancer offers high-availability service for applications and distributes the network traffic load among multiple servers.

The load balancer accepts TCP, UDP, HTTP, or HTTPS requests on the virtual IP address and determines which pool server to use.

Logical load balancer is supported only on the Tier-1 Gateway.

NSX-T Network Requirements and Sizing

NSX-T requirements impact both physical and virtual networks.

Physical Network Requirements

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

Requirement	Comments
Provide an MTU size of 1600 or greater on any network that carries Geneve overlay traffic must.	Geneve packets cannot be fragmented. The MTU size must be large enough to support extra encapsulation overhead. This design uses an MTU size of 9000 for Geneve traffic. See Table 3-3.
Enable dynamic routing support on the upstream Layer 3 devices.	You use BGP on the upstream Layer 3 devices to establish routing adjacency with the Tier-0 SRs.
Provide an NTP server.	The NSX-T Manager requires NTP settings that synchronize it with the rest of the environment.
Establish forward and reverse DNS resolution for all management VMs.	The NSX-T Controllers do not require DNS entries.

NSX-T Component Specifications

When you size the resources for NSX-T components, consider the compute and storage requirements for each component, and the number of nodes per component type.

Size of NSX Edge services gateways might be different according to tenant requirements. Consider all options in such a case.

Table 3-18. Resource Specification of the NSX-T Components

Virtual Machine	vCPU	Memory (GB)	Storage (GB)	Quantity per NSX-T Deployment
NSX-T Manager	12 (Large)	48 (Large)	200 (Large)	3
NSX-T Edge virtual machine	2 (Small)	4 (Small)	120 (Small)	Numbers are different according to the use case. At least two edge devices are required to enable ECMP routing.
	4 (Medium)	8 (Medium)	120 (Medium)	
	8 (Large)	16 (Large)	120 (Large)	

Table 3-19. Design Decisions on Sizing the NSX-T Edge Virtual Machines

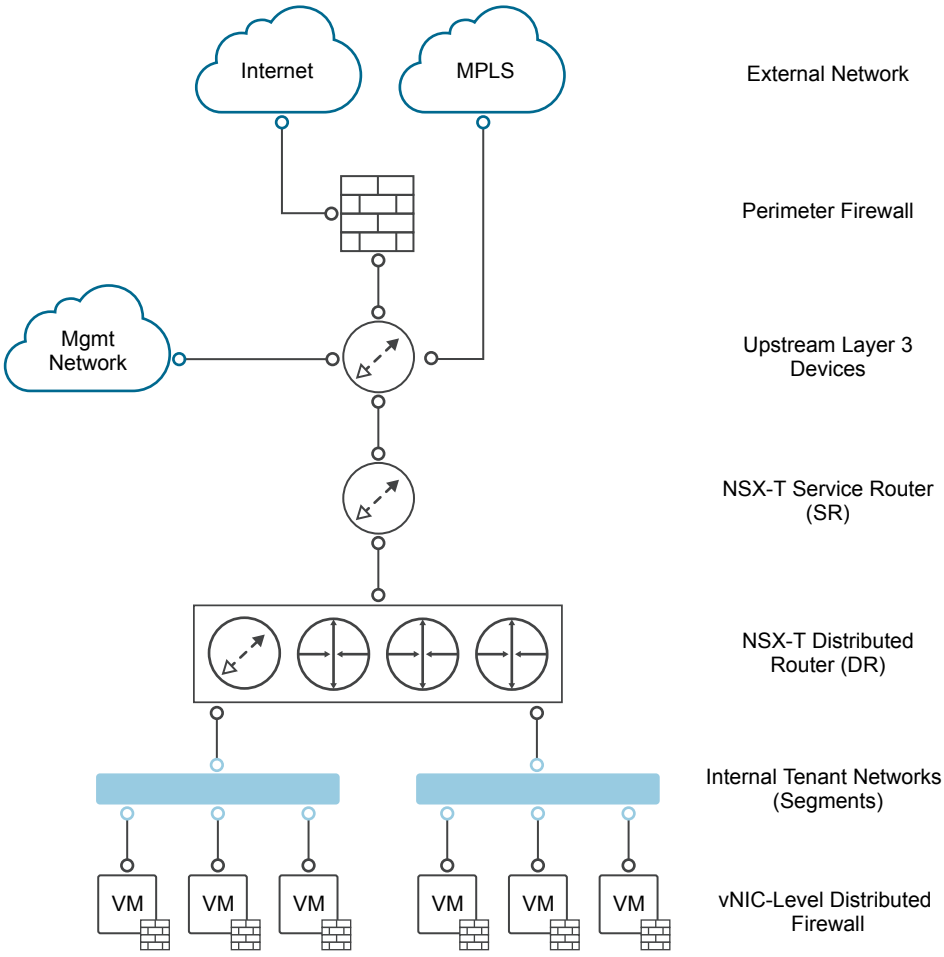
Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-008	Use large-size NSX Edge virtual machines.	The large-size appliance provides all the performance characteristics if a failure occurs.	None.

Network Virtualization Conceptual Design for NSX-T

This conceptual design for NSX-T provides the network virtualization design of the logical components that handle the data to and from tenant workloads in the environment.

The network virtualization conceptual design includes a perimeter firewall, a provider logical router, and the NSX-T Gateway. It also considers the external network, internal workload networks, and the management network.

Figure 3-6. NSX-T Conceptual Overview



The conceptual design has the following components.

External Networks	Connectivity to and from external networks is through the perimeter firewall.
Perimeter Firewall	The firewall exists at the perimeter of the data center to filter Internet traffic.
Upstream Layer 3 Devices	The upstream Layer 3 devices are behind the perimeter firewall and handle North-South traffic that is entering and leaving the NSX-T environment. In most cases, this layer consists of a pair of top of rack switches or redundant upstream Layer 3 devices such as core routers.
NSX-T Gateway (SR)	The SR component of the NSX-T Tier-0 Gateway is responsible for establishing eBGP peering with the Upstream Layer 3 devices and enabling North-South routing.
NSX-T Gateway (DR)	The DR component of the NSX-T Gateway is responsible for East-West routing.

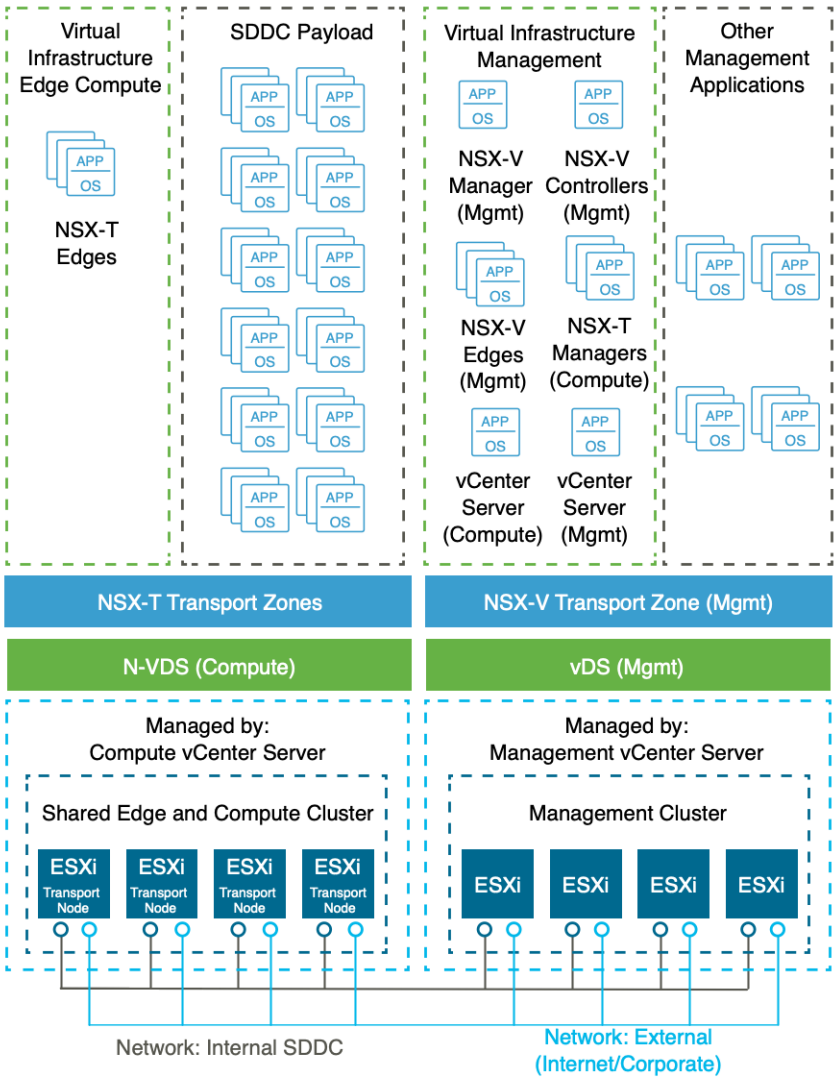
Management Network	The management network is a VLAN-backed network that supports all management components such as NSX-T Manager and NSX-T Controllers.
Internal Workload Networks	Internal workload networks are NSX-T Segments and provide connectivity for the tenant workloads. Workloads are directly connected to these networks. Internal workload networks are then connected to a DR.

Cluster Design for NSX-T

The NSX-T design uses management, and shared edge and compute clusters. You can add more compute clusters for scale-out, or different workload types or SLAs.

The logical NSX-T design considers the vSphere clusters and defines the place where each NSX component runs.

Figure 3-7. NSX-T Cluster Design



Management Cluster

The management cluster contains all components for managing the SDDC. This cluster is a core component of the VMware Validated Design for Software-Defined Data Center. For information about the management cluster design, see the *Architecture and Design* documentation in VMware Validated Design for Software-Defined Data Center.

NSX-T Edge Node Cluster

The NSX-T Edge cluster is a logical grouping of NSX-T Edge virtual machines. These NSX-T Edge virtual machines run in the vSphere shared edge and compute cluster and provide North-South routing for the workloads in the compute clusters.

Shared Edge and Compute Cluster

In the shared edge and compute cluster, ESXi hosts are prepared for NSX-T. As a result, they can be configured as transport nodes and can participate in the overlay network. All tenant workloads, and NSX-T Edge virtual machines run in this cluster.

Table 3-20. Cluster Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-009	For the compute stack, do not dedicate a vSphere edge cluster.	Simplifies configuration and minimizes the number of ESXi hosts required for initial deployment.	The NSX-T Edge virtual machines are deployed in the shared edge and compute cluster. Because of the shared nature of the cluster, you must scale out the cluster as compute workloads are added to avoid an impact on network performance.
NSXT-VI-SDN-010	Deploy at least two large-size NSX-T Edge virtual machines in the shared edge and compute cluster.	Creates NSX-T Edge cluster, and meets availability and scale requirements.	When additional Edge VM's are added, the Resource Pool Memory Reservation must be adjusted.
NSXT-VI-SDN-011	Apply vSphere Distributed Resource Scheduler (vSphere DRS) VM-Host anti-affinity rules to NSX-T Managers.	Prevents managers from running on the same ESXi host and thereby risking their high availability capability.	Requires at least four physical hosts to guarantee the three NSX-T Managers continue to run if an ESXi host failure occurs. Additional configuration is required to set up anti-affinity rules.
NSXT-VI-SDN-012	Apply vSphere DRS VM-Host anti-affinity rules to the virtual machines of the NSX-T Edge cluster.	Prevents the NSX-T Edge virtual machines from running on the same ESXi host and compromising their high availability.	Additional configuration is required to set up anti-affinity rules.

High Availability of NSX-T Components

The NSX-T Managers run on the management cluster. vSphere HA protects the NSX-T Managers by restarting the NSX-T Manager virtual machine on a different ESXi host if a primary ESXi host failure occurs.

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

The NSX-T Edge virtual machines are deployed on the shared edge and compute cluster. vSphere DRS anti-affinity rules prevent NSX-T Edge virtual machines that belong to the same NSX-T Edge cluster from running on the same ESXi host.

NSX-T SRs for North-South routing are configured in equal-cost multi-path (ECMP) mode that supports route failover in seconds.

Replication Mode of Segments

The control plane decouples NSX-T from the physical network, and handles the broadcast, unknown unicast, and multicast (BUM) traffic in the segments (logical switches).

The following options are available for BUM replication on segments.

Table 3-21. BUM Replication Mode of NSX-T Segments

BUM Replication Mode	Description
Hierarchical Two-Tier	In this mode, the ESXi host transport nodes are grouped according to their TEP IP subnet. One ESXi host in each subnet is responsible for replication to a ESXi host in another subnet. The receiving ESXi host replicates the traffic to the ESXi hosts in its local subnet. The source ESXi host transport node knows about the groups based on information it has received from the NSX-T control cluster. The system can select an arbitrary ESXi host transport node as the mediator for the source subnet if the remote mediator ESXi host node is available.
Head-End	In this mode, the ESXi host transport node at the origin of the frame to be flooded on a segment sends a copy to every other ESXi host transport node that is connected to this segment.

Table 3-22. Design Decisions on Segment Replication Mode

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-013	Use hierarchical two-tier replication on all segments.	Hierarchical two-tier replication is more efficient by reducing the number of ESXi hosts the source ESXi host must replicate traffic to.	None.

Transport Zone Design

Transport zones determine which hosts can participate in the use of a particular network. A transport zone identifies the type of traffic, VLAN or overlay, and the N-VDS name. You can configure one or more transport zones. A transport zone does not represent a security boundary.

Table 3-23. Transport Zones Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-014	Create a single transport zone for all overlay traffic.	Ensures all Segments are available to all ESXi hosts and edge virtual machines configured as Transport Nodes.	None.
NSXT-VI-SDN-015	Create a VLAN transport zone for ESXi host VMkernel ports.	Enables the migration of ESXi host VMkernel ports to the N-VDS.	The N-VDS name must match the N-VDS name in the overlay transport zone.
NSXT-VI-SDN-016	Create two transport zones for edge virtual machine uplinks.	Enables the edge virtual machines to use equal-cost multi-path routing (ECMP).	You must specify a VLAN range, that is use VLAN trunking, on the segment used as the uplinks.

Network I/O Control Design for NSX-T

When a Network I/O Control profile is attached to an N-VDS, during contention the switch allocates available bandwidth according to the configured shares, limit, and reservation for each vSphere traffic type.

How Network I/O Control Works

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

Network I/O Control also supports the reservation of bandwidth for system traffic according to the overall percentage of available bandwidth.

Network I/O Control Heuristics

The following heuristics can help with design decisions.

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

Limits on Network Resource Pools — Consider imposing limits on a resource pool. For example, set a limit on vSphere vMotion traffic to avoid oversubscription at the physical network level when multiple vSphere vMotion data transfers are initiated on different ESXi hosts at the same time. By limiting the available bandwidth for vSphere vMotion at the ESXi host level, you can prevent performance degradation for other traffic.

Network I/O Control Design Decisions

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

Table 3-24. Network I/O Control Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-SDN-017	Create and attach a Network I/O Control Policy on all N-DVS switches.	Increases resiliency and performance of the network.	If configured incorrectly, Network I/O Control might impact network performance for critical traffic types.
NSXT-VI-SDN-018	Set the share value for vSphere vMotion traffic to Low (25).	During times of network contention, vSphere vMotion traffic is not as important as virtual machine or storage traffic.	During times of network contention, vMotion takes longer than usual to complete.
NSXT-VI-SDN-019	Set the share value for vSphere Replication traffic to Low (25).	During times of network contention, vSphere Replication traffic is not as important as virtual machine or storage traffic.	During times of network contention, vSphere Replication takes longer and might violate the defined SLA.
NSXT-VI-SDN-020	Set the share value for vSAN traffic to High (100).	During times of network contention, vSAN traffic needs a guaranteed bandwidth to support virtual machine performance.	None.
NSXT-VI-SDN-021	Set the share value for management traffic to Normal (50).	By keeping the default setting of Normal, management traffic is prioritized higher than vSphere vMotion and vSphere Replication but lower than vSAN traffic. Management traffic is important because it ensures that the hosts can still be managed during times of network contention.	None.
NSXT-VI-SDN-022	Set the share value for NFS traffic to Low (25).	Because NFS is used for secondary storage, such as backups and vRealize Log Insight archives, it is not as important as vSAN traffic. By prioritizing it lower, vSAN is not impacted.	During times of network contention, backups are slower than usual.
NSXT-VI-SDN-023	Set the share value for backup traffic to Low (25).	During times of network contention, the primary functions of the SDDC must continue to have access to network resources with priority over backup traffic.	During times of network contention, backups are slower than usual.
NSXT-VI-SDN-024	Set the share value for virtual machines to High (100).	Virtual machines are the most important asset in the SDDC. Leaving the default setting of High ensures that they always have access to the network resources they need.	None.
NSXT-VI-SDN-025	Set the share value for vSphere Fault Tolerance to Low (25).	This design does not use vSphere Fault Tolerance. Fault tolerance traffic can be set the lowest priority.	None.
NSXT-VI-SDN-026	Set the share value for iSCSI traffic to Low (25).	This design does not use iSCSI. iSCSI traffic can be set the lowest priority.	None.

Transport Node and Uplink Policy Design

A transport node can participate in an NSX-T overlay or NSX-T VLAN network.

Several types of transport nodes are available in NSX-T.

ESXi Host Transport Nodes — ESXi host transport nodes are ESXi hosts prepared and configured for NSX-T. N-VDS provides network services to the virtual machines running on these ESXi hosts.

Edge Nodes — NSX-T Edge nodes are service appliances that run network services that cannot be distributed to the hypervisors. They are grouped in one or several NSX-T Edge clusters. Each cluster represents a pool of capacity.

Uplink profiles define policies for the links from ESXi hosts to NSX-T Segments or from NSX-T Edge virtual machines to top of rack switches. By using uplink profiles, you can apply consistent configuration of capabilities for network adapters across multiple ESXi hosts or edge virtual machines. Uplink profiles are containers for the properties or capabilities for the network adapters.

Uplink profiles can use either load balance source or failover order teaming. If using load balance source, multiple uplinks can be active. If using failover order, only a single uplink can be active.

Table 3-25. Design Decisions on Transport Nodes and Uplink Policy

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-027	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 both physical NICs on the ESXi hosts that are configured as transport nodes.	You can use this policy only with ESXi hosts. Edge virtual machines must use the failover order teaming policy.
NSXT-VI-SDN-028	Create an uplink profile with the failover order teaming policy with one active uplink and no standby uplinks for edge virtual machine overlay traffic.	Provides a profile that according to the requirements for Edge virtual machines. Edge virtual machines support uplink profiles only with a failover order teaming policy. VLAN ID is required in the uplink profile. Hence, you must create an uplink profile for each VLAN used by the edge virtual machines.	■ You create and manage more uplink profiles. ■ The VLAN ID used must be different than the VLAN ID for ESXi host overlay traffic.
NSXT-VI-SDN-029	Create two uplink profiles with the failover order teaming policy with one active uplink and no standby uplinks for edge virtual machine uplink traffic.	Enables ECMP because the edge virtual machine can uplink to the physical network over two different VLANs.	You create and manage more uplink profiles.

Table 3-25. Design Decisions on Transport Nodes and Uplink Policy (Continued)

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-030	Create a Transport Node Policy with the VLAN and Overlay Transport Zones, N-VDS settings, Physical NICs and Network mappings to allow NSX-T to migrate the VMKernel adapter to the N-VDS.	Allows the profile to be assigned directly to the vSphere cluster and ensures consistent configuration across all ESXi Host Transport Nodes.	You must create all required Transport Zones and Segments before creating the Transport Node Policy.
NSXT-VI-SDN-031	Add as transport nodes all ESXi hosts in the Shared Edge and Compute cluster by applying the Transport Node Policy to the vSphere cluster object.	Enables the participation of ESXi hosts and the virtual machines on them in NSX-T overlay and VLAN networks.	ESXi hosts VMKernel adapters are migrated to the N-VDS. Ensure the VLAN ID's for the VMKernel Segments are correct to ensure host communication is not lost.
NSXT-VI-SDN-032	Add as transport nodes all edge virtual machines.	Enables the participation of edge virtual machines in the overlay network and the delivery of services, such as routing, by these machines.	None.
NSXT-VI-SDN-033	Create an NSX-T Edge cluster with the default Bidirectional Forwarding Detection (BFD) settings containing the edge transport nodes.	Satisfies the availability requirements by default. Edge clusters are required to create services such as NAT, routing to physical networks, and load balancing.	None.

Routing Design by Using NSX-T

The routing design considers different levels of routing in the environment, such as number and type of NSX-T routers, dynamic routing protocol, and so on. At each level, you apply a set of principles for designing a scalable routing solution.

Routing can be defined in the following directions: North-South and East-West.

- North-South traffic is traffic leaving or entering the NSX-T domain, for example, a virtual machine on an overlay network communicating with an end-user device on the corporate network.
- East-West traffic is traffic that remains in the NSX-T domain, for example, two virtual machines on the same or different segments communicating with each other.

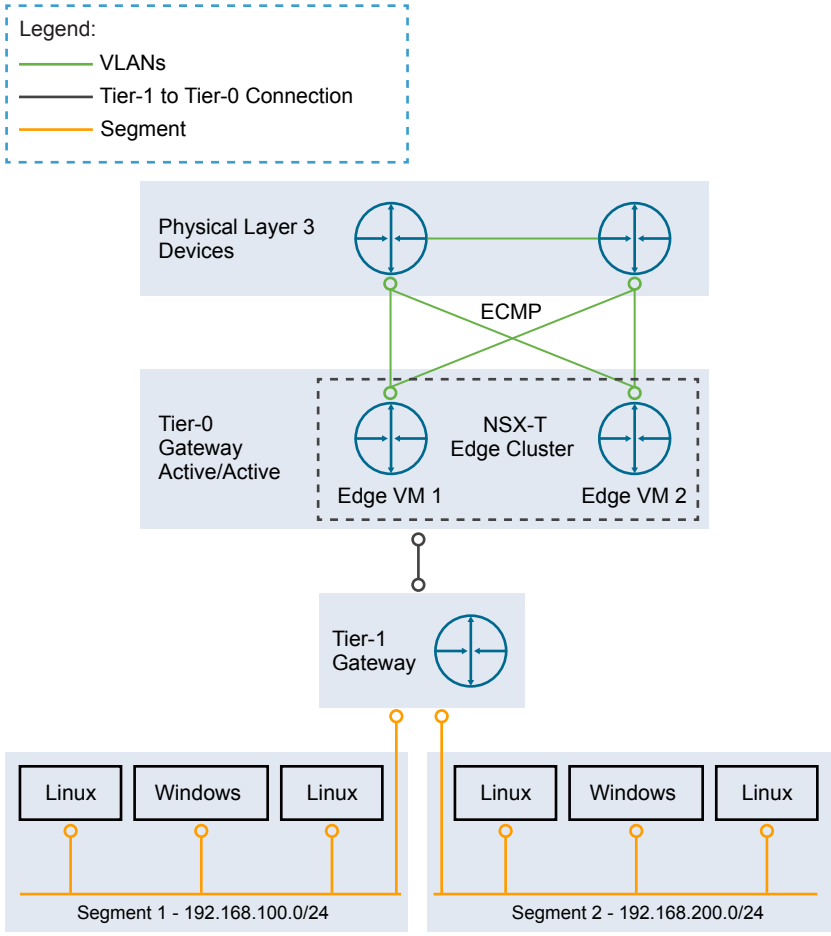
Table 3-26. Design Decisions on Routing Using NSX-T

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-034	Create two VLANs to enable ECMP between the Tier-0 Gateway and the Layer 3 device (ToR or upstream device). The ToR switches or upstream Layer 3 devices have an SVI on one of the two VLANS and each edge virtual machine has an interface on each VLAN.	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.
NSXT-VI-SDN-035	Deploy an Active-Active Tier-0 Gateway.	Supports ECMP North-South routing on all edge virtual machines in the NSX-T 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, such as VMware Pivotal Container Service, you must deploy an additional Tier-0 Gateway in Active-Standby mode.
NSXT-VI-SDN-036	Use BGP as the dynamic routing protocol.	Enables the dynamic routing by using NSX-T. NSX-T supports only BGP .	In environments where BGP cannot be used, you must configure and manage static routes.
NSXT-VI-SDN-037	Configure BGP Keep Alive Timer to 4 and Hold Down Timer to 12 between the ToR switches and the Tier-0 Gateway.	Provides a balance between failure detection between the ToR switches and the Tier-0 Gateway and overburdening the ToRs with keep alive traffic.	By using longer timers to detect if a router is not responding, the data about such a router remains in the routing table longer. As a result, the active router continues to send traffic to a router that is down.
NSXT-VI-SDN-038	Do not enable Graceful Restart between BGP neighbors.	Avoids loss of traffic. Graceful Restart maintains the forwarding table which in turn will forward packets to a down neighbor even after the BGP timers have expired causing loss of traffic.	None.
NSXT-VI-SDN-039	Deploy a Tier-1 Gateway to the NSX-T Edge cluster and connect it to the Tier-0 Gateway.	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.	A Tier-1 Gateway can only be connected to a single Tier-0 Gateway. In scenarios where multiple Tier-0 Gateways are required, you must create multiple Tier-1 Gateways.

Virtual Network Design Example Using NSX-T

Design a setup of virtual networks where you determine the connection of virtual machines to Segments and the routing between the Tier-1 Gateway and Tier-0 Gateway, and then between the Tier-0 Gateway and the physical network.

Figure 3-8. Virtual Network Example



Monitoring NSX-T

Monitor the operation of NSX-T for identifying failures in the network setup. Use the vRealize Log Insight Content Pack for VMware NSX-T to see the logs generated by the components of NSX-T in the user interface of vRealize Log Insight.

vRealize Log Insight saves log queries and alerts, and you can use dashboards for efficient monitoring.

Table 3-27. Design Decisions on Monitoring NSX-T

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-SDN-040	Install the content pack for NSX-T in vRealize Log Insight.	Provides granular monitoring of the NSX-T infrastructure.	Requires manually installing the content pack.
NSXT-VI-SDN-041	Configure each NSX-T component to send log information over syslog to the vRealize Log Insight cluster VIP.	Ensures that all NSX-T components log files are available for monitoring and troubleshooting in vRealize Log Insight.	Requires manually configuring syslog on each NSX-T component.

Use of SSL Certificates in NSX-T

By default, NSX-T Manager uses a self-signed Secure Sockets Layer (SSL) certificate. This certificate is not trusted by end-user devices or Web browsers.

As a best practice, replace self-signed certificates with certificates that are signed by a third-party or enterprise Certificate Authority (CA).

Table 3-28. Design Decisions on the SSL Certificate of NSX-T Manager

Design ID	Design Decision	Design Justification	Design Implication
NSXT-VI-SDN-042	Replace the NSX-T Managers certificate with a certificate that is signed by a third-party Public Key Infrastructure.	Ensures that the communication between NSX-T administrators and the NSX-T Manager is encrypted by using a trusted certificate.	Replacing and managing certificates is an operational overhead.
NSXT-VI-SDN-043	Replace the NSX-T Manager Cluster certificate with a certificate that is signed by a third-party Public Key Infrastructure.	Ensures that the communication between the virtual IP on the NSX-T Manager Cluster and administrators is encrypted by using a trusted certificate.	Replacing and managing certificates is an operational overhead.