VMware NSX-T for Workload Domains

VMware Validated Design 4.2
VMware NSX-T 2.1

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 VMware NSX-T for Workload Domains

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 compute cluster in an SDDC that is compliant with VMware Validated Design for Software-Defined Data Center.

Prerequisites

You must have VMware Validated Design for Software-Defined Data Center 4.2 or later in at least a single-region deployment. 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 4.2 deployment, you must download NSX-T 2.1 or later. You then deploy and configure NSX-T in the compute cluster according to this guide.

Architecture Overview 1

VMware Validated Design for NSX-T enables IT organizations that have deployed VMware Validated Design for Software-Defined Data Center 4.2 or later to create a compute cluster that uses NSX-T capabilities.

This chapter includes the following topics:

- Physical Network Architecture
- Virtual Infrastructure Architecture

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:

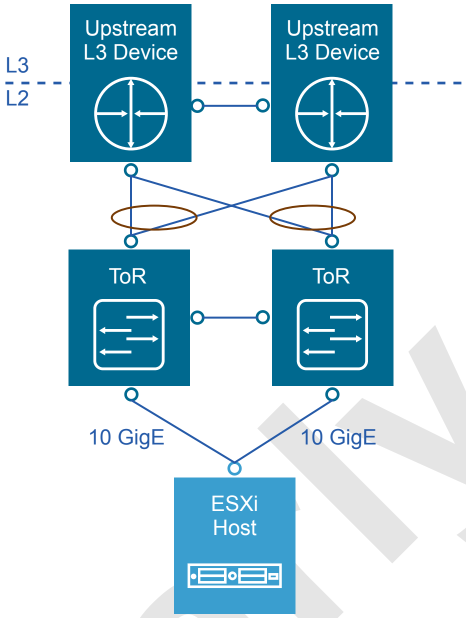
- In a design that uses Layer 2 transport, top of rack switches and upstream Layer 3 devices, such as core switches or routers, form a switched fabric.
- The upstream Layer 3 devices terminate each VLAN and provide default gateway functionality.
- Uplinks from the top of rack switch to the upstream Layer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

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

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

Characteristic	Description
Benefits	■ More design freedom. ■ You can span VLANs across racks, which can be useful in some circumstances.
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 1-1. Example Layer 2 Transport



Benefits and Drawbacks of Layer 3 Transport

A design using Layer 3 transport requires these considerations:

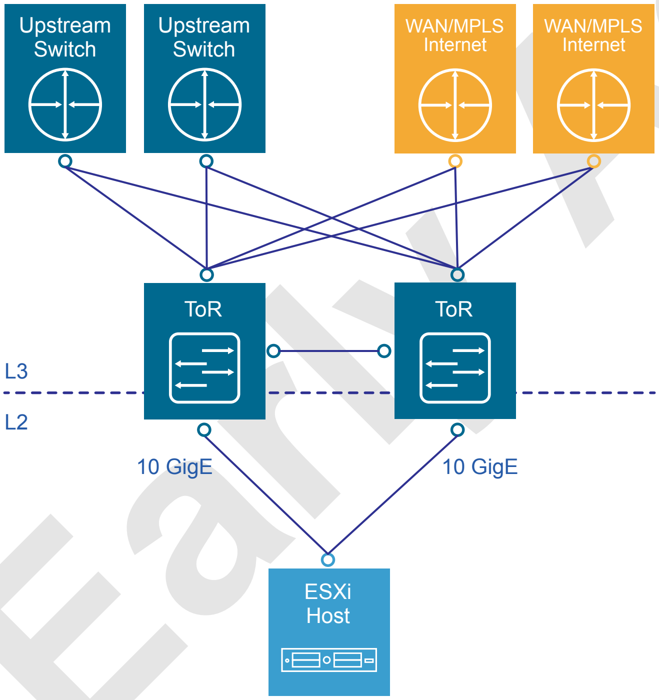
- Layer 2 connectivity is limited within the data center rack up to the top of rack switches.
- The top of rack switch terminates each VLAN and provides default gateway functionality. That is, it has a switch virtual interface (SVI) for each VLAN.
- Uplinks from the top of rack switch to the upstream 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 1-2. 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 1-2. Example Layer 3 Transport

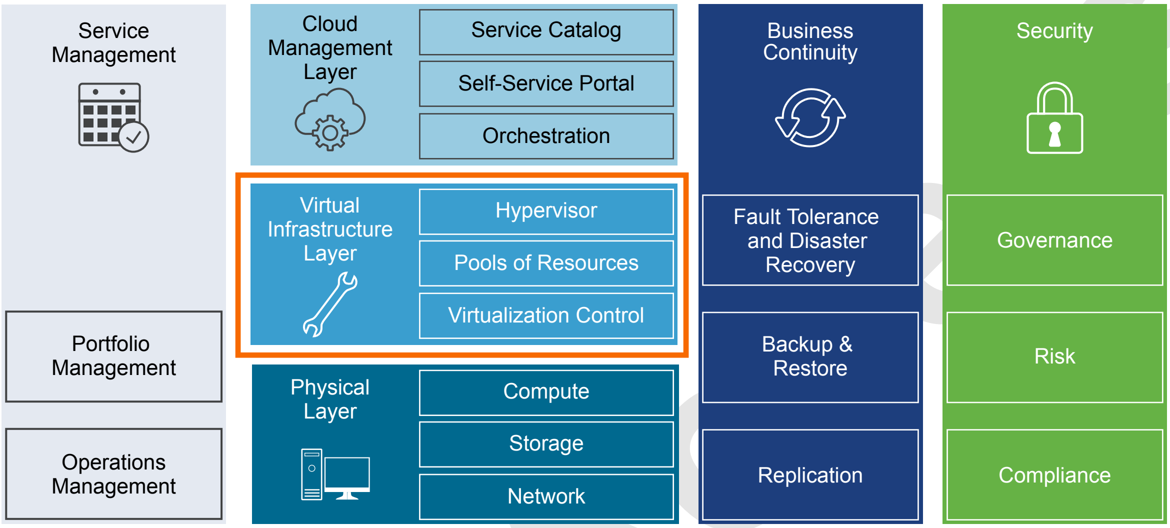


Virtual Infrastructure Architecture

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 1-3. 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 edge and compute clusters.

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.

Edge Cluster

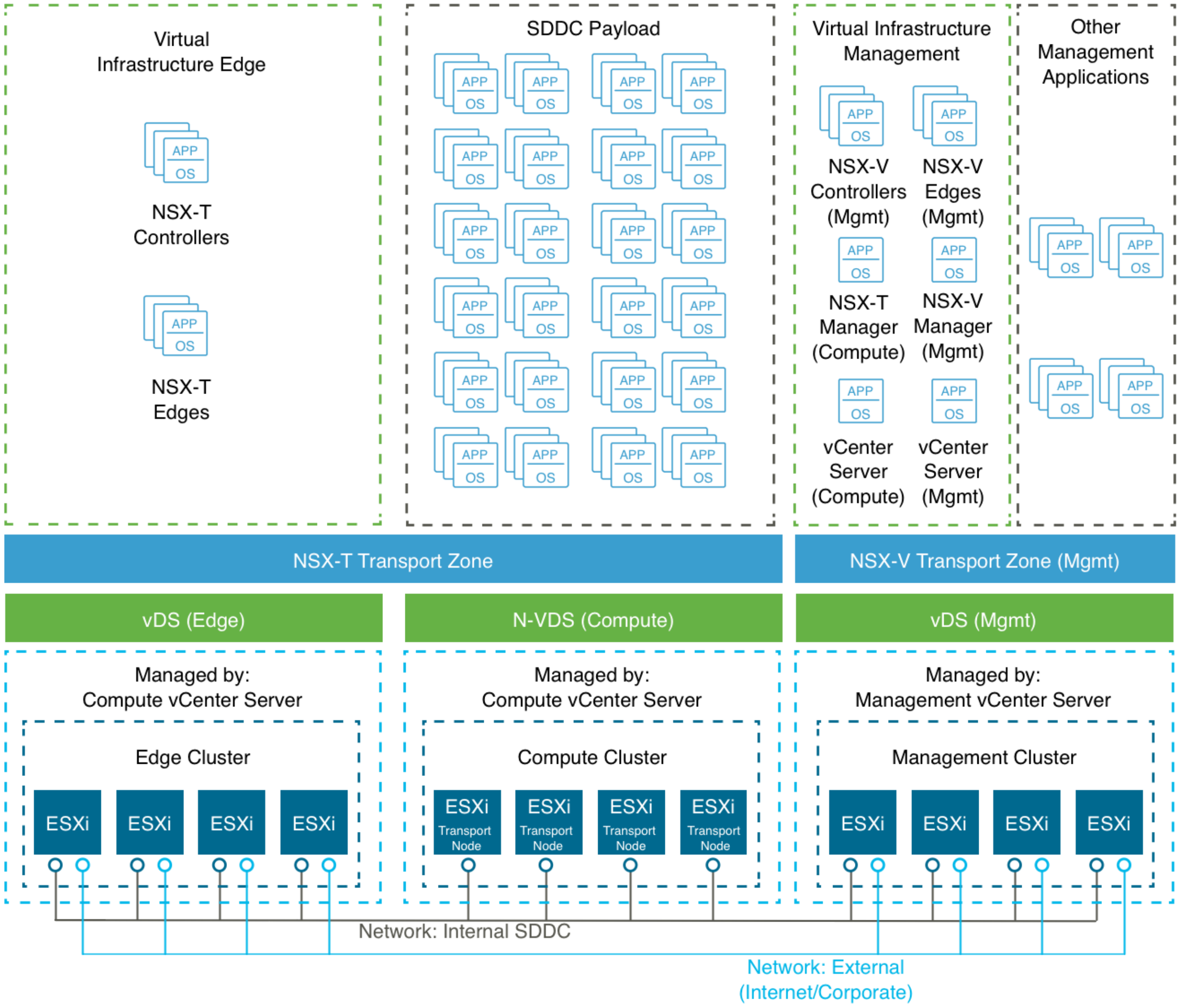
The edge cluster runs the NSX-T controllers and edge virtual machines. The edge virtual machines are responsible for North-South routing between compute workloads and the external network. This is often referred to as the on-off ramp of the SDDC.

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

Compute Cluster

The compute cluster hosts all tenant workloads. The hosts in this cluster provide services such as high availability and resource scheduling to the tenant workloads.

Figure 1-4. 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. All virtual networks are created on top of this layer.

Several components are required to create this network virtualization layer:

- NSX-T Manager
- NSX-T Controllers
- NSX-T Edge Nodes
- NSX-T Distributed Routers (DR)
- NSX-T Service Routers (SR)
- NSX-T Logical Switches

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

Data plane

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

Workload data is entirely in the data plane. The data is carried over designated transport networks in the physical network. The NSX-T N-VDS virtual switch, distributed routing, and the distributed firewall are also implemented in the data plane.

Control plane

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

The control plane computes 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, meaning any failure in the control plane does not affect existing data plane operations. User traffic does not pass through the CCP cluster.

- Local Control Plane (LCP). The LCP runs on transport nodes. It is adjacent 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

The 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 according to its stage of existence. The management plane also handles querying recent status and statistics from the control plane, and sometimes directly from the data plane.

The management plane is the only source of truth for the logical system because users manage it via configuration. You make changes using either a RESTful API or the NSX-T UI.

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 logical switches, routers, firewalls, and other components of NSX-T.

Logical Switch

Logical switches 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 logical switch has a virtual network identifier (VNI), like a VLAN ID. Unlike VLANs, VNIs scale well beyond the limits of VLAN IDs.

An NSX-T logical switch reproduces switching functionality, broadcast, unknown unicast, and multicast (BUM) traffic, in a virtual environment that is decoupled from underlying hardware.

Logical Router

NSX-T logical routers provide North-South connectivity, thereby enabling workloads to 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 can be 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.

A logical router consists of two optional parts : a distributed router (DR) and one or more service routers (SR).

A DR spans ESXi hosts whose virtual machines are connected to this logical router, and edge nodes the logical router is bound to. Functionally, the DR is responsible for one-hop distributed routing between logical switches and logical routers connected to this logical router. An SR is responsible for delivering services that are not currently implemented in a distributed fashion, such as stateful NAT.

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

NSX-T Edge Node

NSX-T Edge nodes provide routing services and connectivity to networks that are external to the NSX-T deployment.

NSX-T Edges are required for establishing external connectivity from the NSX-T domain, through a Tier-0 router over BGP or static routing. Additionally, you must deploy an NSX-T Edge for stateful services at either the Tier-0 or Tier-1 logical routers.

NSX-T Edge Cluster

An NSX-T Edge cluster is a collection of NSX-T Edge nodes that host multiple service routers in highly available configurations. At minimum, a single Tier-0 SR must be deployed to provide external connectivity

A NSX-T Edge cluster is not in a one-to-one relationsihp with a vSphere cluster. A vSphere cluster can run multiple NSX-T Edge Clusters.

Transport Node

A transport node is a node that is capable of participating in an NSX-T overlay or NSX-T VLAN networking. Any node can serve as a transport node if it contains an N-VDS such as ESXi hosts and NSX-T Edges.

An ESXi host can serve as a transport node if it contains at least one NSX managed virtual distributed switch (N-VDS).

Transport Zone

A transport zone controls which transport nodes a logical switch can reach. It can span one or more vSphere clusters. Transport zones dictate which ESXi hosts and, therefore, 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 logical switches and for creating NSX-T logical router uplinks and downlinks.

NSX-T Controller

NSX-T Controllers are an advanced distributed state management system that controls virtual networks and overlay transport tunnels.

To address stability and reliability of data transport, NSX-T Controllers are deployed as a cluster of three highly available virtual appliances. They are responsible for the programmatic deployment of virtual networks across the entire NSX-T architecture. The NSX-T Central Control Plane (CCP) is logically separated from all data plane traffic. A failure in the control plane does not affect existing data plane operations.

Logical Firewall

NSX-T uses firewall rules to specify traffic handling in and out of the network.

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

The default rule, located 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 enforce access control through a positive control model, that is, only traffic defined in a firewall rule is allowed onto 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 distributes incoming service requests evenly among multiple servers in such a way that the load distribution is transparent to users. Load balancing helps in achieving optimal resource utilization, maximizing throughput, minimizing response time, and avoiding overload.

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 a SR on the Tier-1 logical router.

Detailed Design

2

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.

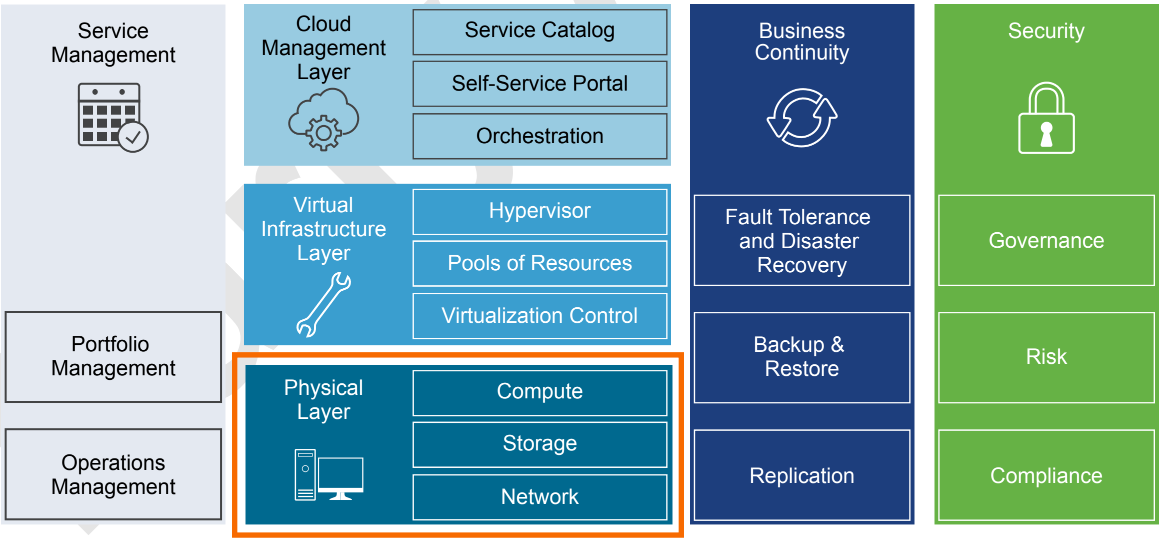
This chapter includes the following topics:

- Physical Infrastructure Design
- Virtual Infrastructure Design

Physical Infrastructure Design

The physical infrastructure design includes details on decisions for the physical network.

Figure 2-1. 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 *NSX-T for Workload Domain* guidance with most enterprise-grade physical network architectures.

Switch Types and Network Connectivity

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

Top of Rack Physical Switches

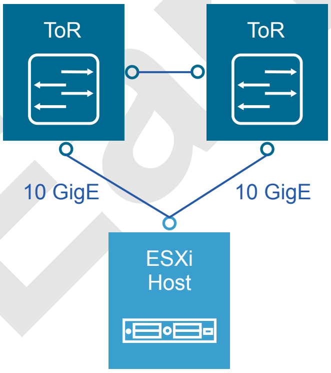
When configuring top of rack (ToR) switches, consider the following best practices.

- 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 it takes to transition ports over to the forwarding state, for example using the Trunk PortFast feature found in a Cisco physical switch.
- Provide DHCP or DHCP Helper capabilities on all VLANs that are used by 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 SDDC network fabric ToR switches by means of two 10 GbE ports. Configure the ToR switches to provide all necessary VLANs via an 802.1Q trunk. These redundant connections use features in the vSphere Distributed Switch and NSX-T to guarantee no physical interface is overrun and available redundant paths are used.

Figure 2-2. 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 .253 as the (floating) interface with .251 and .252 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 come with STP configured by default. Designate the access ports as trunk PortFast.

Trunking Configure the VLANs as members of a 802.1Q trunk with the management VLAN acting as the native VLAN.

MTU Set MTU for all VLANS and SVIs (Management, vMotion, VXLAN and Storage) to jumbo frames for consistency purposes.

DHCP Helper Configure a DHCP helper (sometimes referred to as 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 only supports the BGP routing protocol.

DHCP Helper Set the DHCP helper (relay) to point to a DHCP server by way of its IPv4 address.

Table 2‑1. Physical Network Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-PHY-NET-001	The physical network architecture must support the following requirements: ■ One 10-GbE or faster 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. This design uses vSphere host profiles which are not compatible with link-aggregation technologies. NSX-T only supports BGP as a dynamic routing protocol.	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. NSX-T only supports BGP as a dynamic routing protocol.	Requires BGP configuration in the physical network.

Table 2-1. Physical Network Design Decisions (Continued)

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-PHY-NET-003	Use two ToR switches for each rack.	Supports the use of two 10-GbE 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 2-2. 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 which will be assigned by DHCP.	Provides a consistent access point for management interfaces. NSX-T VTEPs are not accessed by users and are only used for tunnels between ESXi hosts and edge virtual machines. As such, you can use DHCP to ease the IP management of these interfaces.	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. Jumbo frames must be configured end-to-end, which is feasible in a LAN environment. When you enable jumbo frames on an ESXi host, you have to select an MTU that matches the MTU of the physical switch ports.

The workload determines whether 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.

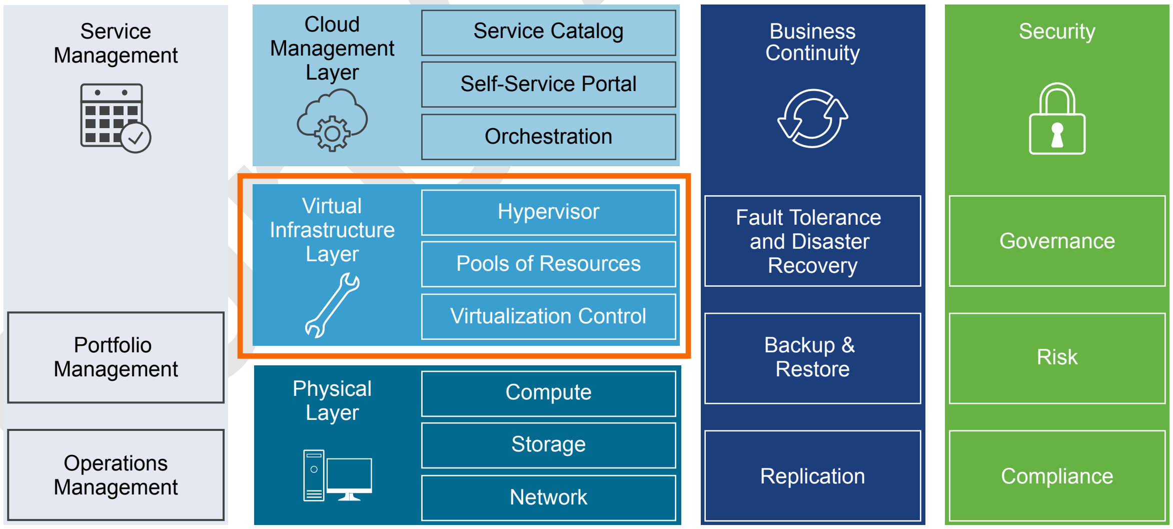
Table 2-3. Jumbo Frames Design Decisions

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. Using an MTU of 9000 bytes increases throughput for Geneve and ensures consistency across port groups that are adjusted from the default MTU size.	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 2-3. Virtual Infrastructure Layer in the SDDC



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

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

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

Best Practices

Follow networking best practices throughout your environment.

- Separate network services from one another to achieve greater security and better performance.
- Use Network I/O Control and traffic shaping to guarantee bandwidth to critical virtual machines. During network contention, these critical virtual machines receive a higher percentage of the bandwidth.
- Separate network services on a single vSphere Distributed Switch in the edge cluster or an NSX Managed Virtual Distributed Switch (N-VDS) in compute clusters by attaching them to port groups with different VLAN IDs.
- Keep vSphere vMotion traffic on a separate network. When migration with vMotion occurs, the contents of the 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 that are connected to the same virtual switch, are also connected to the same physical network.

Network Segmentation and VLANs

Separating different types of traffic is required to reduce contention and latency, and for access security.

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

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 depending on 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 one single pane of glass view for performing virtual network management tasks.

Virtual Switch Design Background

vSphere Distributed Switch and NSX managed virtual distributed switch (N-VDS) offer several enhancements over standard virtual switches.

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, reduces mistakes, and lowers operational costs.

Additional features Distributed switches offer features that are not available on standard virtual switches. Some of these features can be useful to the applications and services that are running in the organization's infrastructure. For example, NetFlow and port mirroring provide monitoring and troubleshooting capabilities to the virtual infrastructure.

Consider the following caveats for distributed switches:

- vSphere Distributed Switches are manageable only when vCenter Server is available. vCenter Server therefore becomes a Tier-1 application.
- N-VDS instances are manageable only when the NSX-T Manager is available. NSX-T therefore 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 the N-VDS for NSX-T based compute clusters.	N-VDS is required to carry overlay traffic.	■ N-VDS is not compatible with vSphere host profiles. ■ vSAN health check does not function when the vSAN VMkernel port is on an N-VDS. ■ Migrating VMkernel ports to N-VDS is available only using the API.
NSXT-VI-NET-002	Use vSphere Distributed Switch for edge clusters.	Because each edge transport node contains an embedded N-VDS, the edge must reside on a distributed or standard switch. vSphere Distributed Switches simplify management.	Migration from a standard switch to a distributed switch requires at least two physical NICs to maintain redundancy.

Edge Cluster Distributed Switches

The edge cluster uses a single vSphere Distributed Switch with certain configuration for handled traffic types, Network I/O Control setup, physical NICs, MTU size, and failover.

Table 2-4. Virtual Switch for the Edge Cluster

vSphere Distributed Switch Name	Function	Network I/O Control	Number of Physical NIC Ports	MTU
sfo01-w01-vds02	■ ESXi Management ■ vSphere vMotion ■ vSAN ■ Geneve Overlay (TEP) ■ Uplinks (2) to enable ECMP	Enabled	2	9000

Table 2-5. sfo01-w01-vds02 Port Group Configuration Settings

Parameter	Setting
Failover detection	Link status only
Notify switches	Enabled
Failback	Yes

Network Switch Design for Edge ESXi Hosts

When you design the switch configuration on the ESXi hosts, consider the physical NIC layout and physical network attributes.

Figure 2-4. Network Switch Design for Edge ESXi Hosts

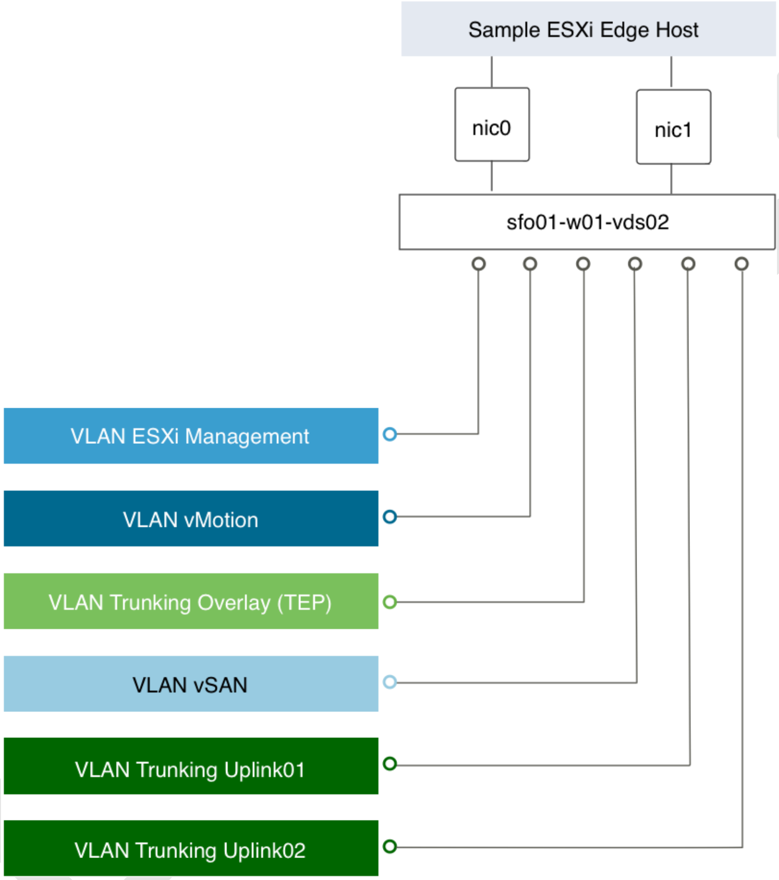


Table 2-6. Edge Cluster Virtual Switches by Physical or Virtual NIC

vSphere Distributed Switch	vmnic	Function
sfo01-w01-vds02	0	Uplink
sfo01-w01-vds02	1	Uplink

Table 2-7. Port Groups and VLANs in the Edge Cluster Virtual Switch

vSphere Distributed Switch	Port Group Name	Teaming Policy	VLAN Type	Active Uplinks
sfo01-w01-vds02	sfo01-w01-vds02-management	Route based on physical NIC load	VLAN	1,2
sfo01-w01-vds02	sfo01-w01-vds02-vmotion	Route based on physical NIC load	VLAN	1,2
sfo01-w01-vds02	sfo01-w01-vds02-vsan	Route based on physical NIC load	VLAN	1,2
sfo01-w01-vds02	sfo01-w01-vds02-overlay	Route based on physical NIC load	VLAN trunking	1,2
sfo01-w01-vds02	sfo01-w01-vds02-uplink01	Route based on originating virtual port	VLAN trunking	1
sfo01-w01-vds02	sfo01-w01-vds02-uplink02	Route based on originating virtual port	VLAN trunking	2

Table 2-8. VMkernel Adapters for the Edge Cluster

vSphere Distributed Switch	Connected Port Group	Enabled Services	MTU
sfo01-w01-vds02	sfo01-w01-vds02-management	Management Traffic	1500 (Default)
sfo01-w01-vds02	sfo01-w01-vds02-vmotion	vMotion Traffic	9000
sfo01-w01-vds02	sfo01-w01-vds02-vsan	vSAN	9000

Health Check

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

Health check helps identify the following common configuration errors.

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

Health check monitors VLAN, MTU, and teaming policies.

VLANs Checks whether the VLAN settings on the distributed switch match the trunk port configuration on the connected physical switch ports.

MTU For each VLAN, determines whether the MTU size setting on the facing physical access switch ports matches the MTU size setting on the distributed switch.

Teaming policies Determines whether the connected access ports of the physical switch that participate in an EtherChannel are paired with distributed ports whose teaming policy is Route based on IP hash.

Health check is limited to the access switch port to which the NICs of the ESXi hosts are connected.

Table 2-9. vSphere Distributed Switch Health Check Design Decisions

Design ID	Design Decision	Design Justification	Design Implication
NSXT-VI-NET003	Enable vSphere Distributed Switch Health Check on all vSphere Distributed Switches.	vSphere Distributed Switch Health Check ensures that all VLANS are trunked to all ESXi hosts attached to the vSphere Distributed Switch and ensures MTU sizes match the physical network.	You must have a minimum of two physical uplinks to use this feature.

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

Compute Cluster Switches

The compute cluster uses a single NSX managed virtual distributed switch (N-VDS) with a certain configuration for handled traffic types, NIC teaming, and MTU size.

Table 2-10. Virtual Switch for the Compute Cluster

N-VDS Switch Name	Function	Number of Physical NIC Ports	Teaming Policy	MTU
sfo01-w01-nvds01	■ ESXi Management ■ vSphere vMotion ■ vSAN ■ Geneve Overlay (TEP)	2	Load balance source	9000

Figure 2-5. Network Switch Design for Compute ESXi Hosts

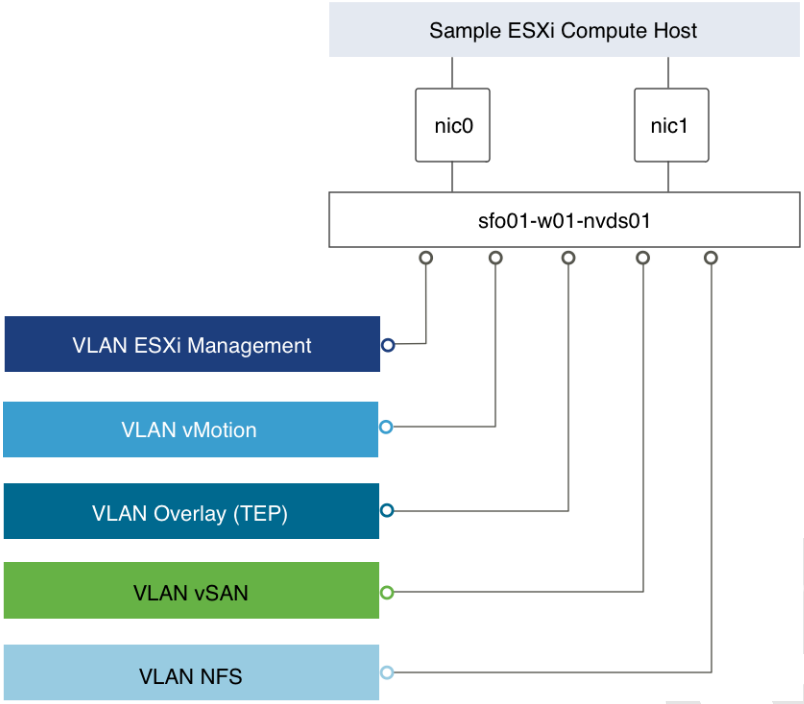


Table 2-11. Compute Cluster Virtual Switches by Physical or Virtual NIC

N-VDS Switch	vmnic	Function
sfo01-w01-nvds01	0	Uplink
sfo01-w01-nvds01	1	Uplink

Table 2-12. Compute Cluster Logical Switches

N-VDS Switch	Logical Switch Name
sfo01-w01-nvds01	sfo01-w01-nvds01-management
sfo01-w01-nvds01	sfo01-w01-nvds01-vmotion
sfo01-w01-nvds01	sfo01-w01-nvds01-vsan
sfo01-w01-nvds01	sfo01-w01-nvds01-nfs

Table 2-13. VMkernel Adapters for the Compute Cluster

N-VDS Switch	Logical Switch Name	Enabled Services
sfo01-w01-nvds01	sfo01-w01-nvds01-management	Management Traffic
sfo01-w01-nvds01	sfo01-w01-nvds01-vmotion	vMotion Traffic
sfo01-w01-nvds01	sfo01-w01-nvds01-vsan	vSAN
sfo01-w01-nvds01	sfo01-w01-nvds01-nfs	--

Note 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 further reduce the risk of a single point of failure, build NIC teams by using ports from multiple NIC and motherboard interfaces.

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

This VMware Validated Design uses a non-LAG based active-active configuration using the route based on physical NIC load and Load balance source algorithms for teaming. In this configuration, idle network cards do not wait for a failure to occur, and they aggregate bandwidth.

NIC Teaming Design Background

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

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

This validated design uses an active-active configuration.

Table 2-14. 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 2-15. NIC Teaming Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-NET-004	In the edge cluster, use the Route based on physical NIC load teaming algorithm for all port groups except for port groups for edge virtual machine uplinks.	Route based on physical NIC load initially places traffic based on originating virtual port but moves traffic from one physical NIC to another when a physical NIC reaches 75% utilization for 30 seconds.	None.
NSXT-VI-NET-005	In the edge cluster, use Route based on originating virtual port for edge virtual machine uplinks.	The two edge virtual machine uplink port groups only contain a single physical uplink, as such the default teaming policy is sufficient.	None.
NSXT-VI-NET-006	In the compute cluster, use the Load balance source teaming policy on N-VDS.	The 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 abstract and pool the networking resources. Just like vSphere abstracts compute capacity from the server hardware to create virtual pools of resources that can be consumed as a service, NSX-T, using the Geneve overlay, abstracts the network into a generalized pool of capacity and separates the consumption of these services from the underlying physical infrastructure. The pool of network capacity can then be optimally segmented into logical networks that are directly attached to specific applications.

Geneve is a tunneling mechanism which provides extensibility while still using the traditional offload capabilities offered by NICs for performance. The ability to insert additional context into the overlay header unlocks doors for future innovations in context awareness, end-to-end telemetry, security, encryption, and more.

Geneve works by creating Layer 2 logical networks that are encapsulated in UDP packets. A Segment ID in every frame differentiates the Geneve logical networks from each other 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 router 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, so that the network administrator can manage the environment in a secure and centralized way.

Table 2-16. Geneve Overlay Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-NET-007	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-008	Use overlay networks with NSX-T edge virtual machines and distributed routing to provide virtualized network capabilities to workloads.	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 1700 bytes.

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, snapshots, deletes, and restores software-based virtual machines (VMs), NSX network virtualization programmatically creates, snapshots, deletes, and restores software-based virtual networks. The result is a transformative approach to networking that not only enables data center managers to achieve orders of magnitude better agility and economics, but also supports a vastly simplified operational model for the underlying physical network. NSX-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 major constraint to cloud deployments where speed, agility, and flexibility are critical requirements. Pre-provisioned physical networks enables rapid creation of virtual networks and faster deployment times of workloads using the virtual network. As long as the physical network that you need is already available on the ESXi host to host the workload, pre-provisioning physical networks works well. However, if the network is not available on a given ESXi host, you must find an ESXi host with the available network and spare capacity to run your workload in your environment.

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

NSX-T Design

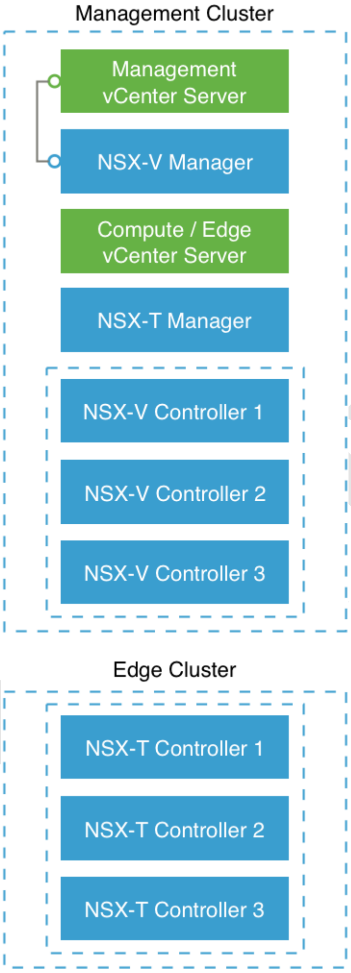
NSX-T components are not dedicated to a specific vCenter Server or vSphere construct. You can share them across different vSphere environments.

NSX-T, while not dedicated to a vCenter Server, supports only single-region deployments in the current release. This design is focused on compute clusters in a single region.

Table 2-17. NSX-T Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-001	Deploy a single NSX-T Manager to configure and manage all NSX-T based compute clusters in a single region.	Software-defined networking (SDN) capabilities offered by NSX, such as load balancing and firewalls, are required to support the required functionality in the compute and edge layers.	You must install and configure NSX-T Manager in a highly available management cluster.
NSXT-VI-SDN-002	In the management cluster, add the NSX-T Manager to the NSX for vSphere Distributed Firewall exclusion list.	Ensures that the management plane is still available if a misconfiguration of the NSX for vSphere Distributed Firewall occurs.	None.

Figure 2-6. NSX-T Architecture



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 REST APIs for creating, configuring, and monitoring NSX-T components, such as logical switches.

NSX-T Manager implements the management plane for the NSX-T eco-system. 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 created by NSX-T. It provides configuration and orchestration of the following.

- Logical networking components – logical switching and routing
- Networking and Edge services
- Security services and distributed firewall

NSX-T Manager also provides REST API entry-points to automate consumption. Because of this architecture, you can automate all configuration and monitoring aspects using any cloud management platform, security vendor platform, or automation framework.

The NSX-T Management Plane Agent (MPA) is an NSX 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.

Table 2-18. NSX-T Manager Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-003	Deploy NSX-T Manager as a large size virtual appliance.	The large-size appliance supports more than 64 ESXi hosts. The small-size appliance is for proof of concepts 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	■ Grant administrators access to both the NSX-T Manager UI and its REST API endpoint. ■ Restrict end-user access to the REST 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 Controller

The NSX-T Controller is an advanced distributed state management system that controls virtual networks and overlay transport tunnels.

For stability and reliability of data transport, the NSX-T Controller is deployed as a cluster of three highly available virtual appliances that are responsible for the programmatic deployment of virtual networks across the entire NSX-T architecture. Because the NSX-T Central Control Plane (CCP) is logically separated from all data plane traffic, existing data plane operations continue running if a failure in the control plane occurs. Instead of handling data traffic, the controller provides configuration to other NSX-T Controller components such as the logical switches, logical routers, and edge virtual machine configuration.

Table 2-19. NSX-T Controller Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-005	Deploy the NSX-T Controller cluster with three members to provide high availability and scale. Provision these three nodes in the edge cluster by deploying the nsx-controller-*version*.ova.	The high availability of NSX-T Controller reduces the downtime period in case of failure of one physical ESXi host.	None.

NSX Managed Virtual Distributed Switch (N-VDS)

Runs on ESXi hosts and provides physical traffic forwarding. N-VDS is invisible to the tenant network administrator and provides the underlying forwarding service that each logical switch relies on. To achieve 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 defined when they created and configured their logical switches.

Each logical broadcast domain is implemented by tunneling VM-to-VM traffic and VM-to-logical router 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 2-20. NSX-T N-VDS Design Decision

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

Logical Switching

NSX-T logical switches create logically abstracted segments to which workload virtual machines can be connected. A single logical switch is mapped to a unique Geneve segment and is distributed across the ESXi hosts in a transport zone. The logical switch supports line-rate switching in the ESXi host without the constraints of VLAN sprawl or spanning tree issues.

Table 2-21. NSX-T Logical Switching Design Decision

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

Logical Routers

NSX-T logical routers 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 can the physical router can handle. Each logical router 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 in a single physical router can perform the operations previously performed by several pieces of equipment.

A logical router consists of two optional parts : a distributed router (DR) and one or more service routers (SR).

A DR spans ESXi hosts whose virtual machines are connected to this logical router, and edge nodes the logical router is bound to. Functionally, the DR is responsible for one-hop distributed routing between logical switches and logical routers connected to this logical router. The SR is responsible for services that are not currently implemented in a distributed fashion, such as NAT.

A logical router always has a DR, and it has an SR when the logical router is a Tier-0 or when the logical router is a Tier-1 and has services configured such as NAT or DHCP.

Tunnel Endpoint

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.

An NSX-T Edge is required for establishing external connectivity from the NSX-T domain, through a Tier-0 router using BGP or static routing. Additionally, an NSX-T Edge must be deployed to support network address translation (NAT) services at either the Tier-0 or Tier-1 logical routers.

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 firewall rules to specify traffic handling in and out of the network.

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

The default rule, located 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 enforce access control through a positive control model, that is, only traffic defined in a firewall rule is allowed onto 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 distributes incoming service requests evenly among multiple servers in such a way that the load distribution is transparent to users. Load balancing helps in achieving optimal resource use, maximizing throughput, minimizing response time, and avoiding overload.

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 logical router.

NSX-T Requirements

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 1700 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 2-3.
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 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 vary according to tenant requirements. Consider all options in such a case.

Virtual Machine	vCPU	Memory (GB)	Storage (GB)	Quantity per NSX-T Deployment
NSX-T Manager	8 (Large)	32 (Large)	140 (Large)	1
NSX-T Controller	4	16	120	3
NSX-T Edge virtual machine	2 (Small)	4 (Small)	120 (Small)	Number varies per 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 2-22. NSX Edge VM Sizing Design Decision

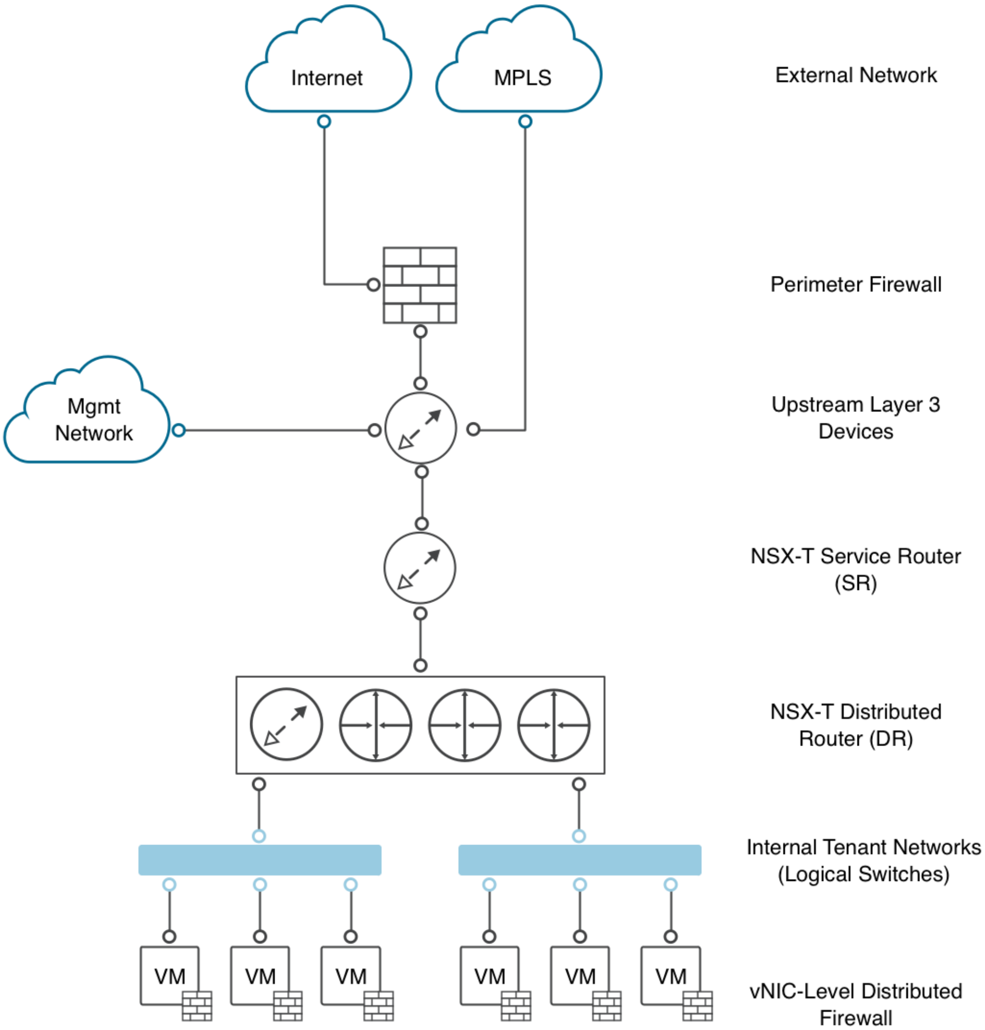
Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-008	Use large-size NSX Edge virtual machines.	The large-size appliance provides all the performance characteristics if a failure occurs.	None.

Network Virtualization Conceptual Design

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 logical router. It also considers the external network, internal workload networks, and the management network.

Figure 2-7. NSX-T Conceptual Overview


Internet
MPLS
External Network
Perimeter Firewall
Mgmt Network
Upstream Layer 3 Devices
NSX-T Service Router (SR)
NSX-T Distributed Router (DR)
Internal Tenant Networks (Logical Switches)
VM
VM
VM
VM
VM
VM
vNIC-Level Distributed Firewall

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 Logical Router (SR) The SR component of the NSX-T Tier-0 Logical Router is responsible for establishing eBGP peering with the PLR and enabling North-South routing.

NSX-T Logical Router (DR) The DR component of the NSX-T Logical Router 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 logical switches 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, edge, and compute clusters. You can add more compute clusters for scale-out, or different workload types or SLAs.

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.

vSphere Edge Cluster

In the edge cluster, the ESXi hosts are not prepared for NSX-T. As a result, they can be connected to a vSphere Distributed Switch and host the NSX-T Edge virtual machines and NSX-T Controllers.

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 edge cluster and provide North-South routing for the workloads in the compute clusters.

Compute Cluster

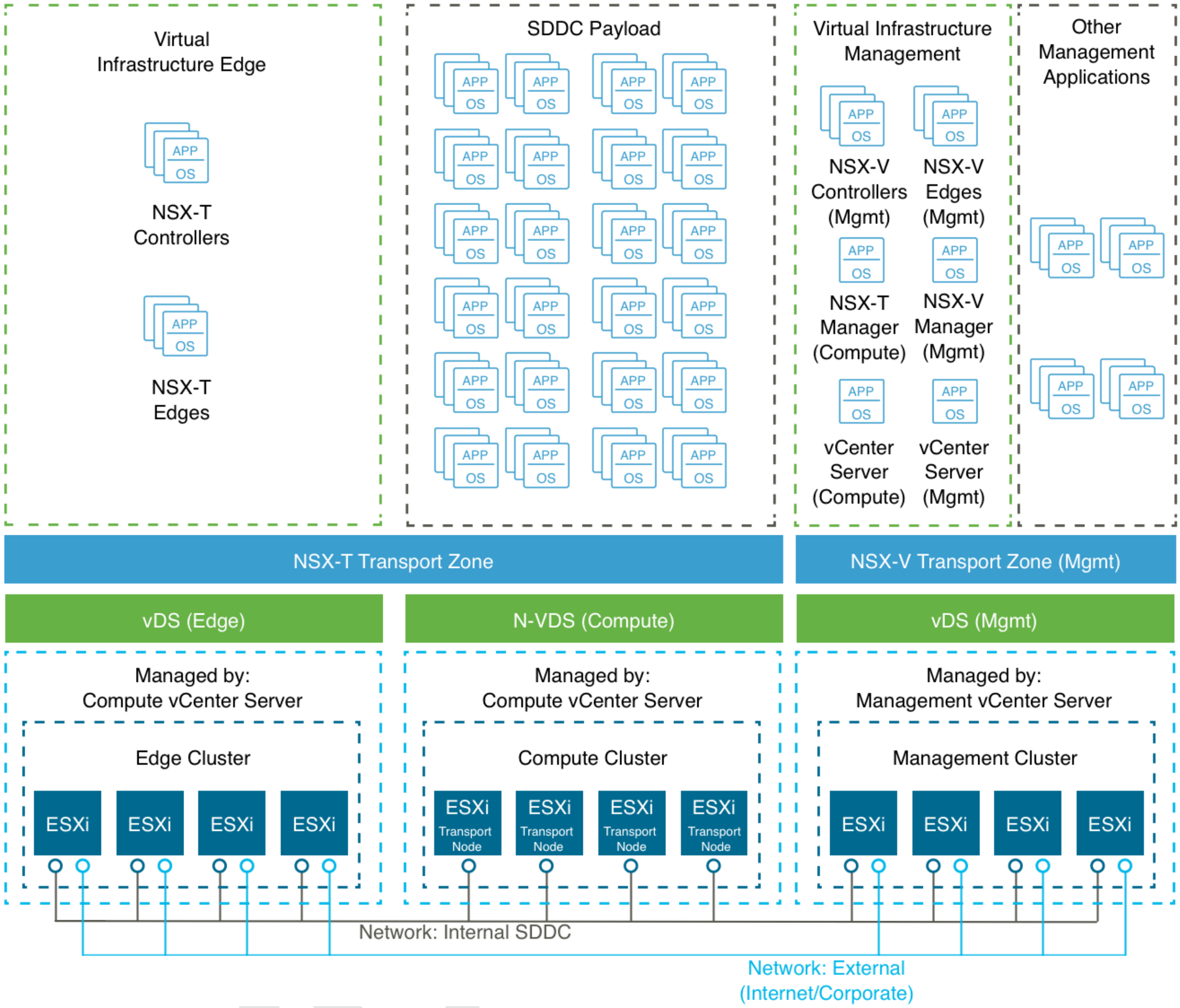
In the compute cluster, ESXi hosts are prepared for NSX-T. As a result, they can be configured as transport nodes and can participate in the overlay network. All tenant workloads run in the compute cluster.

Table 2-23. Cluster Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-009	Use a dedicated vSphere edge cluster for NSX-T Edge virtual machines and NSX-T Controllers.	Offers better separation in terms of security, management, and resources. The NSX-T Edge virtual machines must be connected to a vSphere Distributed Switch or vSphere Standard Switch. In a design where ESXi hosts have two physical NICs and network redundancy is required, do not prepare the cluster for NSX-T.	Requires an additional vSphere cluster to manage the NSX-T Edge virtual machines.
NSXT-VI-SDN-010	Deploy at least two large-size NSX-T Edge virtual machines in the vSphere edge cluster.	Creates NSX-T Edge cluster, and meets availability and scale requirements.	You must add the edge virtual machines as transport nodes before you add them to the NSX-T Edge cluster.
NSXT-VI-SDN-011	Apply vSphere Distributed Resource Scheduler (vSphere DRS) anti-affinity rules to NSX-T Controllers in the vSphere edge cluster.	Prevents controllers 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 Controllers 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 anti-affinity rules to the virtual machines of the NSX-T Edge cluster in the vSphere edge cluster.	Prevents the NSX-T Edge virtual machines from running on the same ESXi host and thereby risking their high availability capability.	Additional configuration is required to set up anti-affinity rules.

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

Figure 2-8. NSX-T Cluster Design



High Availability of NSX-T Components

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

The NSX-T Controller nodes run on the vSphere edge cluster. vSphere DRS anti-affinity rules prevent NSX-T Controller nodes from running on the same ESXi host.

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

The NSX Edge virtual machines are deployed on the vSphere edge 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.

Logical Switch Replication Mode

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

The following options are available for BUM replication on logical switches.

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. It does not matter which ESXi host transport node is selected to perform replication in the remote groups as long as the remote ESXi host transport node is up and available.

Headend

In this mode, the ESXi host transport node at the origin of the frame to be flooded on a logical switch sends a copy to every other ESXi host transport node that is connected to this logical switch.

Table 2-24. Logical Switch Replication Mode Design Decision

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-013	Use hierarchical two-tier replication on all logical switches.	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 can be configured. A transport zone is not meant to delineate a security boundary.

Table 2-25. 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 logical switches are available to all ESXi hosts and edge virtual machines.	None.
NSXT-VI-SDN-015	Create a VLAN transport zone for ESXi host VMkernel ports.	Enables ESXi host VMkernel ports to be migrated to 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 set the VLAN mode to VLAN trunking on the vSphere distributed port groups that back edge virtual machine overlay and VLAN traffic.

Transport Node and Uplink Policy Design

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

There are several types of transport nodes available in NSX-T.

N-VDS N-VDS is a software-defined switch platform that is hypervisor independent. It is the primary component involved in the data plane of the transport nodes. N-VDS forwards traffic between components running on the transport node or between internal components and the physical network.

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 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 representing a pool of capacity.

Uplink profiles define policies for the links from ESXi hosts to NSX-T logical switches or from NSX 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. Instead of configuring individual properties or capabilities for each network adapter, uplink profiles specify the capabilities, which are then applied to the NSX-T transport nodes.

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.

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-017	Create an uplink profile with the load balance source teaming policy with two active uplinks for ESXi hosts.	Supports the concurrent use of both physical NICs on the ESXi hosts that are configured as transport nodes for increased resiliency and performance.	This policy can only be used for ESXi hosts. Edge virtual machines must use the failover order teaming policy.
NSXT-VI-SDN-018	Create an uplink profile with the failover order teaming policy with one active uplink and no standby uplinks for edge virtual machine overlay traffic.	Edge virtual machines support only uplink profiles 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.
NSXT-VI-SDN-019	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.

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-020	Add as transport nodes all ESXi hosts that are prepared for NSX-T.	Enables the participation of ESXi hosts and the virtual machines on them in NSX-T overlay and VLAN networks.	Hosts run N-VDS. You must migrate the existing VMkernel ports from the existing virtual switch to N-VDS. You must apply a staged approach where only a single physical NIC is assigned to the N-VDS instance until all VMkernel ports are migrated to N-VDS.
NSXT-VI-SDN-021	Add as transport nodes all edge virtual machines.	Enables the participation of edge virtual machines in the overlay network and providing of services, such as routing, by these machines.	None.
NSXT-VI-SDN-022	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

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

Routing can be defined in two 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 logical switches communicating with each other.

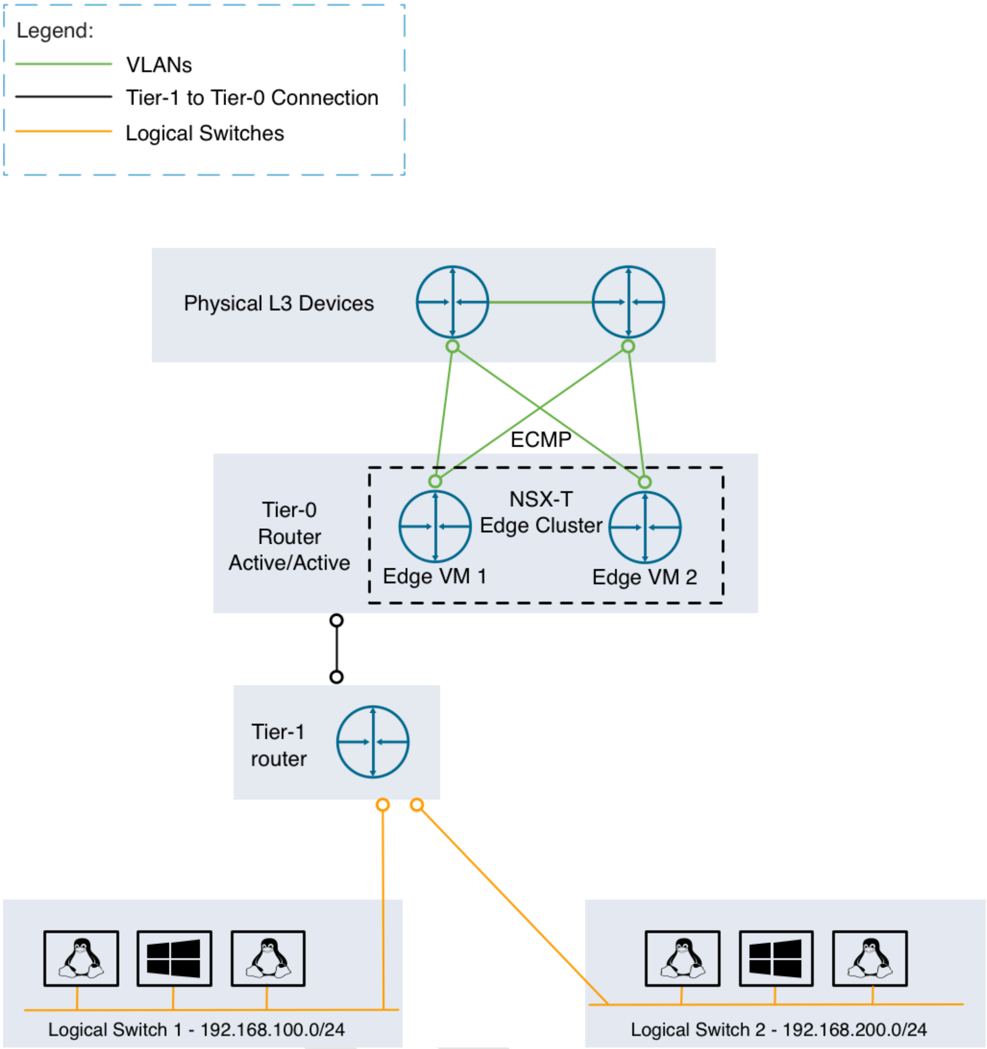
Table 2-26. Routing Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-023	Create two VLANs to enable ECMP between the Tier-0 router 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 Router and provides more resiliency and better bandwidth utilization in the network.	Extra VLANs are required.
NSXT-VI-SDN-024	Deploy an Active/Active Tier-0 router.	Provides support for ECMP North-South routing on all edge virtual machines in the NSX-T Edge cluster	Active/Active Tier-0 routers cannot provide services such as NAT. If you deploy a specific solution that requires stateful services on the Tier-0 router, such as PKS, you must provide an additional Tier-0 router in Active/Standby mode.
NSXT-VI-SDN-025	Use BGP as the dynamic routing protocol.	BGP is the only dynamic routing protocol supported in NSX-T.	In environments where BGP cannot be used, you must configure and manage static routes.
NSXT-VI-SDN-026	Configure BGP Keep Alive Timer to 4 and Hold Down Timer to 12 between the ToR switches and the Tier-0 router.	Provides a good balance between failure detection between the ToR switches and the Tier-0 router 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-027	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-028	Deploy a Tier-1 router to the NSX-T Edge cluster and connect it to the Tier-0 router.	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 router can only be connected to a single Tier-0 router. In scenarios where multiple Tier-0 routers are required, you must create multiple Tier-1 routers.

Virtual Network Design Example

The virtual network design example illustrates the connection of virtual machines to logical switches and the routing between the Tier-1 router and Tier-0 router, and then between the Tier-0 router and the physical network.

Figure 2-9. Virtual Network Example



Use of SSL Certificates

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 2-27. SSL Certificate Design Decision

Design ID	Design Decision	Design Justification	Design Implication
NSXT-VI-SDN-029	Replace the NSX-T Manager certificate with a certificate that is signed by a third-party Public Key Infrastructure.	Ensures that the communication between NSX administrators and the NSX-T Manager is encrypted by a trusted certificate.	Replacing and managing certificates is an operational overhead.