Architecture and Design for VMware NSX-T for Workload Domains with Multiple Availability Zones

Modified on 03 DEC 2019
VMware Validated Design 5.1
VMware NSX-T 2.4.1
VMware Validated Design 5.1.1
VMware NSX-T 2.5

vmware®

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

https://docs.vmware.com/

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

Contents

About Architecture and Design for VMware NSX-T for Workload Domains with Multiple Availability Zones

Architecture and Design for VMware NSX-T for Workload Domains with Multiple Availability Zones provides detailed information about the requirements for software, tools, and external services to implement VMware NSX-T™ Data Center 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 with VMware vSAN stretched clusters for tenant workloads.

Required VMware Software

In addition to the VMware Validated Design for Software-Defined Data Center 5.1 deployment, you must download NSX-T 2.4.1. 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

Update History

This *Architecture and Design for VMware NSX-T for Workload Domains* is updated when necessary.

Revision	Description
24 AUG 2020	At VMware, we value inclusion. To foster this principle within our customer, partner, and internal community, we are replacing some of the terminology in our content. We have updated this guide to remove instances of non-inclusive language.
18 JUL 2019	Initial release.

Applying the Guidance for VMware NSX-T Workload Domains with Multiple Availability Zones

1

The content in *Architecture and Design for VMware NSX-T for Workload Domains with Multiple Availability Zones* 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.

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

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 for VMware NSX-T for Workload Domains with Multiple Availability Zones* and *Architecture and Design*, follow the design guidelines in *Architecture and Design for VMware NSX-T for Workload Domains with Multiple Availability Zones*.

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 5.1 to create a shared edge and compute cluster that uses NSX-T capabilities.

This chapter includes the following topics:

Physical Network Architecture for NSX-T Workload Domains with Multiple Availability Zones

VMware Validated Designs can use most physical network architectures.

Network Transport for NSX-T Workload Domains with Multiple Availability Zones

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.

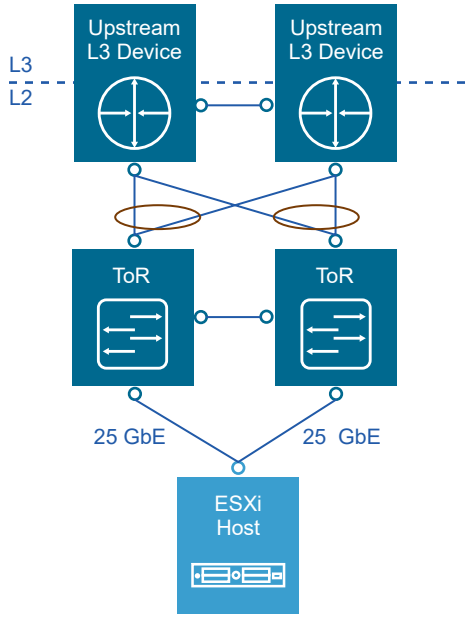
- The upstream Layer 3 device terminates each VLAN and provides default gateway functionality.
- Uplinks from the top of rack switch to the upstream Layer 3 devices are 802.1Q trunks carrying all required VLANs.

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:

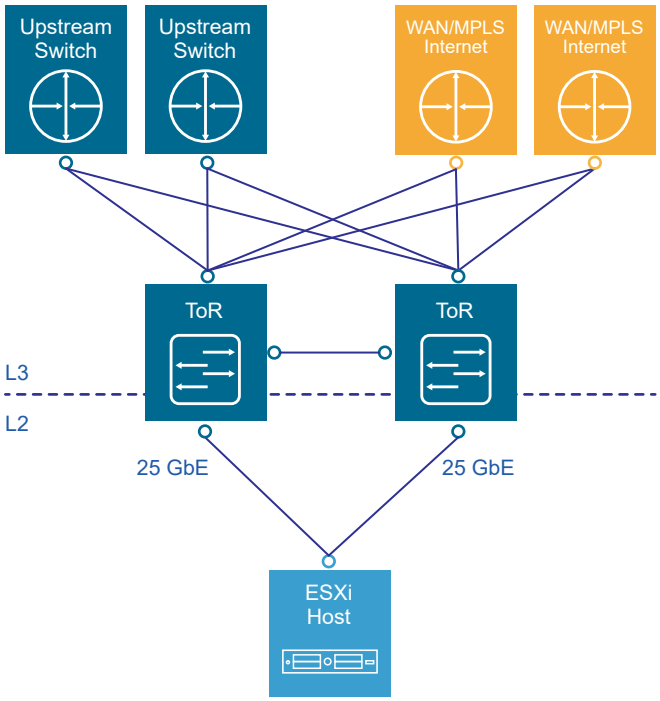
- Layer 2 connectivity is limited within the data center rack up to the top of rack switches.
- The top of rack switch terminates each VLAN and provides default gateway functionality. The top of rack switch has a switch virtual interface (SVI) for each VLAN.

- 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

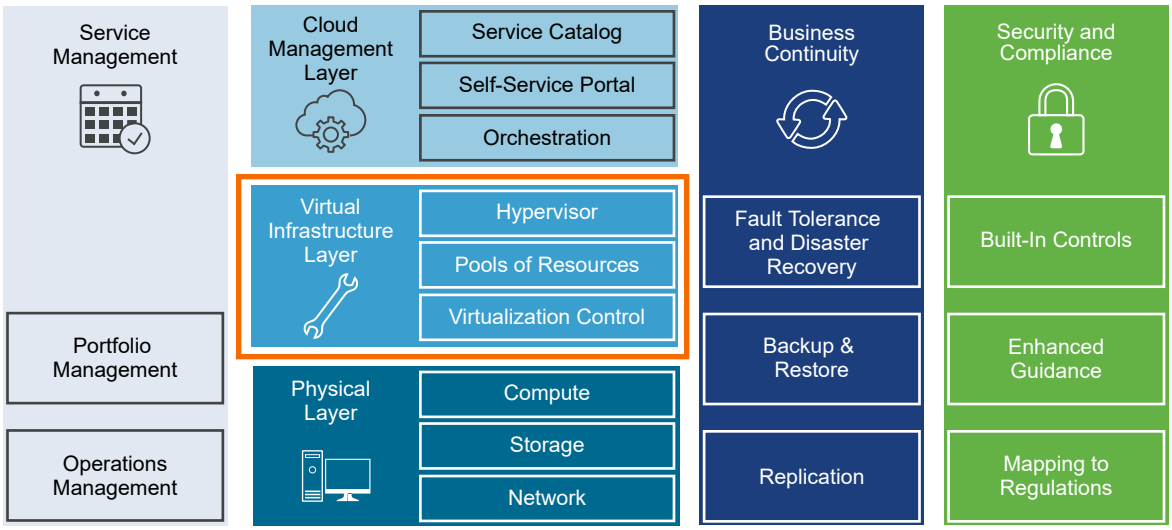


Virtual Infrastructure Architecture for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

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.

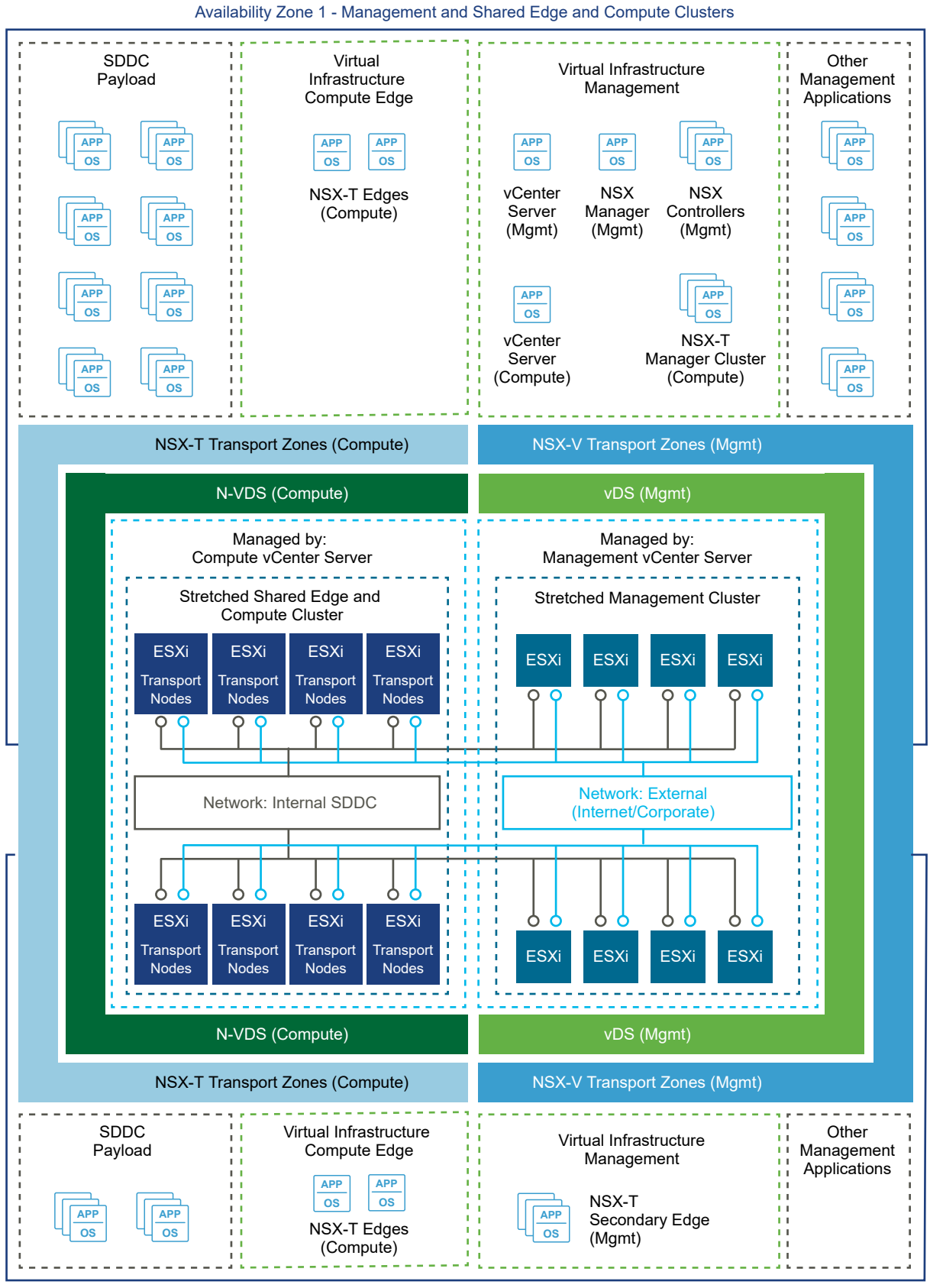
The management cluster architecture and design is covered in the VMware Validated Design for Software-Defined Data Center. The NSX-T validated design does not include the design of the management cluster.

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 NSX-T edge virtual machines also enable services such as load balancers.

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 for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

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 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

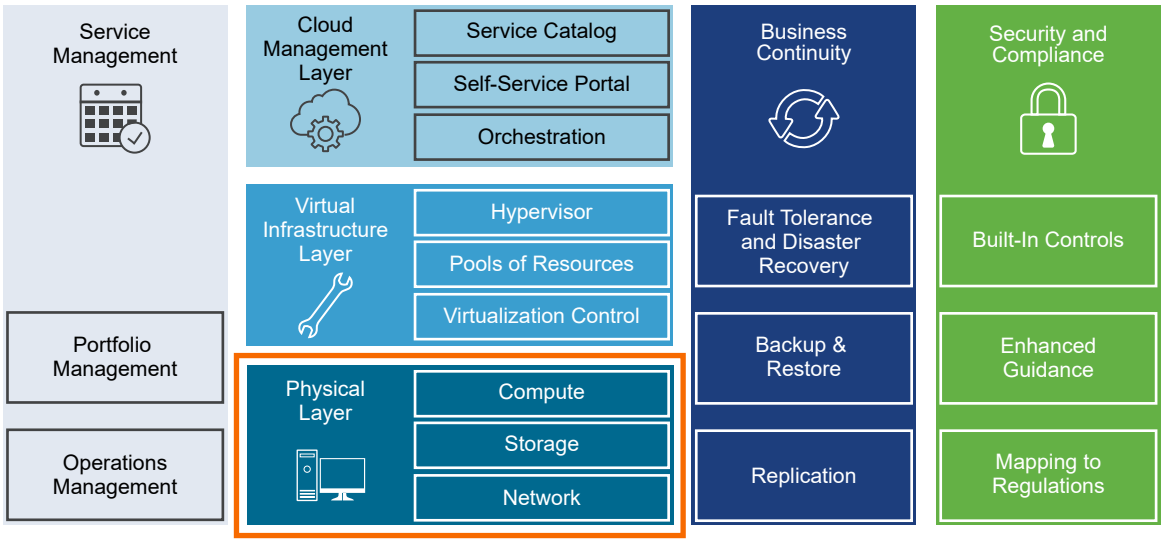
The NSX-T detailed design considers both physical and virtual infrastructure design. It includes numbered design decisions and the justification and implications of each decision.

This chapter includes the following topics:

Physical Infrastructure Design for NSX-T Workload Domains with Multiple Availability Zones

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

Figure 3-1. Physical Infrastructure Design



Physical Networking Design for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

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

- Physical Network Design Decisions for NSX-T Workload Domains with Multiple Availability Zones

  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.

Switch Types and Network Connectivity for NSX-T Workload Domains with Multiple Availability Zones

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

Top of Rack Physical Switches

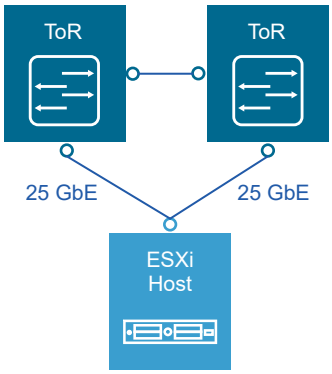
When configuring top of rack (ToR) switches, consider the following best practices:

- Configure redundant physical switches to enhance availability.
- Configure switch ports that connect to ESXi hosts manually as trunk ports.
- 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.

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

Table 3-1. Sample Values for VLANs and IP Ranges

Availability Zone	Function	Sample VLAN	Sample IP Range
Availability Zone 1 Availability Zone 2	Management	1641 (Stretched)	172.16.41.0/24
Availability Zone 1	vSphere vMotion	1642	172.16.42.0/24
Availability Zone 1	vSAN	1643	172.16.43.0/24
Availability Zone 1	Host Overlay	1644	172.16.44.0/24
Availability Zone 1	Uplink01	1647	172.16.47.0/24
Availability Zone 1	Uplink02	1648	172.16.48.0/24
Availability Zone 1	Edge Overlay	1649	172.16.49.0/24
Availability Zone 2 Availability Zone 1	Management	1651 (Stretched)	172.16.51.0/24
Availability Zone 2	vSphere vMotion	1652	172.16.52.0/24
Availability Zone 2	vSAN	1653	172.16.53.0/24
Availability Zone 2	Host Overlay	1654	172.16.54.0/24
Availability Zone 2	Uplink01	1657	172.16.57.0/24

Table 3-1. Sample Values for VLANs and IP Ranges (continued)

Availability Zone	Function	Sample VLAN	Sample IP Range
Availability Zone 2	Uplink02	1658	172.16.58.0/24
Availability Zone 2	Edge Overlay	1659	172.16.59.0/24

Note The management network in each Availability Zone is stretched so that NSX-T Edge devices can fail over between availability zones.

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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-2. 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 3-3. 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. 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 ESXi Hosts management interfaces to enable forward (A), reverse (PTR), 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.

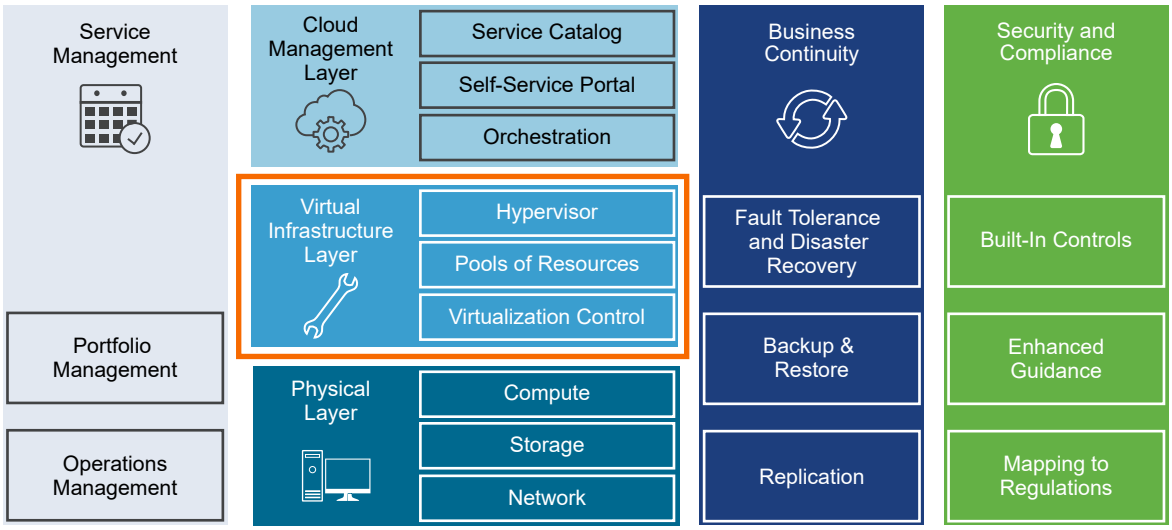
Table 3-4. Jumbo Frames Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-PHY-NET-008	Configure the MTU size to at least 1600 bytes (9000 recommended) on the physical switch ports, vSphere Distributed Switches, vSphere Distributed Switch port groups, N-VDS switches, and physical network link between availability zones, 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 for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

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

- vSAN Storage Design for NSX-T Workload Domains with Multiple Availability Zones

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

- Virtualization Network Design for NSX-T Workload Domains with Multiple Availability Zones

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

- NSX Design for NSX-T Workload Domains with Multiple Availability Zones

  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.

vSphere Cluster Design for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

  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.

- Shared Edge and Compute Cluster Design for NSX-T Workload Domains with Multiple Availability Zones

  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.

- Compute Cluster Design for NSX-T Workload Domains with Multiple Availability Zones

  As the SDDC expands, you can add compute clusters.

vSphere High Availability Design for NSX-T Workload Domains with Multiple Availability Zones

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 primary ESXi host. The primary ESXi host communicates with the vCenter Server system and monitors the virtual machines and secondary ESXi hosts in the cluster.

The primary ESXi host detects different types of failure:

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

Table 3-5. 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 you set HA Isolation Response to Power Off to restart the virtual machines on the 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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-6. 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 percentage-based failover based on half of the ESXi hosts in the cluster. For example, in a cluster with 8 ESXi hosts you configure admission control for 4 ESXi hosts failure 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 virtual machines from the failed ESXi hosts on any remaining ESXi hosts. Only half of a stretched cluster should be used to ensure that all VMs have enough resources in an availability zone outage.	You must add ESXi hosts to the cluster in pairs, one in each availability zone.
NSXT-VI-VC-005	Create a shared edge and compute cluster that consists of a minimum of eight ESXi hosts, four in each availability zone.	Allocating 4 ESXi hosts provides full redundancy for each availability zone within the cluster. Having 4 ESXi hosts in each availability zone guarantees vSAN and NSX redundancy during availability zone outages or maintenance operations.	8 ESXi hosts is the smallest starting point when using two availability zones for the shared edge and compute cluster for redundancy and performance thus increasing cost.

Table 3-6. 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 four large sized edge virtual machines, two per availability zone, with a CPU share level of High, a memory share of normal, and a 128-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 each availability zone in the cluster.	Using host profiles simplifies the configuration of ESXi hosts and ensures that settings are uniform across the cluster.	After NSX-T has been deployed you must update the host profile. Anytime an authorized change to an ESXi host is made, you must update the host profile to reflect the change or the status will show non-compliant. Because of configuration differences between availability zones, two host profiles are required and must be applied on each ESXi host.

Table 3-6. Shared Edge and Compute Cluster Design Decisions (continued)

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-VC-009	Set the cluster isolation addresses for the cluster to the gateway IP address for the vSAN network in both availability zones.	Allows vSphere HA to validate complete network isolation in the case of a connection failure between availability zones.	You must manually configure the isolation address.
NSXT-VI-VC-010	Set the advanced cluster setting das.usedefaultisolationaddress to false.	Ensures that the manual isolation addresses are used instead of the default management network gateway address.	None.

Compute Cluster Design for NSX-T Workload Domains with Multiple Availability Zones

As the SDDC expands, you can add compute 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-011	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.

vSAN Storage Design for NSX-T Workload Domains with Multiple Availability Zones

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

- vSAN Network Design for NSX-T Workload Domains with Multiple Availability Zones

  When you plan your network configuration, consider the overall traffic and decide how to isolate vSAN traffic.

- vSAN Cluster and Disk Group Design for NSX-T Workload Domains with Multiple Availability Zones

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

- vSAN Policy Design for NSX-T Workload Domains with Multiple Availability Zones

  After you enable and configure VMware vSAN, you can create storage policies that define the virtual machine storage characteristics. Storage characteristics specify different levels of service for different virtual machines.

vSAN Network Design for NSX-T Workload Domains with Multiple Availability Zones

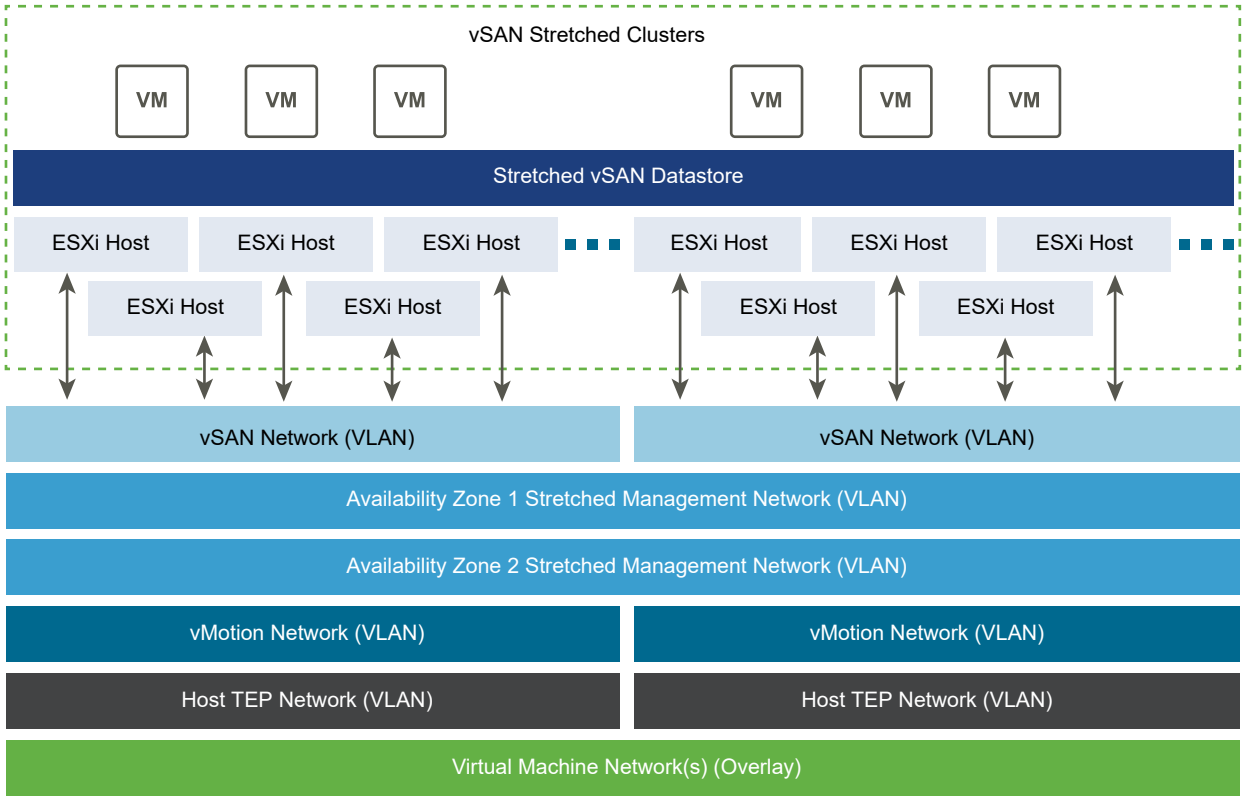
When you plan your network configuration, consider the overall traffic and decide how to isolate vSAN traffic.

vSAN Network Considerations

In case of two availability zones, the management VLAN in each availability zone, which the NSX-T Edge devices management interface connects to, must be stretched across both availability zones. The technology used to stretch the VLAN is out of scope and varies according to your existing infrastructure.

The network infrastructure between availability zones must support jumbo frames, and ensure that latency is less than 5 ms round trip.

Figure 3-4. VMware vSAN Conceptual Network with Two Availability Zones



Network Requirements

Use a minimum of a 10 Gbps Ethernet connection, with 25 Gbps Ethernet recommended, for use with vSAN to ensure the best and most predictable performance (IOPS) for the environment. If you use a slower connection, a significant decrease in array performance appears.

A minimum of 10 Gbps Ethernet also provides support for the use of a vSAN all-flash configurations.

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

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

In a stretched cluster configuration, vSAN requires a 5ms or less round trip latency between ESXi Hosts in each availability zone and 200ms between each ESXi host in each availability zone to the witness appliance.

Table 3-7. Network Design Decisions for vSAN Stretched Cluster

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-Storage-SDS-001	Use a minimum of 10 Gbps Ethernet (25 Gbps Ethernet recommended) for vSAN traffic.	Performance with 10 Gbps Ethernet is sufficient, while with 25 Gbps Ethernet is optimal. If the bandwidth is less than 10 Gbps Ethernet, array performance significantly decreases.	The physical network must support 10 Gb or 25 Gb networking between every ESXi host in the vSAN cluster.
NSXT-VI-Storage-SDS-002	Configure jumbo frames on the VLANs dedicated to vSAN traffic.	Jumbo frames are already used to improve performance of vSphere vMotion and NFS storage traffic.	Every device in the network must support jumbo frames.
NSXT-VI-Storage-SDS-003	Configure a dedicated VLAN in each availability zone, for each vSAN enabled cluster.	VLANs provide traffic isolation. vSAN traffic between availability zones is routed. An additional stretched VLAN is not required.	You have enough VLANs within each cluster and you can use them for traffic segregation. Static routes on the ESXi hosts are required.

vSAN Witness

When you use vSAN in a stretched cluster configuration, you must configure a vSAN stretched cluster witness host. This ESXi host must be configured in a third location that is not local to the ESXi hosts on either side of the stretched cluster.

This vSAN witness can be configured as a physical ESXi host or you can use the vSAN witness appliance.

Table 3-8. Design Decisions on the vSAN Witness

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-Storage-SDS-004	Deploy the vSAN witness appliance in a third physical location.	A location outside of both availability zones is required to provide an appropriate quorum.	A third physically separate location is required when implementing a vSAN stretched cluster between two locations.
NSXT-VI-Storage-SDS-005	Each availability zone must have a unique route to the third physical location.	This ensures connectivity to each physical location in the event of an availability zone failure.	You can not route witness traffic from one availability zone through another.

vSAN Cluster and Disk Group Design for NSX-T Workload Domains with Multiple Availability Zones

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

vSAN Datastore Size

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

Table 3-9. Design Decisions on the vSAN Datastore

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-Storage-SDS-006	On all vSAN datastores, ensure at least 30% of free space is always available.	When vSAN reaches 80% usage, a rebalance task is started which can be resource-intensive.	Increases the amount of available storage needed.

Number of ESXi Hosts Per Cluster

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

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

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

Cluster Size Design Background

Design Quality	3 ESXi Hosts	32 ESXi Hosts	64 ESXi Hosts	Comments
Availability	↓	↑	↑↑	The more ESXi hosts in the cluster, the more failures the cluster can tolerate.
Manageability	↓	↑	↑	The more ESXi hosts in the cluster, the more virtual machines can run in the vSAN environment.

Design Quality	3 ESXi Hosts	32 ESXi Hosts	64 ESXi Hosts	Comments
Performance	↑	↓	↓	A larger cluster can impact performance if there is an imbalance of resources. Consider performance as you make your decision.
Recoverability	o	o	o	Neither design option impacts recoverability.
Security	o	o	o	Neither design option impacts security.

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

Table 3-10. Design Decision on the Cluster Size for vSAN

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-Storage-SDS-007	Two Availability Zones requires a cluster with a minimum of 8 ESXi hosts (4 in each availability zone) to support a stretched vSAN configuration.	Having 8 ESXi hosts addresses the availability and sizing requirements, and allows you to take an availability zone offline for maintenance or upgrades without impacting the overall vSAN cluster health.	The availability requirements for the cluster might cause underutilization of the cluster's ESXi hosts.

Number of Disk Groups Per ESXi Host

Disk group sizing is an important factor during volume design.

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

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

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

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

Disk Groups Design Background

The number of disk groups can affect availability and performance.

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

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

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-Storage-SDS-008	Configure vSAN with a minimum of one disk group per ESXi host.	Single disk group provides a cost effective solution with the minimum required performance and usable space for the datastore.	If an SSD in an ESXi host becomes unavailable, the host takes the disk group offline. If you use two or more disk groups, you can increase availability and performance.

vSAN Policy Design for NSX-T Workload Domains with Multiple Availability Zones

After you enable and configure VMware vSAN, you can create storage policies that define the virtual machine storage characteristics. Storage characteristics specify different levels of service for different virtual machines.

The default storage policy tolerates a single failure and has a single disk stripe. If you configure a custom policy, vSAN must guarantee its application. However, if vSAN cannot guarantee a policy, you cannot provision a virtual machine that uses the policy, unless you enable force provisioning.

VMware vSAN Policy Options

A storage policy includes several attributes, which can be used alone or combined to provide different service levels. Policies can be configured for availability and performance conservatively, to balance the consumed space and recoverability properties. In most scenarios, the default system policy is adequate and you do not need additional policies. Policies allow any configuration to be as customized as needed for the business requirements of the application .

Policy Design Background

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

Table 3-11. VMware vSAN Policy Options

Capability	Use Case	Default Value	Maximum Value	Comments
Number of failures to tolerate	Redundancy	1	3	A standard RAID 1 mirrored configuration that provides redundancy for a virtual machine disk. The higher the value, the more failures can be tolerated. For n failures tolerated, n+1 copies of the disk are created, and 2n+1 ESXi hosts contributing storage are required. A higher n value indicates that more replicas of virtual machines are made, which can consume more disk space than expected.
Number of disk stripes per object	Performance	1	12	Use a standard RAID 0 stripe configuration to increase performance for a virtual machine disk. The setting defines the number of HDDs on which each replica of a storage object is striped. If the value is higher than 1, you get an increased performance. However you can also get an increase in system resource usage.

Table 3-11. VMware vSAN Policy Options (continued)

Capability	Use Case	Default Value	Maximum Value	Comments
Flash read cache reservation (%)	Performance	0%	100%	Flash capacity reserved as read cache for the storage is a percentage of the logical object size that is reserved for that object. Only use the setting for workloads if you must address read performance issues. The downside is that other objects cannot use a reserved cache. VMware recommends to not use these reservations unless it is absolutely necessary because unreserved flash is shared fairly among all objects.

Table 3-11. VMware vSAN Policy Options (continued)

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

By default, policies are configured based on application requirements. However, you apply them differently per object.

Table 3-12. Object Policy Defaults

Object	Policy	Comments
Virtual machine namespace	Failures to tolerate: 1	Configurable. Changes are not recommended.
Swap	Failures to tolerate: 1	Configurable. Changes are not recommended.

Table 3-12. Object Policy Defaults (continued)

Object	Policy	Comments
Virtual disks	User-configured storage policy	Can be any storage policy configured on the system.
Virtual disk snapshots	Uses virtual disk policy	Same as virtual disk policy by default. Changes are not recommended.

Note If you do not specify a user-configured policy, vSAN uses a default system policy of 1 failure to tolerate and 1 disk stripe for virtual disks and virtual disk snapshots. To ensure protection for these critical virtual machine components, policy defaults for the VM namespace and swap are set statically and are not configurable. Configure policies according to the business requirements of the application. By using policies, vSAN can adjust the performance of a disk on demand.

Policy Design Recommendations

Policy design starts with assessment of business needs and application requirements. Assess use cases for VMware vSAN to determine the necessary policies. First, assess the following application requirements:

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

After assessment, configure the software-defined storage module policies for availability and performance in a conservative manner so that you balance consumed space and recoverability. Default system policy might be adequate and you might not need additional policies, unless you have specific requirements for performance or availability.

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-Storage-SDS-009	For vSAN Stretched clusters, edit the default VMware vSAN storage policy and change the Site Disaster Tolerance setting from None to Dual Site Mirroring (stretched cluster).	Provides protection for virtual machines in each availability zone, with the ability to recover from an availability zone outage.	You might need additional policies if third-party VMs must be hosted in these clusters due to performance or availability requirements that differ from what the default VMware vSAN policy supports.
NSXT-VI-Storage-SDS-010	Leave the default virtual machine swap file as a sparse object on VMware vSAN.	Creates virtual swap files as a sparse object on the vSAN datastore. Sparse virtual swap files only consume capacity on vSAN when accessed. As a result, you can reduce the consumption on the vSAN datastore if virtual machines do not experience memory over-commitment which requires the use of the virtual swap file.	None.

Virtualization Network Design for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

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

- Virtual Switches for NSX-T Workload Domains with Multiple Availability Zones

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

- NIC Teaming for NSX-T Workload Domains with Multiple Availability Zones

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

- Geneve Overlay for NSX-T Workload Domains with Multiple Availability Zones

  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.

- vMotion TCP/IP Stack for NSX-T Workload Domains with Multiple Availability Zones

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

Virtual Network Design Guidelines for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

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

- Virtual Switch Design Background for NSX-T Workload Domains with Multiple Availability Zones

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

- Virtual Switch Design Decisions for NSX-T Workload Domains with Multiple Availability Zones

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

- Shared Edge and Compute Cluster Switches for NSX-T Workload Domains with Multiple Availability Zones

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

Virtual Switch Design Background for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

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.	The N-VDS is required for overlay traffic.	Management is shifted from the vSphere Client to the NSX Manager.

Shared Edge and Compute Cluster Switches for NSX-T Workload Domains with Multiple Availability Zones

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 3-5. Virtual Switch Design for ESXi Hosts in the Shared Edge and Compute Cluster

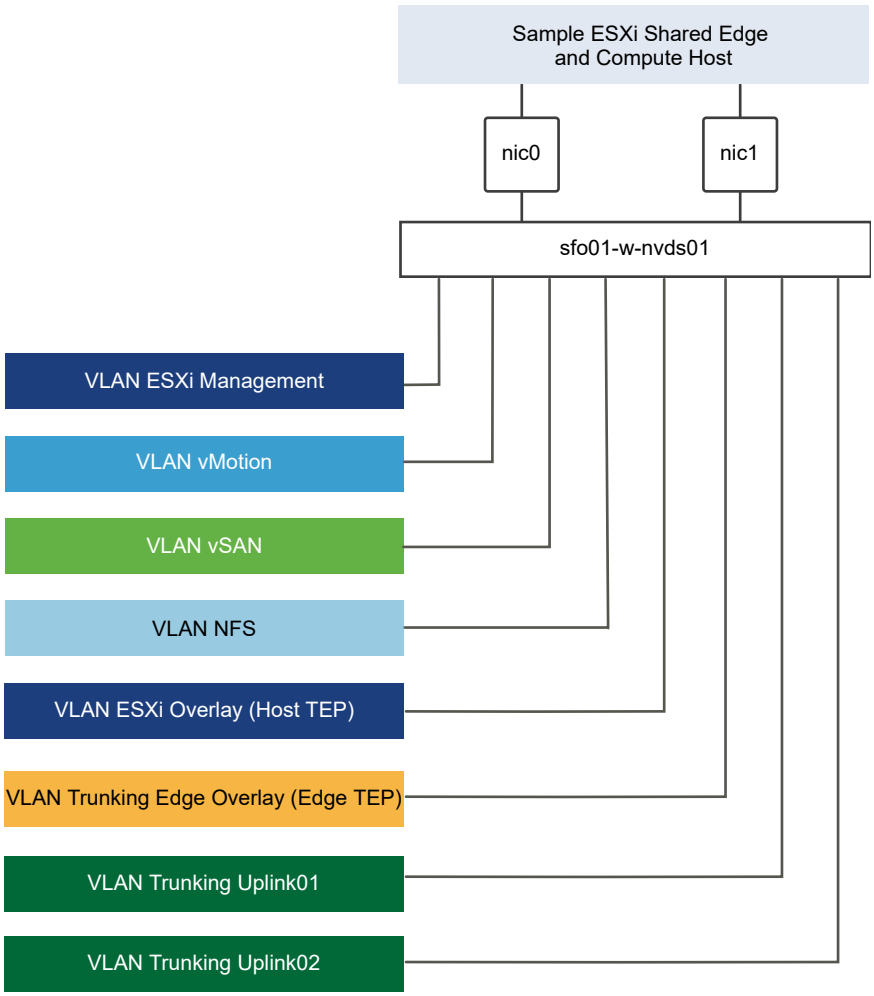


Table 3-13. Virtual Switches for the Shared Edge and Compute Cluster

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 the ESXi traffic ■ Failover order for the 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 3-14. 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 3-6. Segment Configuration on an ESXi Host That Runs an NSX-T Edge Node

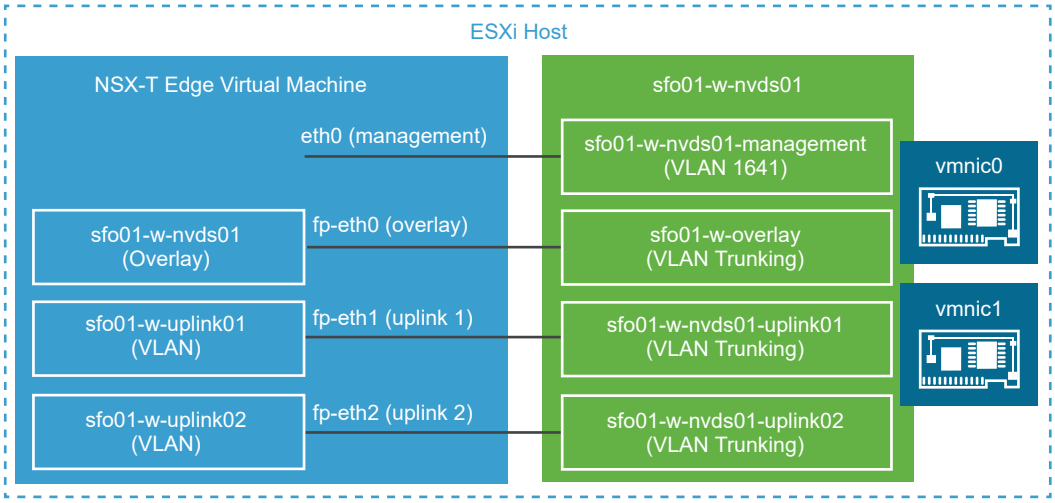


Table 3-15. Segments in the Shared Edge and Compute Cluster

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

Table 3-15. Segments in the Shared Edge and Compute Cluster (continued)

N-VDS Switch	Segment Name
sfo01-w-nvds01	sfo01-w-nvds01-overlay
sfo01-w-nvds01	sfo01-w-nvds01-uplink01
sfo01-w02-uplink01	sfo01-w02-uplink01
sfo01-w02-uplink02	sfo01-w02-uplink02
sfo01-w-nvds01	sfo02-w-nvds01-management
sfo01-w-nvds01	sfo02-w-nvds01-vmotion
sfo01-w-nvds01	sfo02-w-nvds01-vsan
sfo01-w-nvds01	sfo02-w-nvds01-nfs
sfo02-w02-uplink01	sfo02-w02-uplink01
sfo02-w02-uplink02	sfo02-w02-uplink02

Table 3-16. 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	--
sfo01-w-nvds01	sfo02-w-nvds01-management	Management Traffic
sfo01-w-nvds01	sfo02-w-nvds01-vmotion	vMotion Traffic
sfo01-w-nvds01	sfo02-w-nvds01-vsan	vSAN
sfo01-w-nvds01	sfo02-w-nvds01-nfs	--
sfo01-w-nvds01	*auto created* (Host TEP)	--
sfo01-w-nvds01	*auto created* (Host TEP)	--
sfo01-w-nvds01	*auto created* (Hyperbus)	--

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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-17. 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 3-18. 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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-19. 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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-20. 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 for NSX-T Workload Domains with Multiple Availability Zones

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 for NSX-T Workload Domains with Multiple Availability Zones

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

- NSX-T Components for NSX-T Workload Domains with Multiple Availability Zones

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

- NSX-T Network Requirements and Sizing for NSX-T Workload Domains with Multiple Availability Zones

  NSX-T requirements impact both physical and virtual networks.

- NSX-T Network Virtualization Conceptual Design for NSX-T Workload Domains with Multiple Availability Zones

  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.

- Cluster Design for NSX-T Workload Domains with Multiple Availability Zones

  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.

- Replication Mode of Segments for NSX-T Workload Domains with Multiple Availability Zones

  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).

- Transport Zone Design for NSX-T Workload Domains with Multiple Availability Zones

  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.

- Network I/O Control Design for NSX-T Workload Domains with Multiple Availability Zones

  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.

- Transport Node and Uplink Policy Design for NSX-T Workload Domains with Multiple Availability Zones

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

- Routing Design by Using NSX-T for NSX-T Workload Domains with Multiple Availability Zones

  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.

- Virtual Network Design Example Using NSX-T for NSX-T Workload Domains with Multiple Availability Zones

  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.

- Monitoring NSX-T for NSX-T Workload Domains with Multiple Availability Zones

  Monitor the operation of NSX-T for identifying failures in the network setup by using vRealize Log Insight and vRealize Operations Manager.

- Use of SSL Certificates in NSX-T for NSX-T Workload Domains with Multiple Availability Zones

  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.

NSX-T Design for NSX-T Workload Domains with Multiple Availability Zones

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 3-21. NSX-T Design Decisions

Design ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-001	Deploy three NSX-T Manager appliances 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. As of NSX-T 2.4, the the NSX-T Manager also serves the role of the NSX-T Controllers. You deploy three nodes in the cluster for availability of services.	You must install and configure NSX-T Manager in a highly available management cluster. Management cluster requires four physical ESXi hosts for redunancy.
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.

NSX-T Components for NSX-T Workload Domains with Multiple Availability Zones

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 3-22. 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 greater 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 VMware Enterprise 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 3-23. 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 3-24. 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, vRealize Log Insight, vRealize Operations Manger, 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 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 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 for NSX-T Workload Domains with Multiple Availability Zones

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.	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 3-4. Jumbo Frames Design Decisions.
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.	Enables administrators to access the NSX-T envrionement via FQDN as opposed to memorizing IP addresses.

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 3-25. 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, PoC only)	4 (Small, PoC only)	200 (Small, PoC only)	Numbers are different according to the use case. At least two edge devices are required to enable ECMP routing.
	4 (Medium)	8 (Medium)	200 (Medium)	
	8 (Large)	32 (Large)	200 (Large)	

Table 3-26. Design Decisions on Sizing the NSX-T Edge Virtual Machines

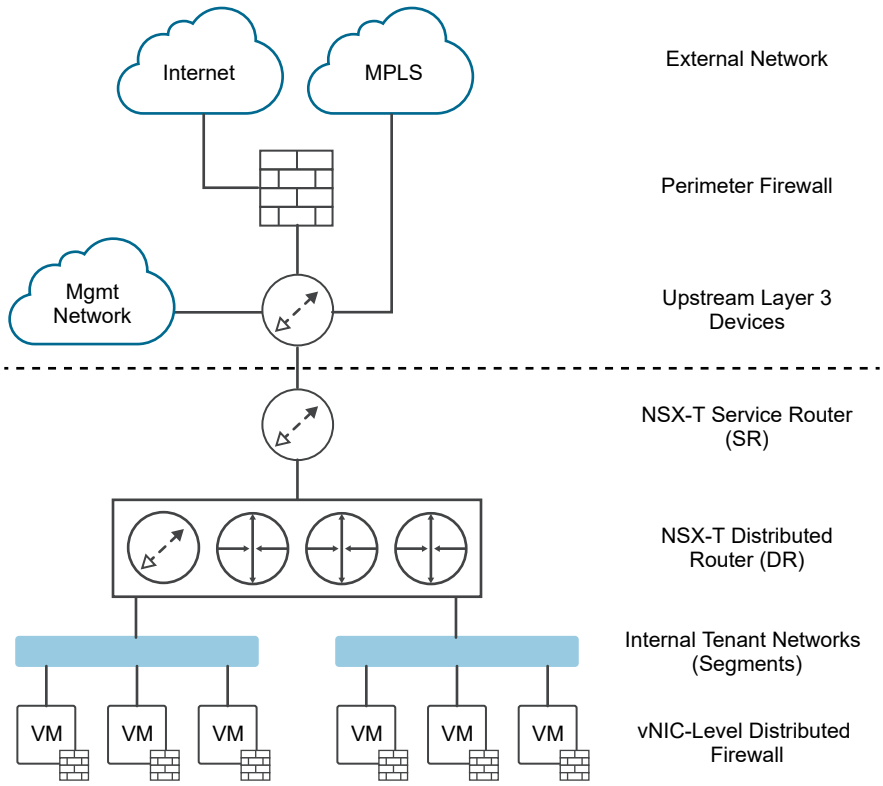
Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-008	Use large-size NSX-T Edge virtual machines.	The large-size appliance provides the required performance characteristics if a failure occurs. Virtual Edges provide simplied lifecycle management.	Large size Edges consume more CPU and Memory resources.

NSX-T Network Virtualization Conceptual Design for NSX-T Workload Domains with Multiple Availability Zones

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 3-7. 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 Service Router (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 Distributed Router (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 Tenant Networks

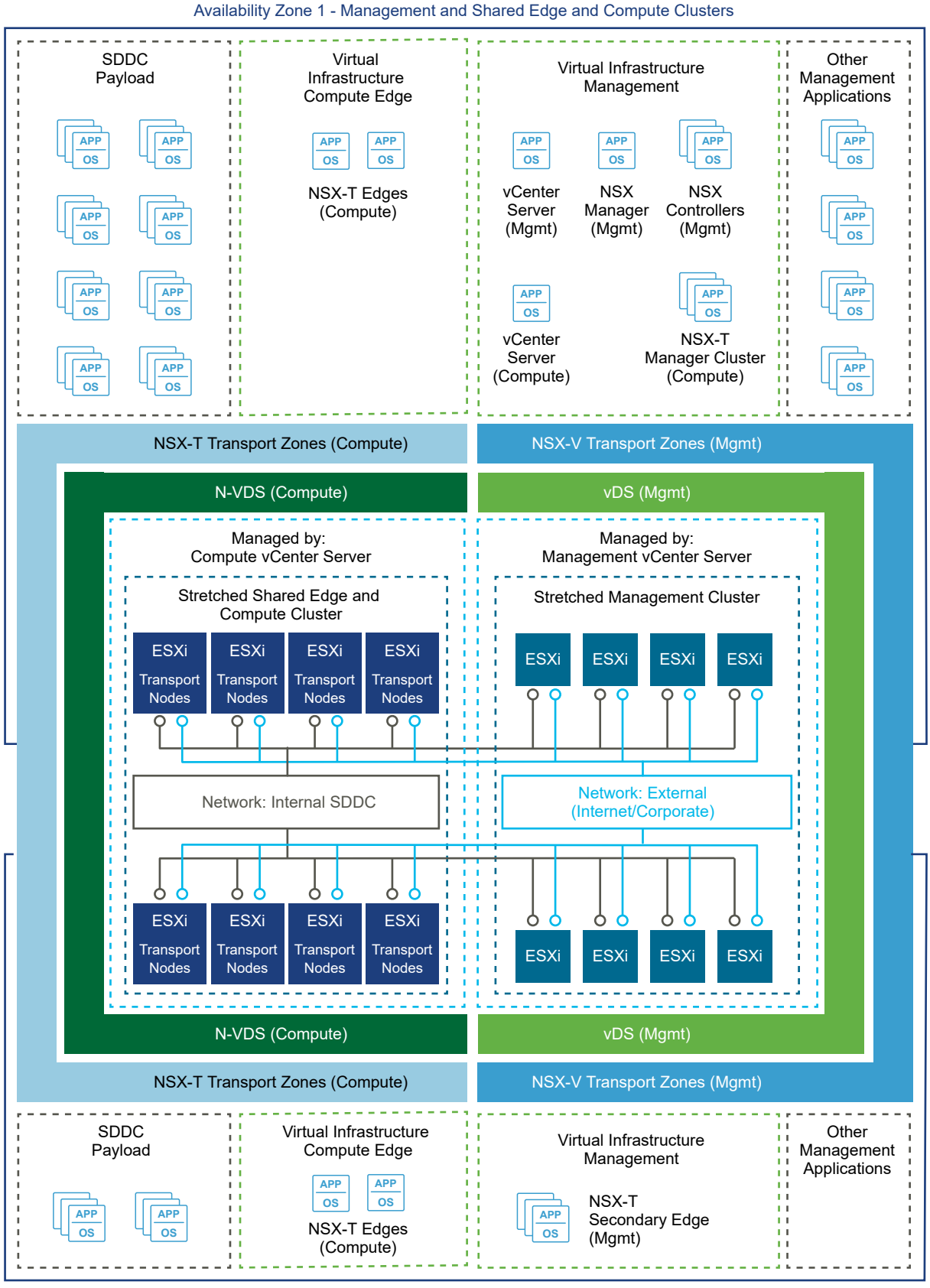
Internal tenant networks are NSX-T Segments and provide connectivity for the tenant workloads. Workloads are directly connected to these networks. Internal tenant networks are then connected to a DR.

Cluster Design for NSX-T Workload Domains with Multiple Availability Zones

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 3-8. 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 and network services for 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 3-27. Cluster Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-009	For virtual infrastructure workload domains, do not dedicate an edge cluster in vSphere.	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, to avoid an impact on network performance, you must scale out the cluster as you add tenant workloads.
NSXT-VI-SDN-010	Deploy at least four large-size NSX-T Edge virtual machines, two in each availability zone, in the shared edge and compute cluster.	Creates the 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 VM-VM anti-affinity rules in vSphere Distributed Resource Scheduler (vSphere DRS) to the NSX-T Manager appliances.	Keeps the NSX-T Manager appliances running on different ESXi hosts for high availability.	Requires at least four physical hosts to guarantee the three NSX-T Manager appliances continue to run if an ESXi host failure occurs. Additional configuration is required to set up anti-affinity rules.

Table 3-27. Cluster Design Decisions (continued)

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-012	Create two VM-VM anti-affinity rules for vSphere DRS, one per availability zone, for the edge virtual machines in each availability zone.	Keeps the NSX-T Edge virtual machines in an availability zone running on different ESXi hosts for high availability.	You must perform additional configuration to set up the anti-affinity rules.
NSXT-VI-SDN-013	Create a virtual machne group that contains the edge and tenant virtual machines in Availability Zone 1.	Ensures that virtual machines are located only in the assigned availability zone.	You must add VMs to the allocated group manually to ensure they are not powered-on in or migrated to the wrong availability zone.
NSXT-VI-SDN-014	Create a VM group that contains the edge virtual machines and tenant workloads in Availability Zone 2.	Ensures that virtual machines are located only in the assigned availability zone.	You must add VMs to the allocated group manually to ensure they are not powered-on in or migrated to the wrong availability zone.
NSXT-VI-SDN-015	Create a host group that contains the ESXi hosts in Availability Zone 1.	Makes it easier to manage which virtual machines should run in which availability zone.	You must create and maintain VM/Host DRS groups.
NSXT-VI-SDN-016	Create a host group that contains the ESXi hosts in Availability Zone 2.	Makes it easier to manage which virtual machines should run in which availability zone.	You must create and maintain VM/Host DRS groups.
NSXT-VI-SDN-017	Create a VM/Host DRS rule that specifies that VM's in the VM Availability Zone 1 group should run on ESXi hosts in the Availability Zone 1 group.	Ensures VM's that should run in Availability Zone 1 do not get migrated to Availability Zone 2. The should rule allows the VM's to failover to Availability Zone 2 in the case of a failure.	You must create VM to Host DRS rules.
NSXT-VI-SDN-018	Create a VM/Host DRS rule that specifies that VM's in the VM Availability Zone 2 group should run on ESXi hosts in the Availability Zone 2 group.	Ensures VM's that should run in Availability Zone 2 do not get migrated to Availability Zone 1. The should rule allows the VM's to failover to Availability Zone 2 in the case of a failure.	You must create VM to Host DRS 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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-28. 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 3-29. Design Decisions on Segment Replication Mode

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-019	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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-30. Transport Zones Design Decisions

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-020	Create a single transport zone for all overlay traffic across all workload domains.	Ensures all Segments are available to all ESXi hosts and edge virtual machines configured as Transport Nodes.	None.
NSXT-VI-SDN-021	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-022	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 Workload Domains with Multiple Availability Zones

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 3-31. Network I/O Control Design Decisions

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-SDN-023	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-024	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-025	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-026	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-027	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.

Table 3-31. Network I/O Control Design Decisions (continued)

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-SDN-028	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, its priority is lower than the priority of the vSAN traffic.	During times of network contention, backups are slower than usual.
NSXT-VI-SDN-029	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-030	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-031	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-032	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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-32. Design Decisions on Transport Nodes and Uplink Policy

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-033	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-034	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 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-035	Create two uplink profiles per Availability Zone with the failover order teaming policy with one active uplink and no standby uplinks for edge virtual machine uplink traffic.	Enables ECMP in each Availability Zone because the edge virtual machine can uplink to the physical network over two different VLANs.	You create and manage more uplink profiles.
NSXT-VI-SDN-036	Configure each ESXi host as a Transport Node without the use of Transport Node Profiles.	Enables the participation of ESXi hosts and the virtual machines on them in NSX-T overlay and VLAN networks. Transport Node Profiles can only be applied at the cluster level, because each Availability Zone is a member of different VLAN's a Tranport Node Profile can not be used.	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.

Table 3-32. Design Decisions on Transport Nodes and Uplink Policy (continued)

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-037	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-038	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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-33. Design Decisions on Routing Using NSX-T

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-039	Create two VLANs per Availability Zone 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-040	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 Enterprise PKS, you must deploy a Tier-0 Gateway in Active-Standby mode.
NSXT-VI-SDN-041	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-042	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-043	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-044	Create an IP Prefix List that consists of 'any' for the network and Permit as the action instead of using the default IP Prefix List.	Will be used in a Route Map to configure AS Path Prepend for Availability Zone 2 BGP neighbors.	Requires manually creating an IP Prefix List that is identical to the default.

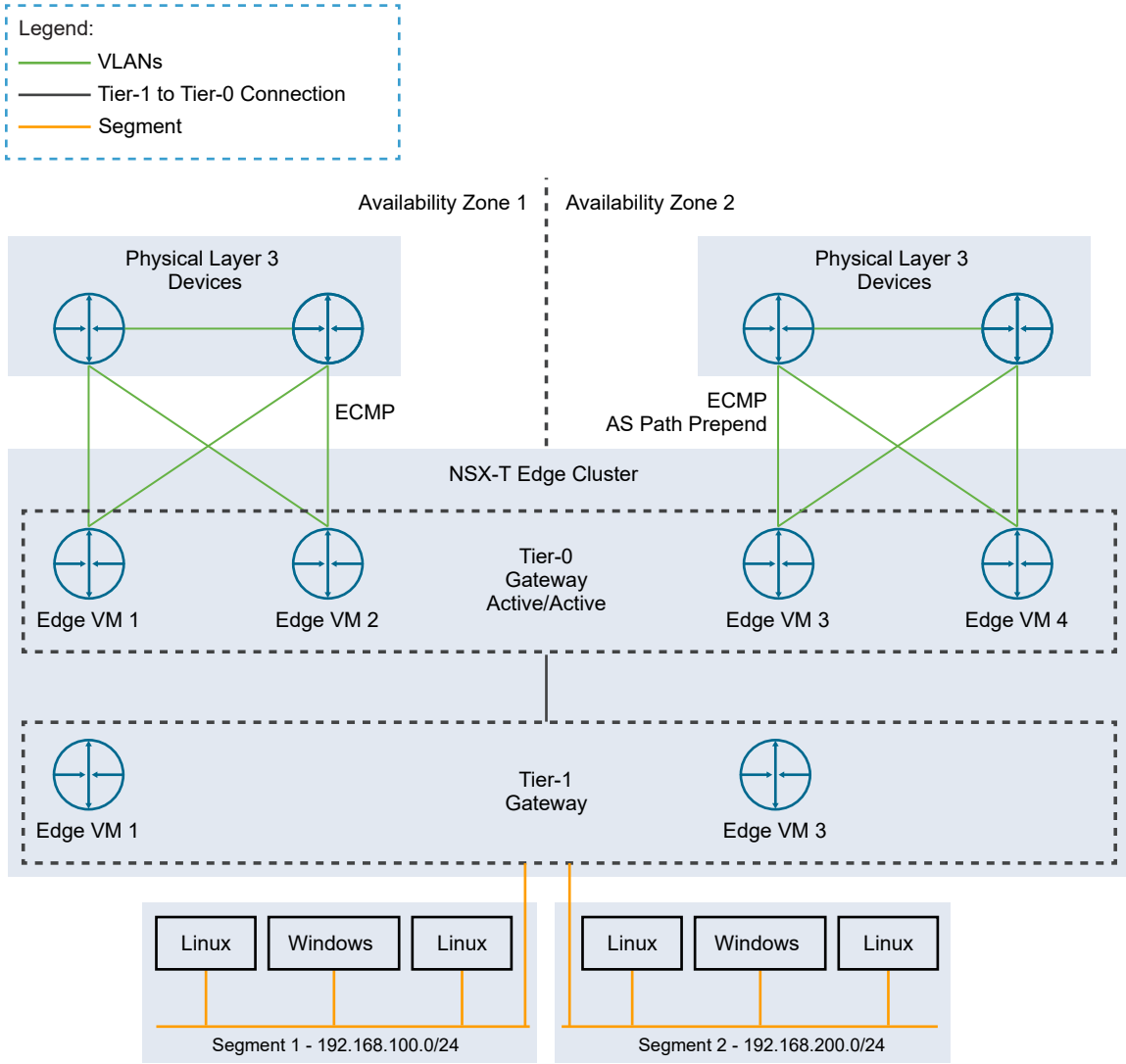
Table 3-33. Design Decisions on Routing Using NSX-T (continued)

Decision ID	Design Decision	Design Justification	Design Implications
NSXT-VI-SDN-045	Create a Route Map that contains the newly created IP Prefix and the Tier-0 Local AS added twice as the AS Path Prepend value.	Will be used when configuring neighbor relationships with Availability Zone 2 L3 devices. Ensures all ingress/egress traffic goes through Availability Zone 1.	Requires manually creating the Route Map. The two edge VMs in Availability Zone 2 will not route north/south traffic unless Availability Zone 1 edge VMs lose their BGP neighbors, such as ToR or Availability Zone Failures.
NSXT-VI-SDN-046	Configure Availability Zone 2 neighbors to use the newly created Route Map as In and Out filters.	Makes the path in and out of Availability Zone 2 less preferred as the AS Path is longer, which results in all traffic going through Availability Zone 1.	The two edge VMs in Availability Zone 2 will not route north/south traffic unless Availability Zone 1 edge VMs lose their BGP neighbors, such as ToR or Availability Zone Failures.
NSXT-VI-SDN-047	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.
NSXT-VI-SDN-048	Deploy a Tier-1 Gateway to the NSX-T Edge cluster, choosing one Edge Transport Node from each Availability Zone 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. Ensures that during an Availability Zone outage any services running on the Tier-1, such as a load balancer, will quickly failover to the standby edge.	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.
NSXT-VI-SDN-049	Deploy Tier-1 Gateways with the Non-Premptive setting.	Ensures that when the failed Edge Transport Node comes back online it doesn't move services back to itself resulting in a small service outage.	None.

Virtual Network Design Example Using NSX-T for NSX-T Workload Domains with Multiple Availability Zones

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 3-9. Virtual Network Example



Monitoring NSX-T for NSX-T Workload Domains with Multiple Availability Zones

Monitor the operation of NSX-T for identifying failures in the network setup by using vRealize Log Insight and vRealize Operations Manager.

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

- vRealize Operations Manger provides alerts, capacity management and custom views and dashboards.

Table 3-34. Design Decisions on Monitoring NSX-T

Decision ID	Design Decision	Design Justification	Design Implication
NSXT-VI-SDN-050	Install the content pack for NSX-T in vRealize Log Insight.	Add a dashboard and metrics for granular monitoring of the NSX-T infrastructure.	Requires manually installing the content pack.
NSXT-VI-SDN-051	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 for NSX-T Workload Domains with Multiple Availability Zones

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 3-35. Design Decisions on the SSL Certificate of NSX-T Manager

Design ID	Design Decision	Design Justification	Design Implication
NSXT-VI-SDN-052	Replace the certificate of the NSX-T Manager instances 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 instance is encrypted by using a trusted certificate.	Replacing and managing certificates is an operational overhead.
NSXT-VI-SDN-053	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 address of the NSX-T Manager cluster and administrators is encrypted by using a trusted certificate.	Replacing and managing certificates is an operational overhead.