VMware Cloud Foundation 4.5 on VxRail 7.0 Architecture Guide

October 2022

Abstract

This guide introduces the architecture of the VMware Cloud Foundation (VCF) on VxRail solution. It describes the different components within the solution and also acts as an aid to selecting the configuration needed for your business requirements.

DELL Technologies

Copyright

Contents

Chapter 1 Executive Summary

This chapter presents the following topics:

VMware Cloud Foundation on VxRail

VMware Cloud Foundation (VCF) on VxRail™ is a Dell and VMware jointly engineered integrated solution. It contains features that simplify, streamline, and automate the operations of your entire Software-Defined Datacenter (SDDC) from Day 0 through Day 2. The new platform delivers a set of software-defined services for compute (with vSphere and vCenter), storage (with vSAN), networking (with NSX), security, and cloud management (with vRealize Suite). These services apply to both private and public environments, making it the operational hub for your hybrid cloud.

VCF on VxRail provides the simplest path to the hybrid cloud through a fully integrated hybrid cloud platform. This platform leverages native VxRail hardware and software capabilities and other VxRail-unique integrations (such as vCenter plugins and Dell networking). These components work together to deliver a new turnkey hybrid cloud user experience with full-stack integration. Full-stack integration means you get both HCI infrastructure layer and cloud software stack in one complete automated life-cycle turnkey experience.

Document purpose

This guide introduces the architecture of the VCF on VxRail solution. It describes the different components within the solution. It is also an aid to selecting the configuration needed for your business requirements.

Intended audience

This architecture guide is intended for executives, managers, cloud architects, network architects, and technical sales engineers who are interested in designing or deploying an SDDC or Hybrid Cloud Platform to meet the needs or the business requirements. Readers should be familiar with the VMware vSphere, NSX, vSAN, and vRealize product suites in addition to general network architecture concepts.

Revisions

Date	Description
April 2019	Initial release
March 2020	Updated to support VCF 3.9.1 and PKS, DR guidance removed.
May 2020	Updated to support VCF 4.0.
July 2020	Updated to support VCF 4.0.1
November 2020	Updated to support VCF 4.1 and VxRail 7.0.100
March 2021	Updated to support VCF 4.2 and VxRail 7.0.131
October 2021	Updated to support VCF 4.3 and VxRail 7.0.241
February 2022	Updated to support VCF 4.4 and VxRail 7.0.320
October 2022	Updated to support VCF 4.5 and VxRail 7.0.400

Chapter 2 Architecture Overview

This chapter presents the following topics:

Introduction

You can virtualize all your infrastructure and deploy a full VMware SDDC with the benefit of automated SDDC life cycle management (LCM) by implementing a standardized VMware SDDC architecture on VxRail with Cloud Foundation. This solution includes NSX for Network Virtualization and Security, vSAN for SDS, vSphere 7 for Kubernetes, Tanzu Kubernetes Grid, and SDDC Manager for SDDC LCM.

By virtualizing all your infrastructure, you can take advantage of what a fully virtualized infrastructure can provide, such as resource utilization, workload and infrastructure configuration agility, and advanced security. With SDDC software life-cycle automation provided by Cloud Foundation (and in particular SDDC Manager which is a part of Cloud Foundation on top of VxRail), you can streamline the LCM experience for the full SDDC software and hardware stack.

You no longer need to worry about performing updates and upgrades manually using multiple tools for all the SDDC SW and HW components of the stack. These processes are now streamlined using a common management toolset in SDDC Manager with VxRail Manager. You can begin to leverage the data services benefits that a fully virtualized infrastructure can offer along with SDDC infrastructure automated LCM. An example of data services is using software-defined networking features from NSX like microsegmentation, which before software-defined networking tools, was nearly impossible to implement using physical networking tools.

Another important aspect is the introduction of a standardized architecture for how these SDDC components are deployed together using Cloud Foundation, an integrated cloud software platform. Having a standardized design incorporated as part of the platform provides you with a guarantee that these components have been certified with each other and are backed by Dell Technologies. You can then be assured that there is an automated and validated path forward to get from one known good state to the next across the end-to-end stack.

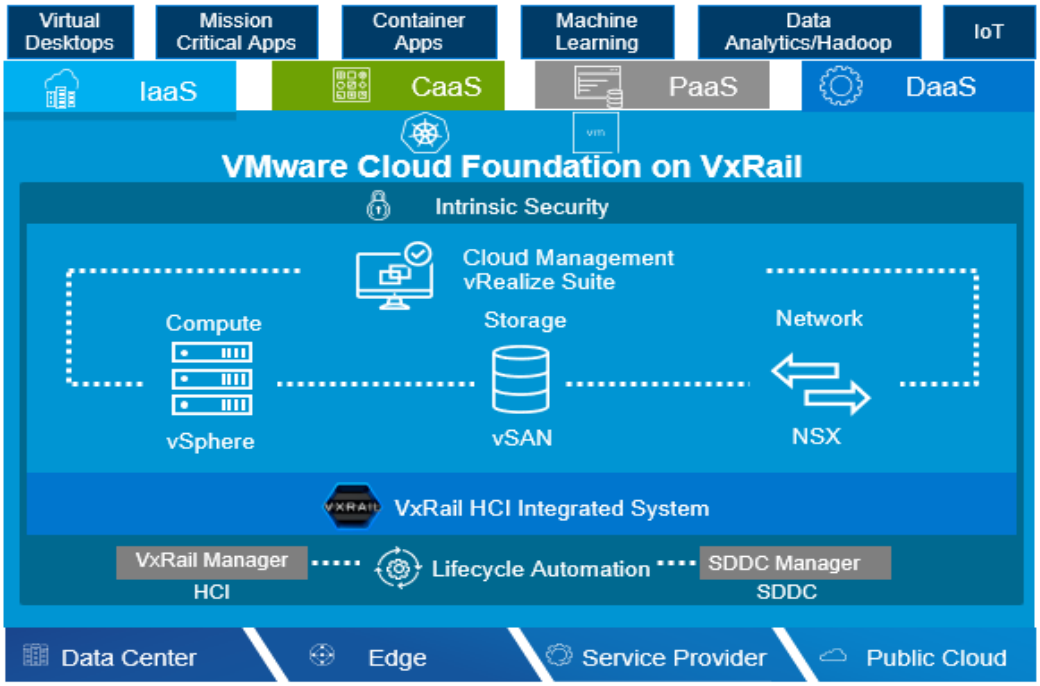

Figure 1. **Architecture Overview**

VxRail Manager

VCF on VxRail uses VxRail Manager to deploy and configure VxRail clusters that are powered by vSAN or external storage. It is also used to perform the LCM of ESXi, vSAN, and HW firmware using a fully integrated and seamless SDDC Manager orchestrated process. It monitors the health of hardware components and provides remote service support. This level of integration provides a truly unique turnkey hybrid cloud experience not available on any other infrastructure.

VxRail Manager provides the glue for the HCI hardware and software and is all life cycle managed together. Focusing on the glue and automation across the deployment, updating, monitoring, and maintenance phases of product life cycle, VxRail Manager delivers value by removing the need for heavy operational staffing. This automation improves operational efficiency. It reduces LCM risk, and significantly changes the focus of staff by providing value back to the business rather than expending time on maintaining the infrastructure.

SDDC Manager

SDDC Manager orchestrates the deployment, configuration, and (LCM) of vCenter and NSX above the ESXi and vSAN layers of VxRail. It unifies multiple VxRail clusters as workload domains (WLDs) or as multiple WLD. For multiple-availability zones (Multi-AZs), SDDC Manager creates the VxRail vSAN stretched cluster configuration for a dual-availability zone (AZ) WLD.

Network virtualization

VMware NSX Data Center is the network virtualization and security platform that enables the virtual cloud network. It is a software-defined approach to networking that extends across data centers, clouds, endpoints, and edge locations. With NSX Data Center, network functions—including switching, routing, firewalling, and load balancing—are brought closer to the application and distributed across the environment. Similar to the operational model of virtual machines, networks can be provisioned and managed independent of underlying hardware.

NSX Data Center reproduces the entire network model in software, enabling any network topology—from simple to complex multitier networks—to be created and provisioned in seconds. You can create multiple virtual networks with diverse requirements, leveraging a combination of the services that NSX offers. These services include microsegmentation or from a broad ecosystem of third-party integrations ranging from next-generation firewalls to performance management solutions to build inherently more agile and secure environments. These services can then be extended to several endpoints within and across clouds.

Operations Management

vRealize Operations is a self-driving operations management platform for the VMware Cloud and beyond. It incorporates AI and predictive analytics to deliver continuous performance optimization, efficient capacity and cost management, intelligent troubleshooting and remediation, and integrated compliance.

Logging and analytics

Another component of the VMware SDDC is VMware vRealize Log Insight™. It delivers heterogeneous and highly scalable log management with intuitive, actionable dashboards, sophisticated analytics, and broad third-party extensibility, providing deep operational visibility and faster troubleshooting.

Self-service cloud

vRealize Automation is the main consumption portal for the VMware Cloud and beyond. You use vRealize Automation to author, administer, and consume application templates and blueprints. As an integral component of VCF, vRealize Automation provides a unified service catalog that gives IT or end-users the ability to select and perform requests to instantiate specific services.

Chapter 3 Workload Domain Architecture

This chapter presents the following topics:

Introduction

A Workload Domain (WLD) consists of one or more Dell VxRail clusters that are managed by one vCenter Server instance. WLDs are connected to a network core that distributes data between them. WLDs can include different combinations of VxRail clusters and network equipment which can be set up with varying levels of hardware redundancy.

From the VxRail clusters, you can organize separate pools of capacity into WLDs, each with its own set of specified CPU, memory, and storage requirements to support various workload types such as Horizon or business-critical apps like Oracle databases. As new VxRail physical capacity is added by the SDDC Manager, it is made available for consumption as part of a WLD.

There are two types of WLDs that can be deployed:

- A Management WLD (Mgmt WLD), single per VCF instance
- A Virtual Infrastructure (VI) WLD, also known as a tenant WLD

More detail about each type of WLD is provided in the next section.

Management WLD

The VCF Management WLD VxRail cluster requires a minimum of four hosts on which the infrastructure components used to instantiate and manage the private cloud infrastructure run. The Management WLD is created during initial system install (or bring-up) using the VCF Cloud Builder tool.

In the Management WLD VxRail cluster, vSphere runs with a dedicated vCenter server that is backed by vSAN storage. It hosts the SDDC Manager, VxRail Manager VMs, and NSX-T Managers. vRealize Log Insight for Management domain logging and vRealize Operations and vRealize Automation are optional and must be manually deployed following VVD guidance. The management VxRail cluster must have a minimum of four hosts to provide vSAN FTT=1 during maintenance operations.

While the deployment and configuration of the management VxRail cluster is fully automated, once it is running, you manage it just like you would any other VxRail cluster using the vSphere HTML5 client.

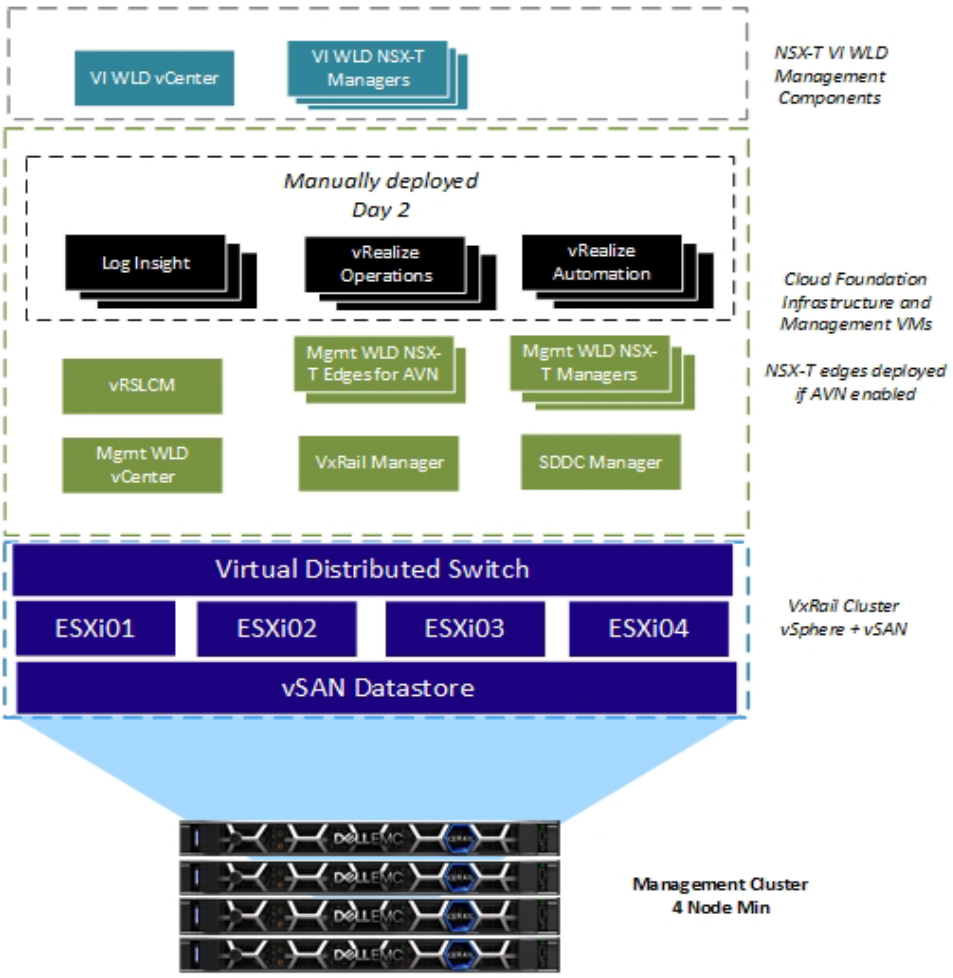

Figure 2. **Management Domain Components**

vCenter design

The management domain vCenter is deployed using the standard VxRail cluster deployment process using internal VCSA deployment. During the SDDC deployment, the vCenter is configured as an external vCenter to the VxRail Manager. This conversion is performed for two reasons:

- It establishes a common identity management system that can be linked between vCenters.
- It allows the SDDC Manager LCM process to life cycle all vCenter components in the solution.

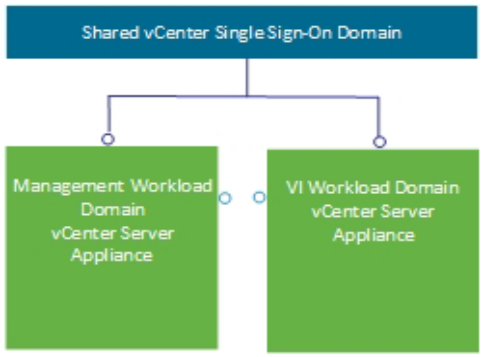

Figure 3. **vCenter Design**

VI workload domain

The VI WLD can consist of one or more VxRail clusters. The VxRail cluster is the building block for the VI WLD. The VxRail clusters in the VI WLD can start with three hosts, but four hosts are recommended to maintain FTT=1 during maintenance operations. This can be selected when adding the first VxRail cluster to the WLD. The vCenter and the NSX-T Managers for each VI WLD are deployed into the Mgmt WLD.

When the first VxRail cluster is added to the first VI WLD, the NSX-T Managers (3 in a cluster) are deployed to the Mgmt WLD. Subsequent NSX-T based VI WLDs can either use the existing NSX-T and existing NSX-T Managers, or a new NSX-T instance with three new NSX-T Managers can be deployed.

Typically, the first VxRail cluster can be considered a compute-and-edge VxRail cluster as it contains both NSX and compute components. NSX Edge nodes can be deployed to this first VxRail cluster. The second and subsequent VxRail clusters in a VI WLD can be considered compute-only clusters as they do not need to host any NSX Edge nodes.

You can dedicate a VxRail cluster just for the Edge node components if either dedicated compute or bandwidth is required for the Edge node cluster.

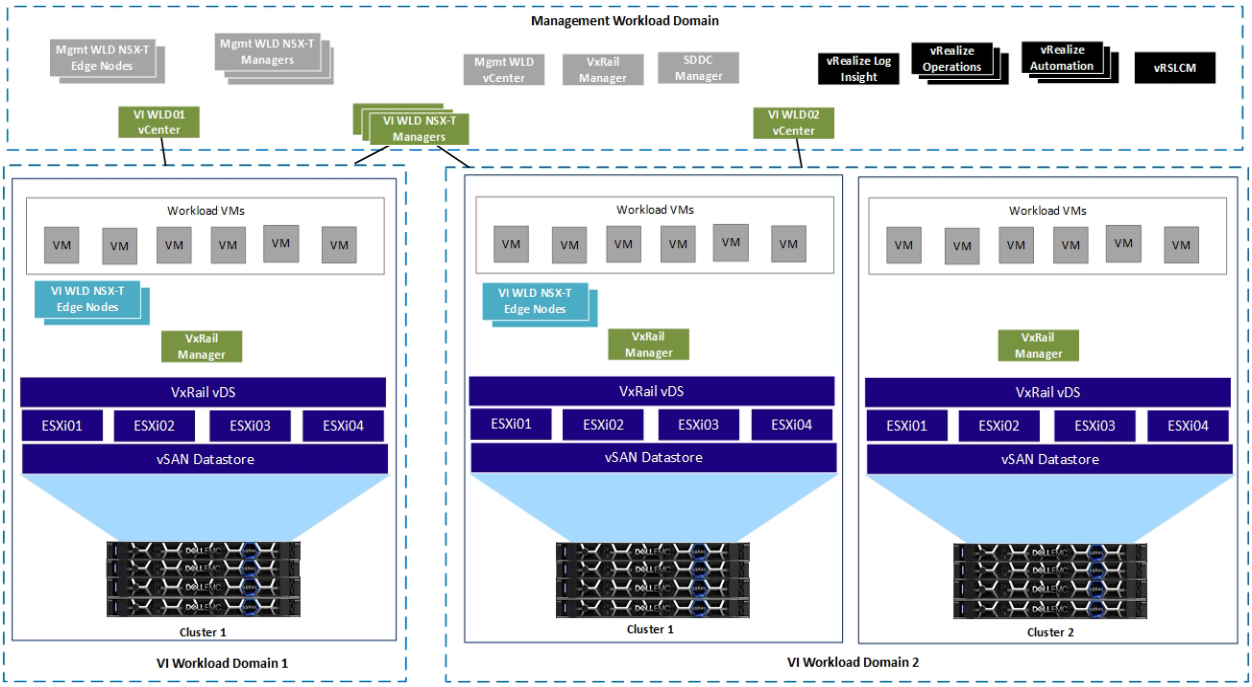

Figure 4. **VI WLD Component Layout with NSX-T everywhere**

vCenter design

The VI WLD vCenter is deployed by the SDDC Manager when creating a VI WLD. It is deployed in the Mgmt WLD as shown in the preceding figure. During deployment, it is added to the existing SSO domain, allowing a single pane of glass to manage both the management and VI WLD vCenters.

Consolidated architecture

In a standard deployment, the Cloud Foundation management WLD consists of workloads supporting the virtual infrastructure, cloud operations, cloud automation, business continuity, and security and compliance components for the SDDC. Using SDDC Manager, separate WLDs are allocated to tenant or containerized workloads. In a consolidated architecture, the Cloud Foundation management WLD runs both the management workloads and tenant workloads.

There are limitations to the consolidated architecture model that must be considered:

- The conversion of consolidated to standard requires a new VI WLD domain to be created. The tenant workloads must be migrated to the new VI WLD. The recommended method for this migration is to use HCX.
- Use cases that require a VI WLD to be configured to meet specific application requirements cannot run on a consolidated architecture. The singular management WLD cannot be tailored to support management functionality and these use cases. If your plans include applications that require a specialized VI WLD (such as Horizon VDI or PKS), plan to deploy a standard architecture.
- Life-cycle management can be applied to individual VI WLDs in a standard architecture. If the applications targeted for Cloud Foundation on VxRail have strict dependencies on the underlying platform, consolidated architecture 4 is not an option.
- Autonomous licensing can be used in a standard architecture, where licensing can be applied to individual VI WLDs. In a consolidated architecture, this is not an option.
- Scalability in a consolidated architecture has less flexibility than a standard architecture. Expansion is limited to the underlying VxRail cluster or clusters supporting the single management WLD, as all resources are shared.
- If a VxRail cluster was built using two network interfaces, consolidating VxRail traffic and NSX-T traffic, additional nodes added to a VxRail cluster are limited to two Ethernet ports being used for Cloud Foundation for VxRail.

Remote VxRail clusters

VCF 4.1 introduced remote clusters as a feature that extends a VCF WLD or a VCF VxRail cluster in order to operate at a site that is remote from the central VCF instance from which it is managed. All the Cloud Foundation operational management can be administered from the central or the regional data center out to the remote sites. Central administration and management are an important aspect because:

- It eliminates the need to have technical or administrative support personnel at the remote locations resulting in better efficiencies with much lower operating expenses.
- Edge compute processing also allows customers to comply with data locality requirements that are driven by local government regulations.
- VCF Remote VxRail clusters establish a means to standardize operations and centralize the administration and software updates to all the remote locations.

The following diagram illustrates the remote VxRail cluster feature with three different edge sites where the remote VxRail clusters are located.

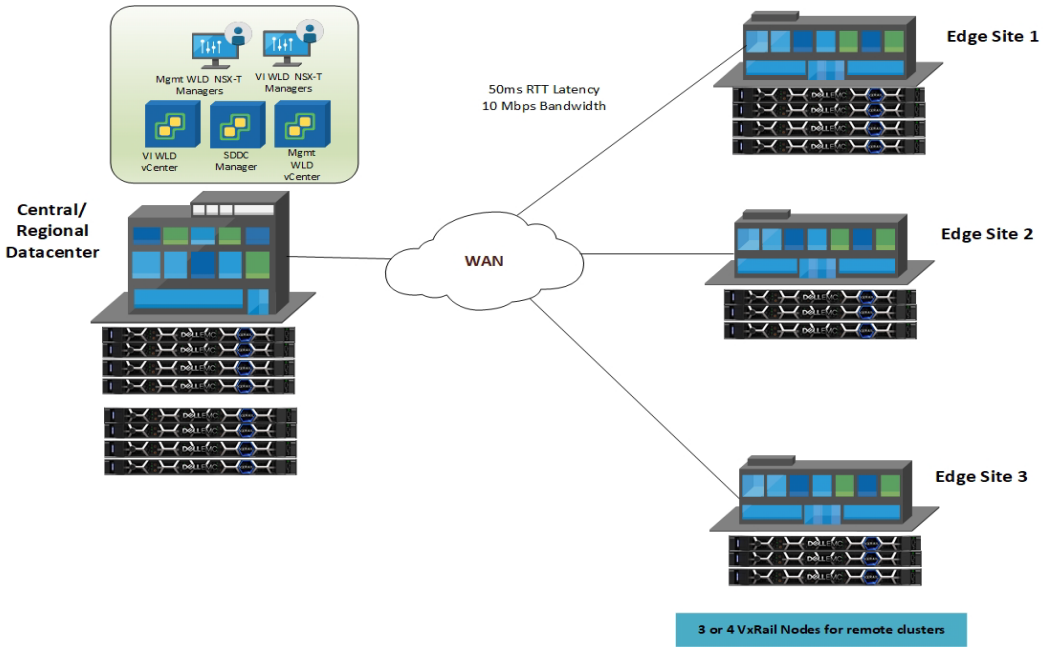

Figure 5. **Remote VxRail Cluster Deployment**

Remote VxRail Cluster Deployments

The following requirements must be met for remote VxRail cluster deployments:

- Bandwidth of 10 Mbps
- Latency is 50 millisecond RTT.
- Support only 3-4 Nodes at the Edge (ROBO) sites.
- Primary and secondary active WAN links are highly recommended.
- DNS and NTP Server is available locally or are reachable to Edge site from Central site.
- A DHCP server must be available for the NSX-T host overlay (Host TEP) VLAN of the WLD. When NSX-T creates Edge Tunnel End Points (TEPs) for the VI WLD, they are assigned IP addresses from the DHCP server. The DHCP server should be available locally at the Edge site.

Failure to adhere to these requirements will lead to system integrity, instability, resiliency, and security issues of the edge workload.

There are essentially two ways to deploy remote VxRail clusters. You can either use a dedicated WLD per site with one or more VxRail clusters per WLD or deploy VxRail clusters at the remote location in a WLD with an existing VxRail cluster in the central location. The following diagram shows a WLD deployed for each remote site with two VxRail clusters in Edge site 1 VI WLD 02 and one VxRail cluster deployed at Edge site 2 in VI WLD 03.

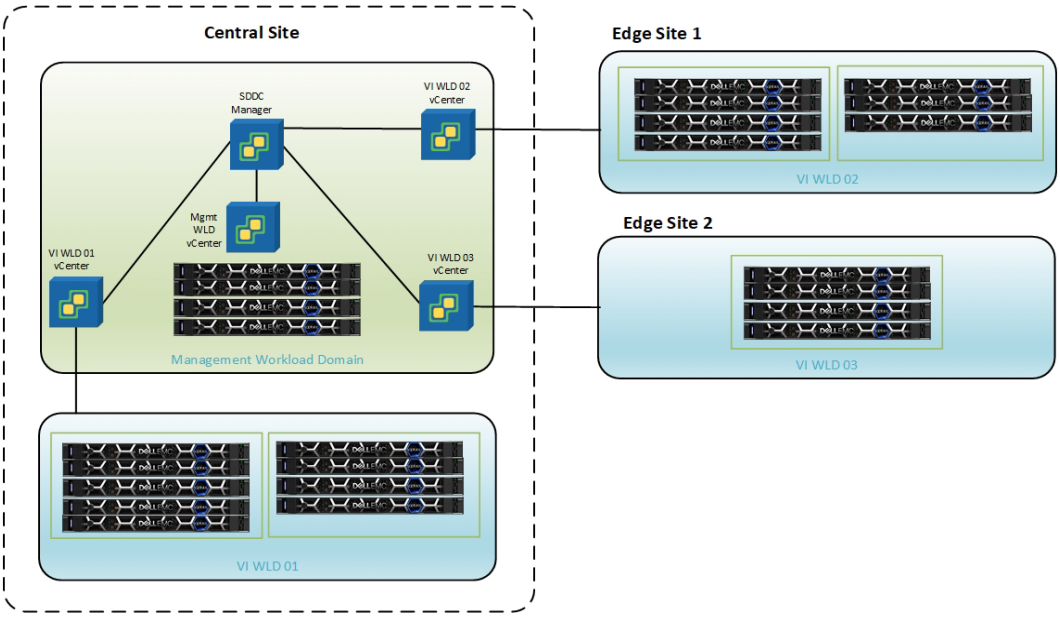

Figure 6. **Remote WLD deployment model**

The second deployment option is to deploy each site as a remote VxRail cluster in an existing VI WLD. This option reduces the number of VI WLDs and vCenters needed for the remote deployments as shown in the following diagram. In this scenario, we have an existing VI WLD 02 with a VxRail cluster from the central site and remote VxRail clusters from two different edge sites have been added to this WLD.

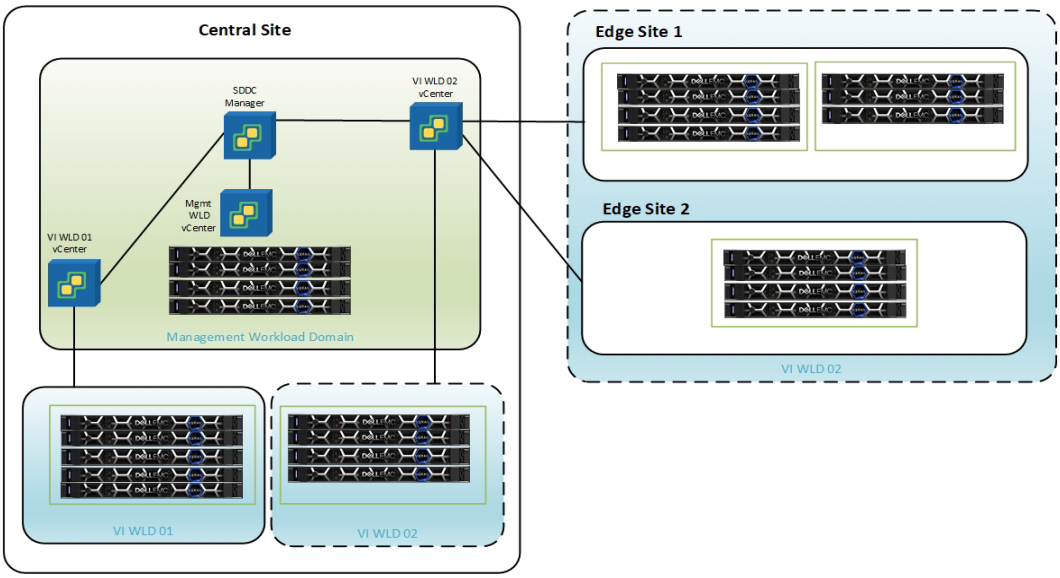

Figure 7. **Remote VxRail cluster deployment model**

Remote VxRail cluster network design

The remote sites require NSX-T edges to be deployed at each site for North/South connectivity. Also, connectivity from the central site to the remote site must be maintained to ensure connectivity of management components such as vCenter, SDDC Manager, NSX-T Manager, and so forth. As mentioned in the requirements, if DNS and NTP servers are running in the central site, they must be reachable from the Edge site.

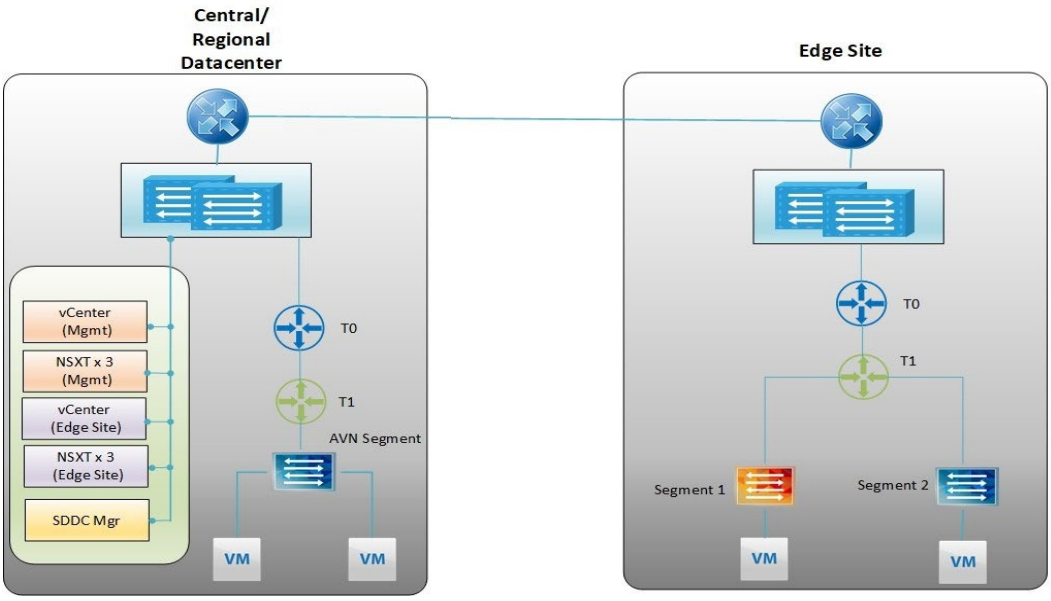

Figure 8. **Remote VxRail cluster network design**

Physical WLD layout

A WLD represents a logical boundary of functionality, managed by a single vCenter server instance. Although a WLD usually spans one rack, you can aggregate multiple WLDs in a single rack in smaller setups. In larger configurations, WLDs can span racks.

The following figure shows how one rack can be used to host two different WLDs, the Mgmt WLD and one tenant WLD. A tenant WLD can consist of one or more VxRail clusters; this will be discussed later.

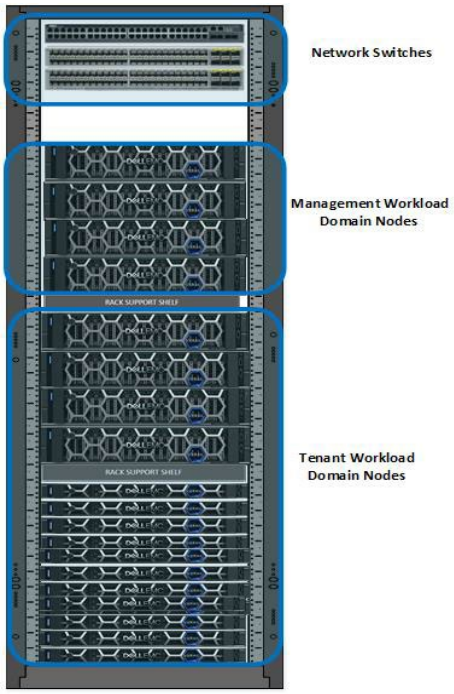

Figure 9. **Single Rack WLD Mapping**

A single WLD can stretch across multiple adjacent racks. For example, a tenant WLD that has more VxRail nodes than a single rack can support, or the need for redundancy might require stretching across multiple adjacent racks, as shown in the following figure.

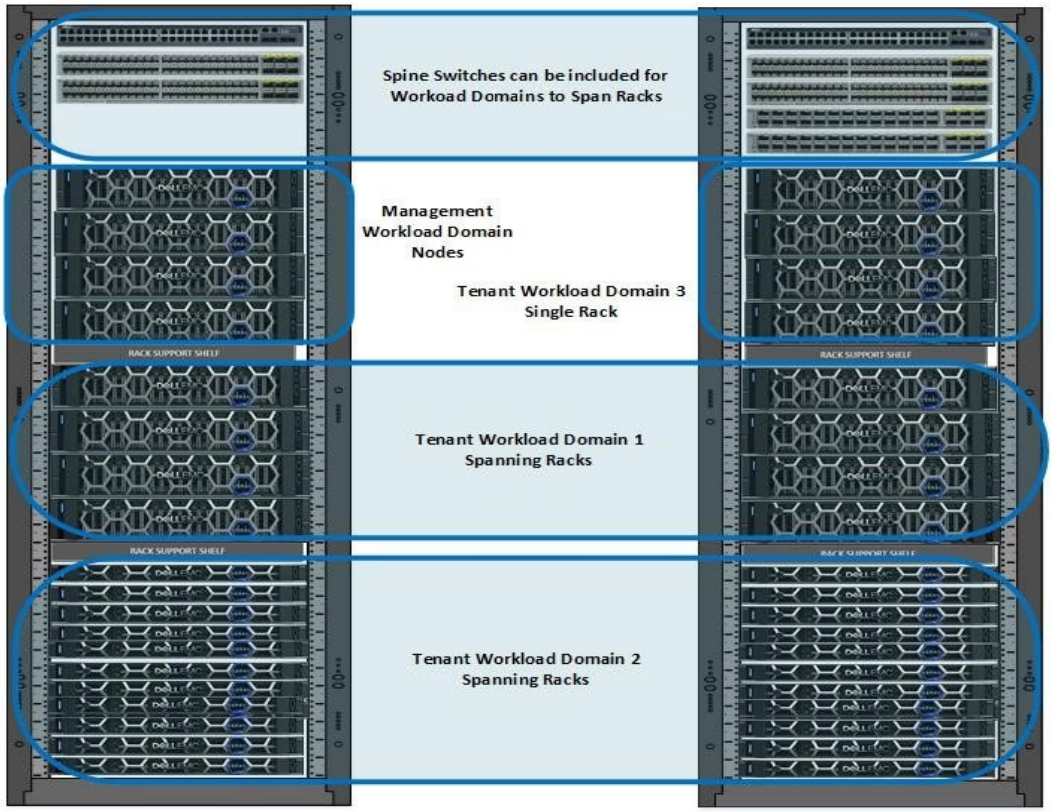

Figure 10. **WLDs Spanning Racks**

VxRail hardware options

VCF on VxRail Supports 14G/15G VxRail platforms. See the VCF on VxRail Support Matrix for detailed information.

Depending on the customer workload and application requirements, the correct VxRail hardware platform must be selected. See the VxRail sizing tool for guidance on sizing for the SDDC components for the different VxRail hardware platforms.

Chapter 4 VxRail Virtual Network Architecture

This chapter presents the following topics:

Introduction

The solution uses the network virtualization inherent in vSphere for deployment and operations of the VxRail cluster. VCF also depends on the underlying vSphere network to support a comprehensive virtualized network with its rich set of features.

VxRail virtual distributed switch (system vDS)

The VxRail Appliance is the building block for each VxRail cluster, either Mgmt WLD or VI WLD. The VxRail virtual distributed switch (vDS) also known as the system vDS provides the virtual network layer for the system network services that are needed for the VCF solution. vDS can also provide the underlying networks for NSX-based WLDs if no additional vDS will be deployed. The virtual port groups on each vDS should be separated using a dedicated VLAN for best performance and security. The VxRail cluster bring-up process requires the following VLANs:

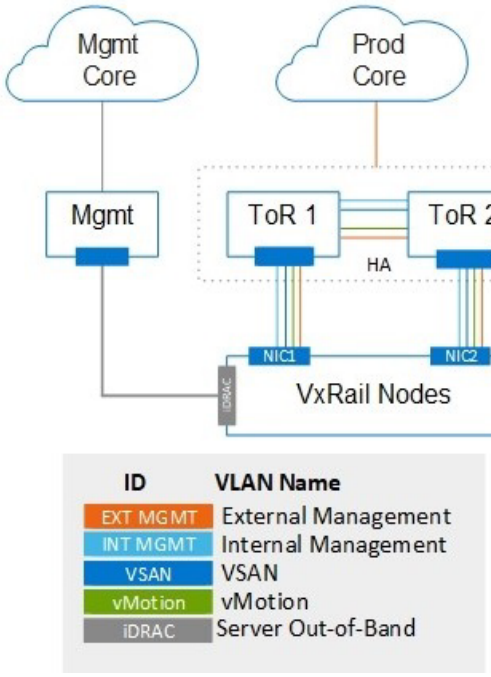

Figure 11. **VxRail Cluster VLANs**

VxRail vDS NIC teaming

There is a mixture of teaming algorithms for the port groups on the vDS. The VxRail management network that is used for node discovery uses route-based on the originating virtual port with one active and one standby adapter. This configuration is also used for the vCenter Server network where the VxRail Manager is connected. The vSAN, vMotion, and external management (vSphere) network use load-based teaming policy.

VxRail Predefined Profiles

VxRail has several predefined network profiles that can be used to deploy the VxRail in various configurations depending on the required network design and the physical networking requirements. The following tables show the teaming policies for each port group for a VxRail deployed with 2x10 or 2x25 GbE profile when using predefined network profiles.

Table 1. Predefined 2x10 or 2x25 GbE profile

Port Group	Teaming Policy	VMNIC0	VMNIC1
VxRail Management	Route based on the originating virtual port	Active	Standby
vCenter Server	Route based on the originating virtual port	Active	Standby
External Management	Route based on Physical NIC load	Active	Active
vMotion	Route based on Physical NIC load	Active	Active
vSAN	Route based on Physical NIC load	Active	Active

You can also deploy a VxRail cluster with a 4x10 network profile for either a Mgmt WLD or a VI WLD. The following table shows the teaming policy for each port group that is created with this profile.

Table 2. Predefined 4x10 profile

Port Group	Teaming Policy	VMNIC0	VMNIC1	VMNIC2	VMNIC3
VxRail Management	Route based on the originating virtual port	Active	Standby	Unused	Unused
vCenter Server	Route based on the originating virtual port	Active	Standby	Unused	Unused
External Management	Route based on Physical NIC load	Active	Active	Unused	Unused
vMotion	Route based on Physical NIC load	Unused	Unused	Active	Active
vSAN	Route based on Physical NIC load	Active	Unused	Active	Active

Finally, VxRail version 7.0.100 introduced a new 4x25 profile which is available with the following network layout.

Table 3. Predefined 4x25 profile

Port Group	Teaming Policy	VMNIC0	VMNIC1	VMNIC2	VMNIC3
VxRail Management	Route based on the originating virtual port	Active	Unused	Standby	Unused
vCenter Server	Route based on the originating virtual port	Active	Unused	Standby	Unused
External Management	Route based on Physical NIC load	Active	Unused	Active	Unused
vMotion	Route based on Physical NIC load	Unused	Active	Unused	Active
vSAN	Route based on Physical NIC load	Unused	Active	Unused	Active

VxRail vDS custom profiles

VxRail 7.0.130 introduced a new feature which allows you to create custom profiles. With this new feature, you can essentially select what uplinks/vmnics pairings to use for each type of system traffic. Custom profiles are discussed in more detail in section VxRail vDS and custom profiles.

Additional VCF NSX networks

VCF requires the following additional VLANs created and configured on the TOR switches connecting to VxRail nodes in the management WLD VxRail cluster.

Note: From VCF 4.3, the edge cluster for the management workload domain is deployed as a Day-2 operation from the SDDC Manager. It is no longer deployed by cloud builder during the initial management workload domain deployment.

Table 4. VCF VLANs for management WLD deployment

Workload Domain	Network Traffic	Sample VLAN
Management WLD	NSX-T Host TEP	103
Management WLD	NSX-T Edge TEP	104
Management WLD	Edge Uplink01	105
Management WLD	Edge Uplink02	106

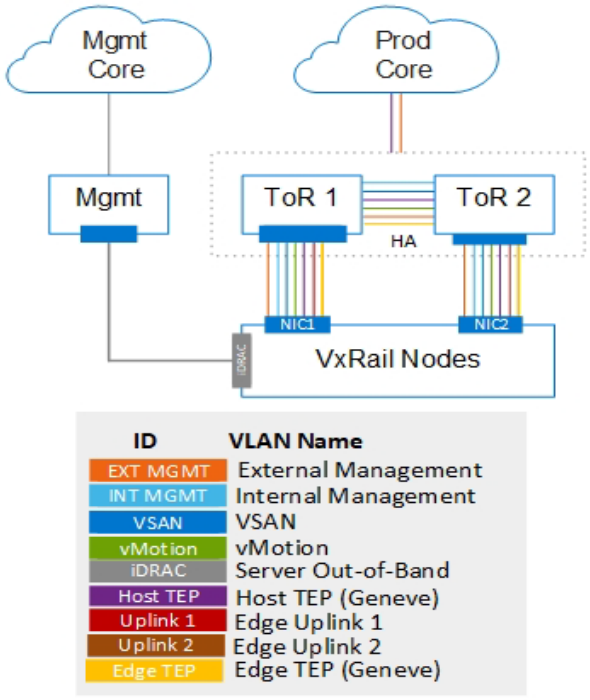

Figure 12. **VxRail Management WLD VxRail cluster VLANs (Uplink and Edge with AVN enabled)**

VCF requires the following additional VLANs created and configured on the TOR switches before deploying a VI WLD.

Table 5. VCF VLANs for VI WLD deployment

Workload Domain Type	Network Traffic	Sample VLAN
VI WLD	NSX-T Host TEP	203
VI WLD (edge deployment only)	NSX-T Edge TEP	204
VI WLD (edge deployment only)	Edge Uplink01	205
VI WLD (edge deployment only)	Edge Uplink02	206

Note: The Edge deployment is a Day 2 operation that can be achieved using either the Edge automation or a manual deployment after the VI WLD has been deployed.

VxRail vDS with NSX networks

VxRail vDS and predefined network profiles

If a single vDS is used for the deployment, all system traffic and NSX traffic share the same vDS. There are four predefined VxRail vDS network profiles for the deployment, two uplinks with 2x10 GbE, two uplinks with 2x 25 GbE, four uplinks with 4x10, and four uplinks with 4x25 profiles. A two-uplink profile can either be 2x10 or 2x25. The following two diagrams illustrate the connectivity and teaming configuration for the VxRail vDS used for system traffic and NSX traffic with the 2x10/2x25 network profile and the 4x10 network profile.

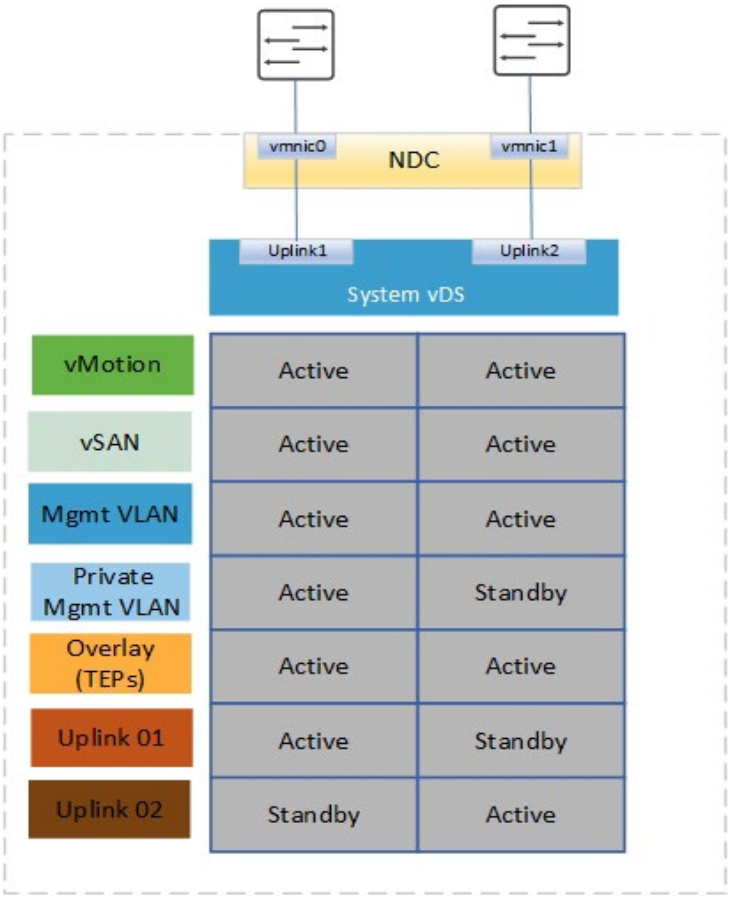

Figure 13. **Single vDS using 2x10/2x25 predefined network profile**

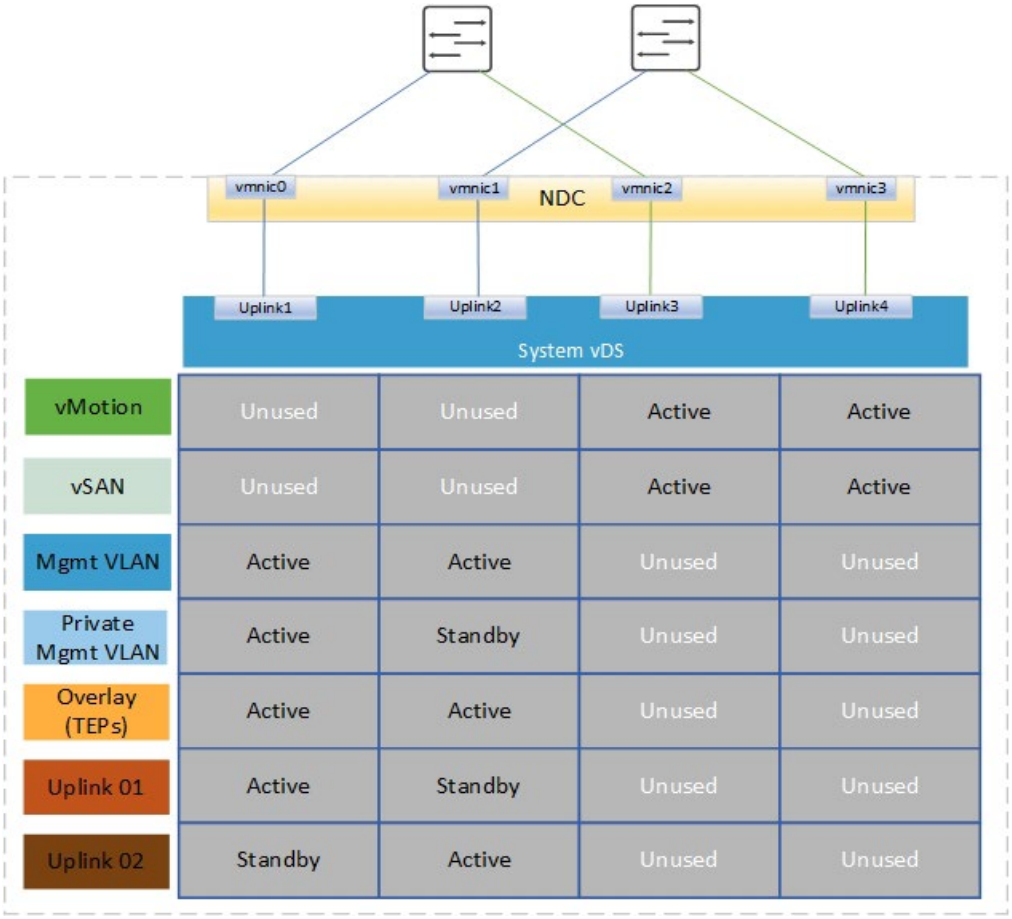

Figure 14. **Single vDS with 4x10 predefined network profile**

The next diagram illustrates the 4x25 profile that was introduced in VxRail version 7.0.100. It uses both an NDC/OCP and PCIe to achieve NIC level redundancy for the system traffic.

From 7.0.130 onwards, using this profile is not recommended as it results in a nonstandard wiring configuration as shown in the following figure.

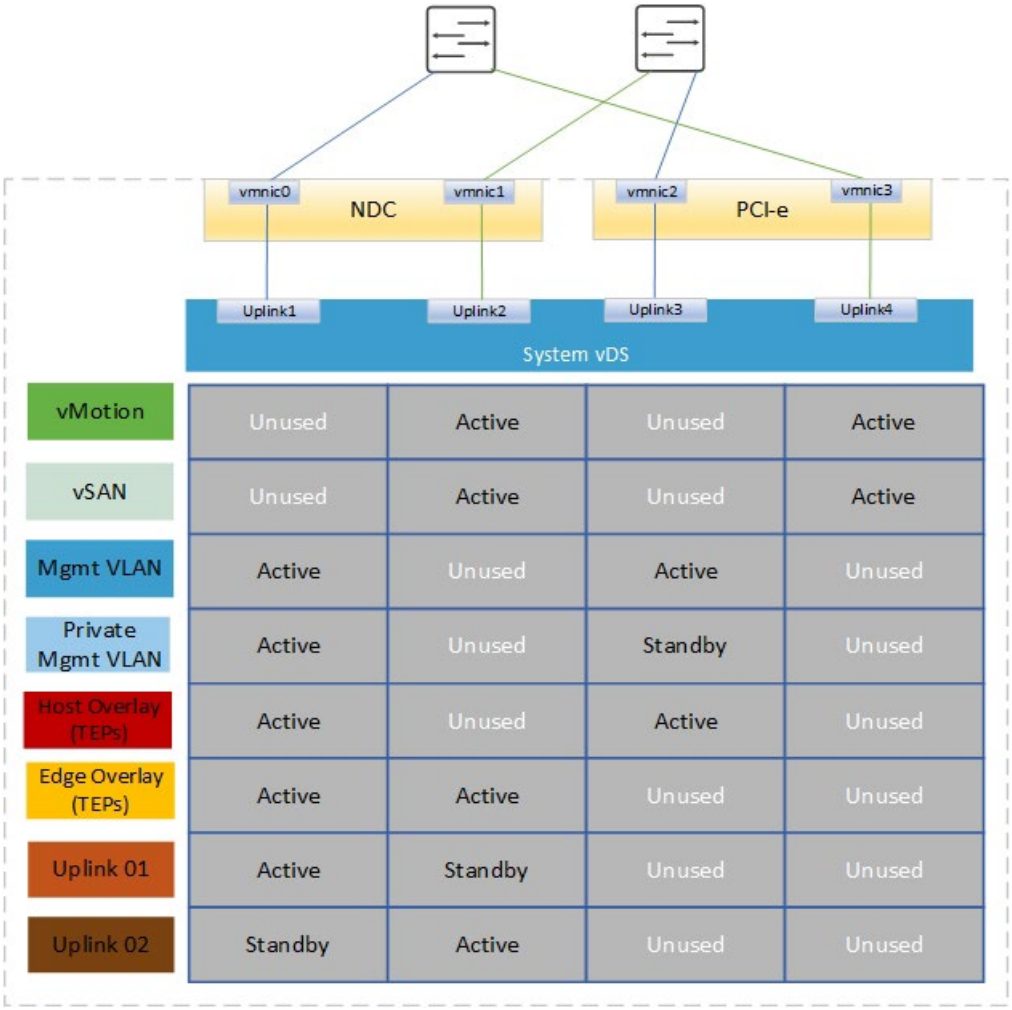

Figure 15. **Single vDS with 4x25 predefined network profile**

VxRail vDS and custom profiles

Notice that in the previous diagram, the cabling of the NIC ports to the switches is a little unorthodox. We normally have vmnic2 going to Fabric A and vminc3 going to Fabric B.

Note: Starting with VCF version 4.3, you can select the vmnics used for external Mgmt, vSAN or vMotion, to be used for the host TEP traffic and Edge traffic.

Starting with VCF 4.2 and VxRail 7.0.131, the recommended method to achieve NIC level redundancy for a VCF on VxRail cluster with 4x25 GbE is to configure the custom profile for the VxRail vDS using the configuration of the vmnic/uplink mappings and the uplink to port group mapping as shown in the following two tables. This must be done when creating the json before the VxRail cluster deployment for any VCF clusters.

Table 6. VxRail vDS uplink to pNIC mapping

vDS Uplink	Physical NIC
Uplink1	vmnic0 – NDC - port 1
Uplink2	vmnic3 – PCIe - port 2
Uplink3	vmnic1 – NDC - port 2
Uplink4	vmnic2 – PCIe - port 1

Table 7. VxRail vDS port group uplink mapping

Port Group	Teaming Policy	Active	Standby
VxRail Management	Route based on the originating virtual port	Uplink1	Uplink2
vCenter Server	Route based on the originating virtual port	Uplink1	Uplink2
External Management	Route based on Physical NIC load	Uplink1	Uplink2
vMotion	Route based on Physical NIC load	Uplink3	Uplink4
vSAN	Route based on Physical NIC load	Uplink3	Uplink4

Note: During VCF deployment of management WLD or when a VxRail cluster is ingested into a VI WLD, all port groups are configured as active/active except VxRail management, which remain active/standby.

The configuration of the vDS and uplink to pNIC mapping that is shown in the following diagram provides NIC level redundancy for a 4x25 GbE deployment when using the VxRail custom profile feature in 7.0.131.

Note: The uplink to vmnic mappings on the vDS as a misconfiguration could cause a deployment failure.

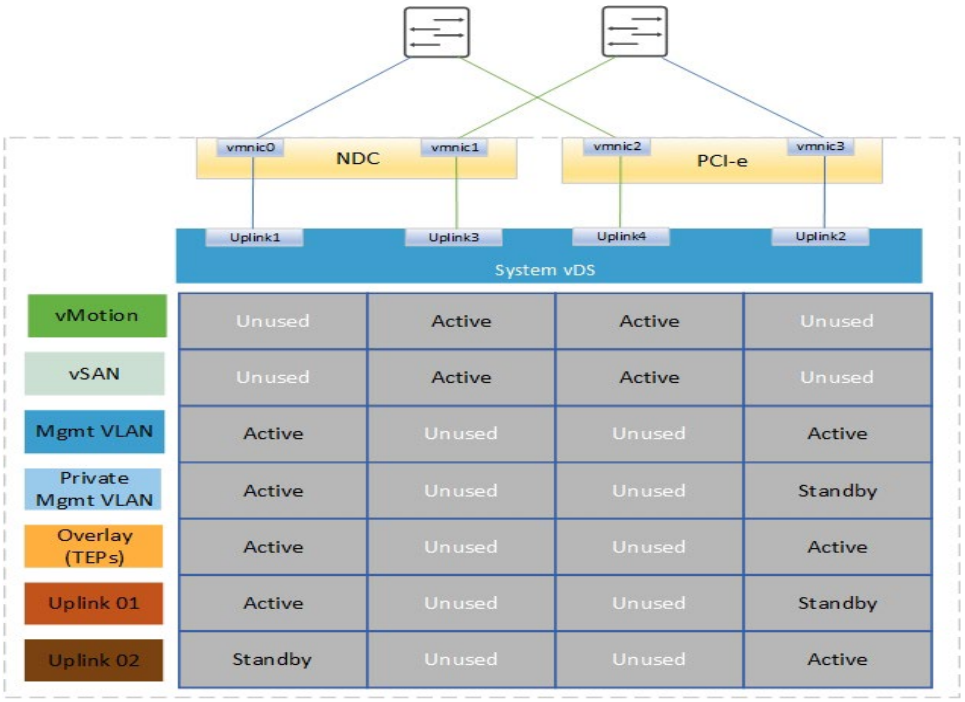

Figure 16. **Single vDS using custom profile for vSAN traffic isolation**

Note: The preceding design can also be achieved using 10 GbE network cards when a custom profile is used to create the configuration.

Another variation of the preceding design might be to separate vSAN onto a dedicated pair of physical NICs, or a different pair of TOR switches to ensure that maximum bandwidth can always be allocated to vSAN traffic. This design would require one change in the custom profile where vMotion would use uplink1/uplink2 leaving vSAN only using uplink3/uplink4.

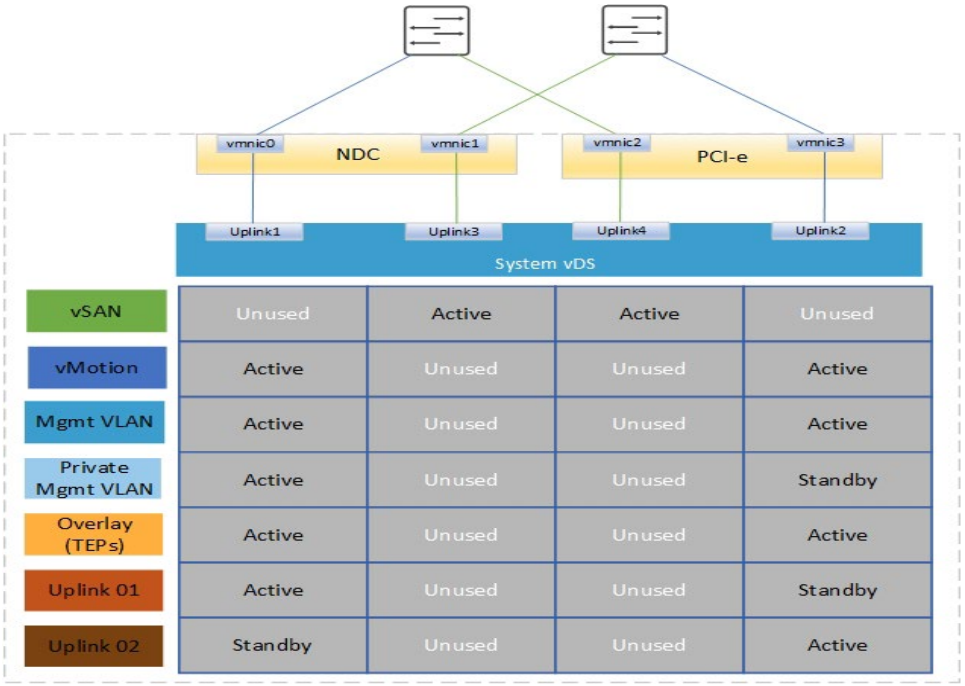

Figure 17. **Two vDS with custom profile and NIC-level redundancy**

NSX vDS

Starting with VCF 4.0.1, you can fully separate system traffic and NSX-T traffic using a second vDS dedicated for NSX traffic. This can be achieved for both the management WLD and VI WLD. For the management WLD, additional optional inputs are required. Cloud builder creates and configures the second vDS during VCF bring-up. For the VI WLD, a script available in Developer Center in the SDDC Manager must be used to add a VxRail cluster with a second vDS to the VI WLD. The additional input that is needed for the vDS must be provided to the script during execution.

Note: VCF orders the vmnic to uplink mapping on the vDS lexicographically (lowest to highest), even if the order in the input spreadsheet for the Mgmt WLD or the input to the script in the VI WLD are not ordered lexicographically.

Two vDS (system and NSX) network topologies

The second vDS provides several different network topologies. Some of these topologies are covered in this section. Note in these examples, we focus on the connectivity from the vDS and do not take the NIC card to vDS uplink into consideration. With the new feature for custom profiles, there are too many combinations to cover in this guide.

Two vDS (system and NSX) – 4 pNIC topology

The first option uses four pNICs, two uplinks on the VxRail (system) vDS and two uplinks on the NSX vDS.

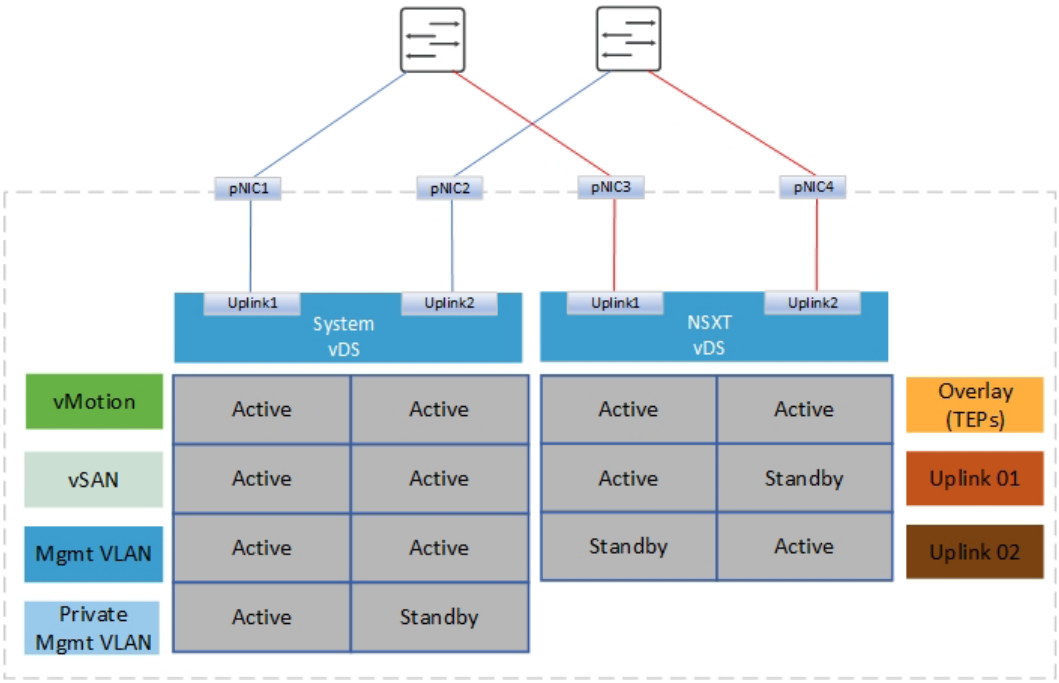

Figure 18. **Two vDS with four pNICs**

If we now consider the two-vDS design and NIC level redundancy, the VxRail vDS must be deployed using a custom profile using uplink1/uplink2 for all traffic with uplink1 mapped to a port

on the NDC and the second uplink2 mapped to a port on the PCIe providing NIC level redundancy for the system traffic. When the VxRail cluster is added to VCF, the remaining two pNICs (one from NDC and one from PCIe) can be selected to provide NIC level redundancy for the NSX traffic. The next diagram illustrates this network design.

Note: Both ports from NDC must connect to switch A, and both ports from the PCIe must connect to switch B. This is a requirement for VCF vmnic lexicographic ordering.

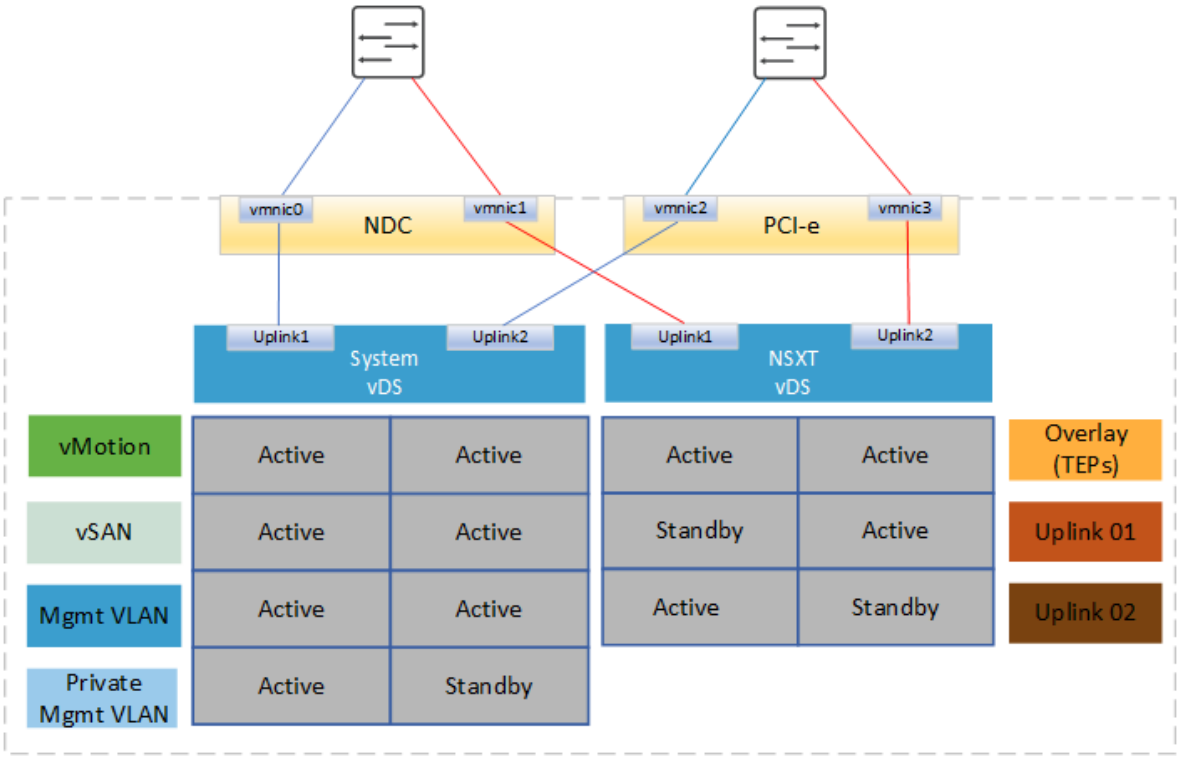

Figure 19. **Two vDS with custom profile and NIC level redundancy**

Two vDS (system and NSX) - 6 pNIC topologies

There are two options in a six-pNIC design with two vDS. For the first option, we have four pNICs on the VxRail vDS and use two additional pNICs dedicated for NSX traffic on the NSX vDS. This might be required to keep Mgmt and vSAN/vMotion on different physical interfaces and also if NSX-T needs its own dedicated interfaces.

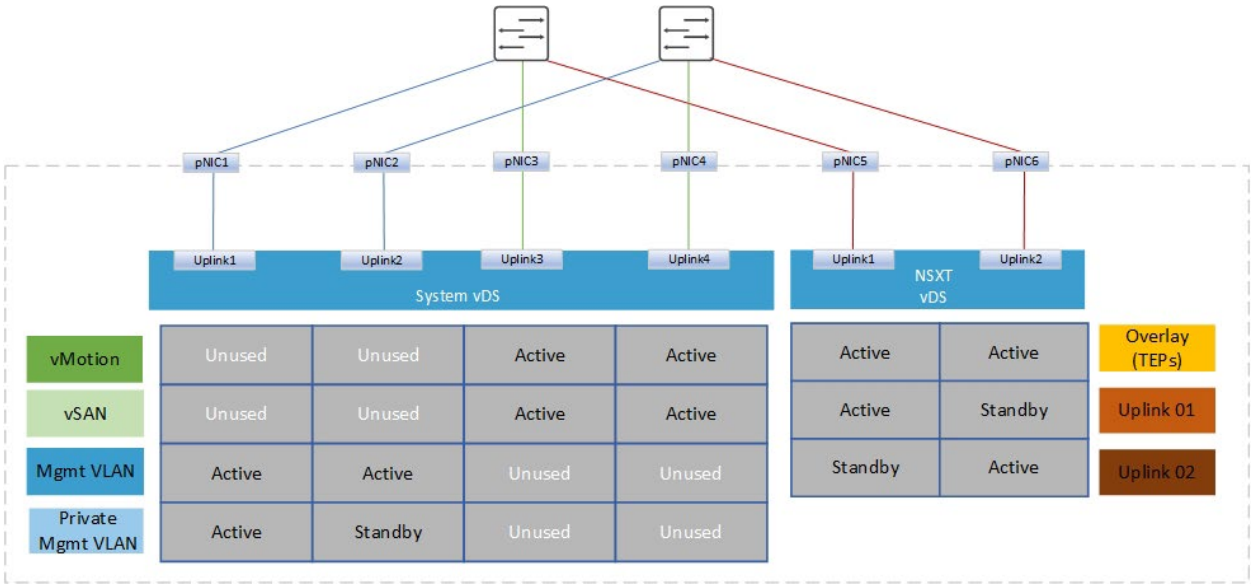

Figure 20.**Two vDS with six pNICs option 1**

The second option with six pNICs uses a system vDS with two pNIC and the NSX vDS with four pNICs. This increases the bandwidth for NSX East/West traffic between transport nodes. The use case for this design might be when the East/West bandwidth requirement scales beyond two pNICs. The host overlay traffic uses all four uplinks on the NSX vDS, load-balanced using source ID teaming. The Edge VMs by default use uplink 1 and 2 on NSX vDS. This includes the Edge overlay and Edge uplink traffic.

Note: Starting with VCF 4.3, you can select what uplinks to use on the vDS for the Edge VM traffic when a dedicated NSXT vDS has been deployed with four uplinks.

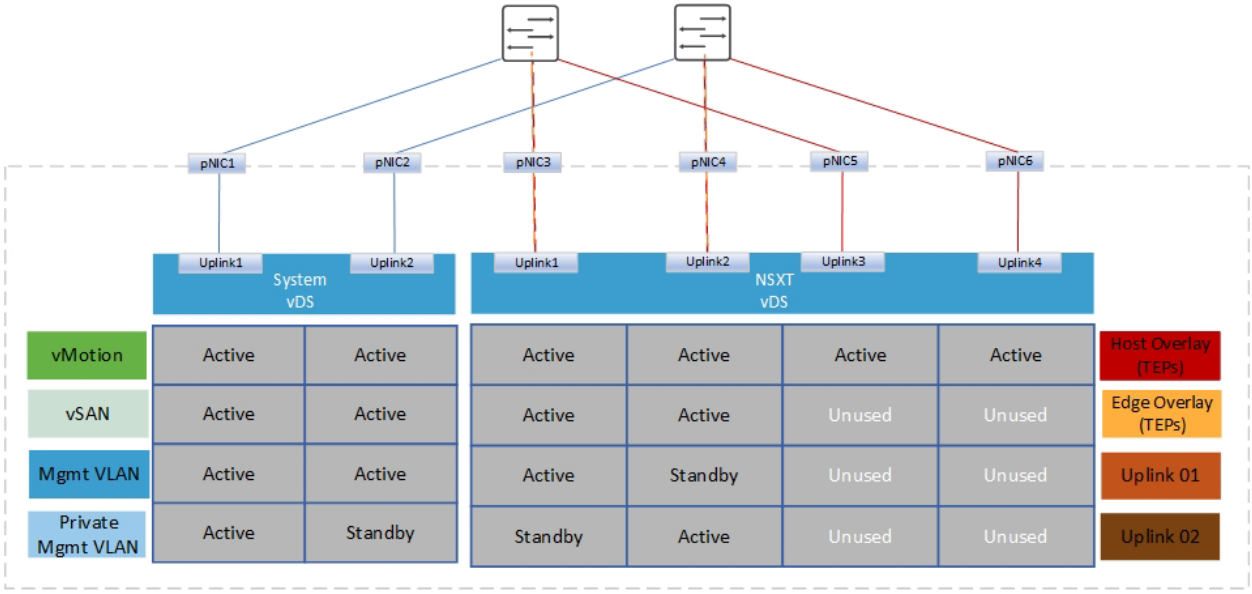

Figure 21. **Two vDS with six pNICs option 2**

Two vDS (System and NSX) – 8 pNIC topologies

The eight-pNIC option that is illustrated in the following diagram provides a high level of network isolation and also the maximum bandwidth for NSX East/West between host transport nodes. At the cost of a large port count on the switches, each host requires four ports per switch. The VxRail vDS (system) uses four uplinks and the NSX vDS also uses four uplinks.

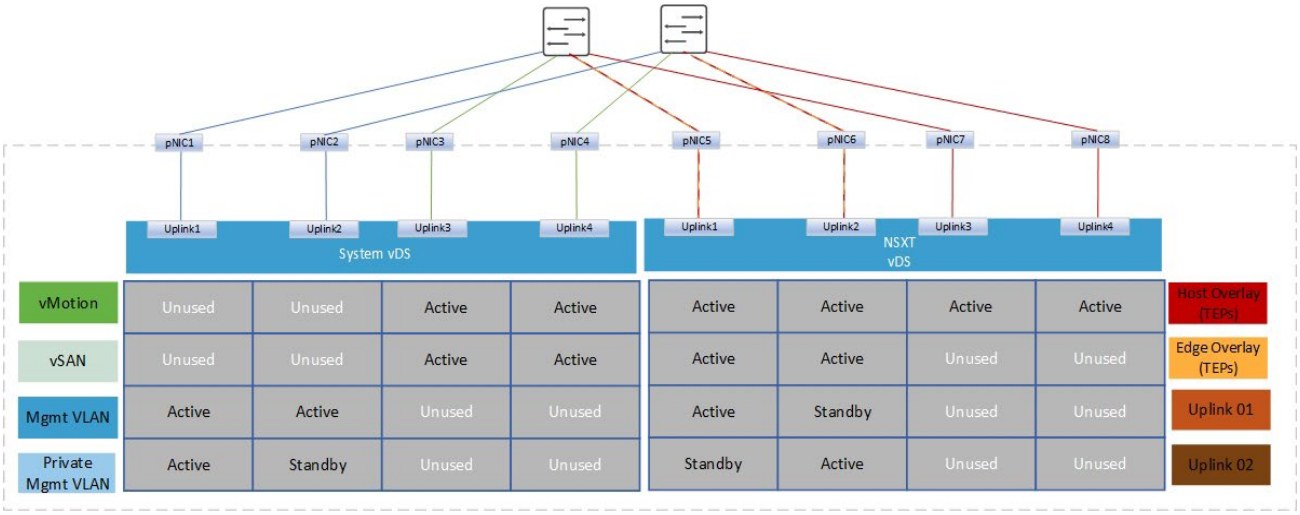

Figure 22. **Two vDS (System and NSX) design with eight pNICs**

Two System vDS

Starting with VCF 4.3.0, you can have two VxRail system vDS to separate system traffic onto two different vDS. For example, vMotion and external management traffic can be on one vDS and vSAN on another vDS. Either one of the two VxRail vDS can be used for NSXT traffic. Alternatively, you can use a dedicated NSXT vDS which results in three vDS in the network design. In some cases, physical separation is needed for management/vMotion, vSAN, and NSX-T traffic. The three-vDS design provides this capability. Sample topologies are described in this section.

Two system vDS – 4 pNIC

In this first example, two system vDS are used. The first vDS is used for management and NSX-T traffic, the second system vDS for vMotion and vSAN traffic. This design also incorporates NIC-level redundancy.

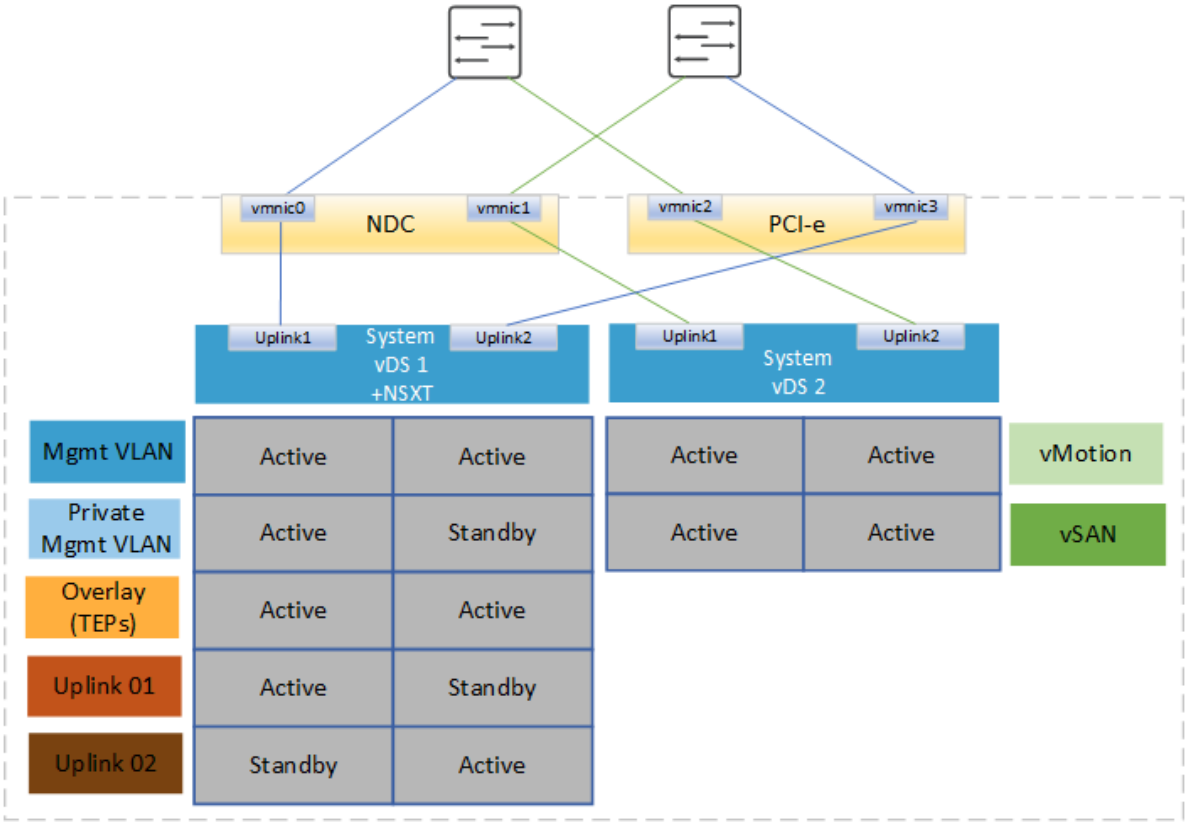

Figure 23. **Two system vDS design with four pNIC**

In the next example, vSAN is isolated to a dedicated network fabric. This might be needed if there is a requirement for physical separation of storage traffic from management and workload production traffic.

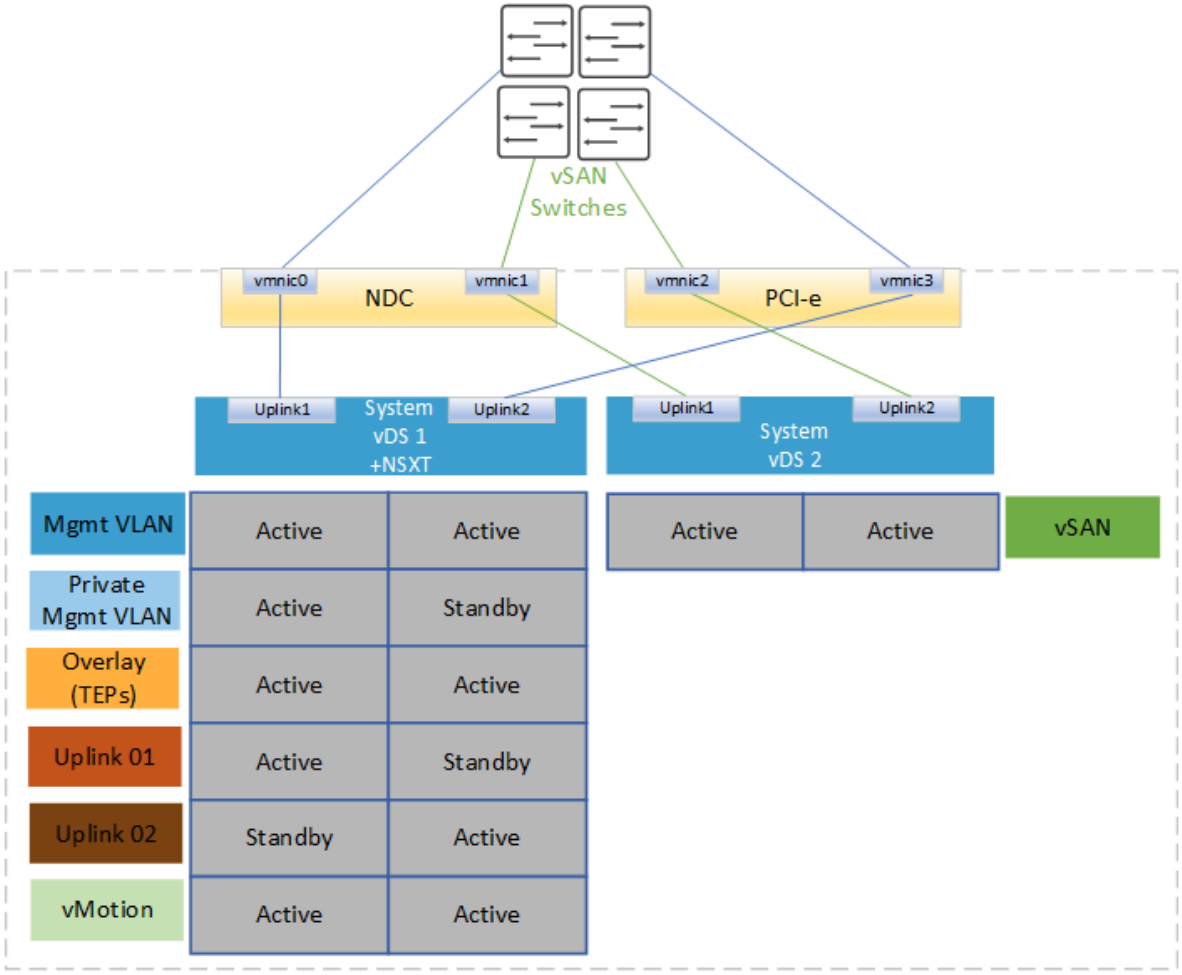

Figure 24. **Two system vDS design with 4 pNIC and isolated vSAN**

Two System vDS –six pNIC

Another option is to deploy the first system vDS with four uplinks. This option allows separation of vMotion from management and workload production traffic and allows vSAN on its own dedicated vDS. This option requires six pNICs per host. With VCF 4.3, you can pin the NSX-T traffic to the same NICs using either management, vMotion, or vSAN on either vDS. In this example, management NICs have been selected on system vDS1 which is the default behavior in the previous release.

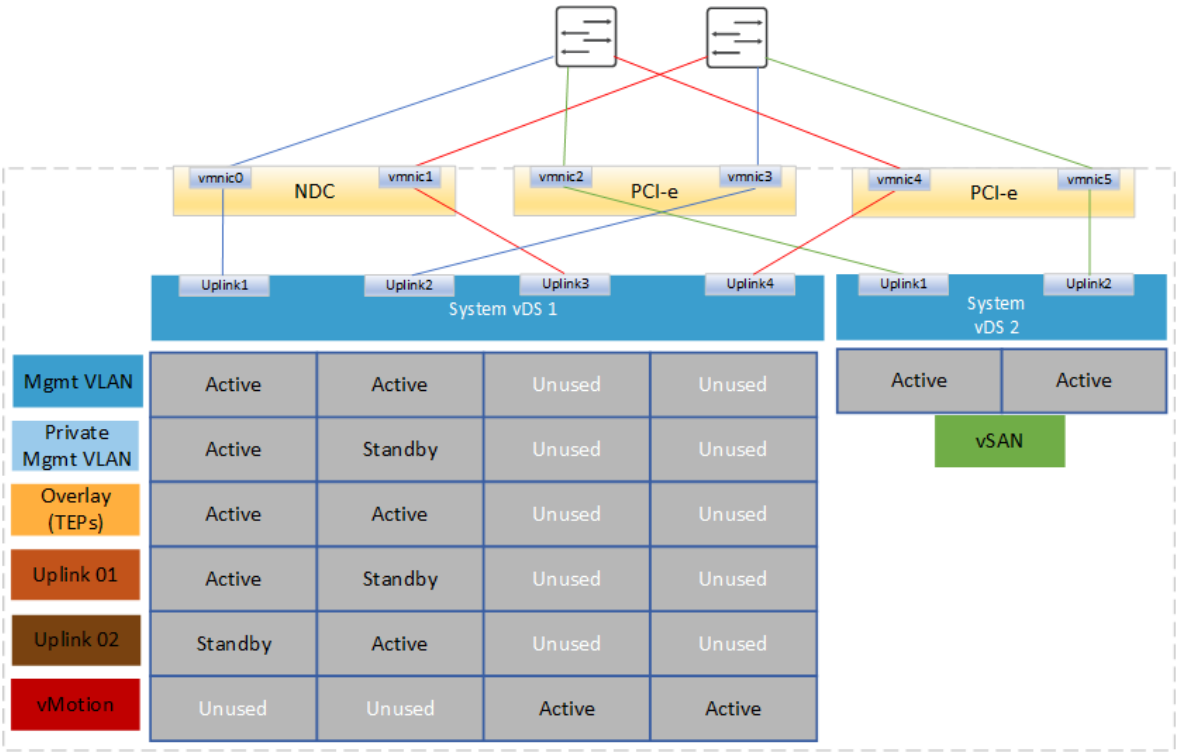

Figure 25. **Two-system vDS design with six pNIC**

Note: VCF 4.3 provides the option to place NSX-T traffic onto the same two pNICs used for external management, vMotion or vSAN on either system vDS.

A third vDS is required if there is a requirement to isolate NSX-T traffic and vSAN traffic. In this case, two system vDS and an NSX-T vDS are required. The first system vDS is used for management and vMotion, the second system vDS is used for vSAN, and the third vDS is dedicated for NSX-T traffic. This results in six pNICs per host in this network design.

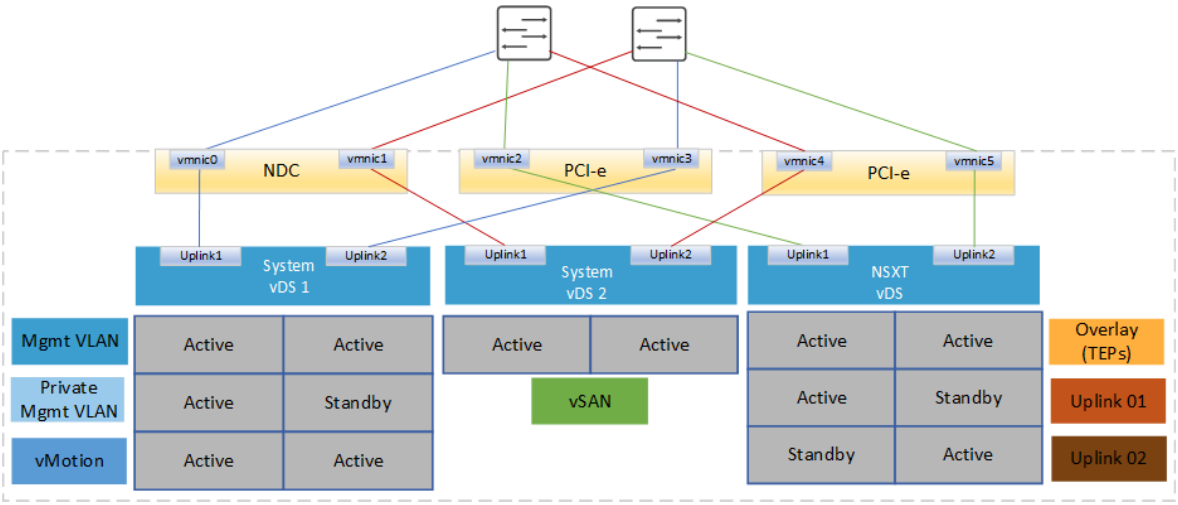

Figure 26. **Two system vDS and NSX-T vDS**

Chapter 5 Network Virtualization

This chapter presents the following topics:

Introduction

The foundation of the Network virtualization layer for VCF on VxRail is provided by NSX-T. The solution provides a software-defined networking approach that delivers Layer 2 to Layer 7 networking services (for example, switching, routing, firewalling, and load balancing) in software. These services can then be programmatically assembled in any arbitrary combination, producing unique, isolated virtual networks in a matter of seconds. NSX-T is considered the next generation and provides additional features that NSX-V did not provide. For multicloud connectivity and security, NSX-T provides the best possible features and it provides native support for Kubernetes, PKS, and Cloud Native applications.

NSX-T architecture

NSX-T reproduces the complete set of networking services (such as switching, routing, firewalling, QoS) all in a network virtualization layer which is an abstraction between the physical and virtual networks. The NSX-T platform consists of several components that operate across three different planes: management, control, and data.

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

Management plane

The management plane provides a single API entry point to the system. It maintains user configuration, handles user queries, and performs operational tasks on all management, control, and data plane nodes. It provides an aggregated system view and is the centralized network management component of NSX-T. NSX-T Manager is delivered in a virtual machine form factor and is clustered with three VMs to provide High Availability of the Management plane.

Note: Bare Metal NSX-T servers and edges are not supported.

Control plane

The control plane computes the runtime state of the system based on configuration from the management plane. It also disseminates topology information that is reported by the data plane elements and pushes stateless configuration to forwarding engines. It runs on VLAN-backed networks that are isolated from the transport networks for the data plane. NSX-T splits the control plane into two parts:

- Central Control Plane (CCP)–The CCP is implemented on the NSX-T cluster of managers. The cluster form factor provides both redundancy and scalability of resources. The CCP is logically separated from all data plane traffic, meaning any failure in the control plane does not affect existing data plane operations.
- Local Control Plane (LCP)–The LCP runs on transport nodes. It is next to the data plane it controls and is connected to the CCP. The LCP programs the forwarding entries of the data plane.

Data plane The data plane performs stateless forwarding or transformation of packets, based on tables that are populated by the control plane. It reports topology information to the control plane and maintains packet level statistics.

The transport nodes are the hosts running the local control plane daemons and forwarding engines implementing the NSX-T data plane. The N-VDS is responsible for switching packets according to the configuration of available network services.

NSX-T network services

NSX-T provides all the Layer 2 to Layer 7 services that are required to build virtualized networks in the software layer for modern user applications. The following sections describe these different services, and the functions they provide.

Segments (logical switch) The segment, previously known as logical switch, is a Layer 2 construct similar to a VLAN backed network except that it is decoupled from the physical network infrastructure. Segments can be created in a VLAN transport zone or an overlay transport zone. Segments that are created in an overlay transport zone have a Virtual Network Identifier (VNI) associated with the segment. VNIs can scale far beyond the limits of VLAN IDs.

Gateway (logical router) A logical router, also known as a gateway, consists of two components: distributed router (DR) and services router (SR).

A DR is essentially a router with logical interfaces (LIFs) connected to multiple subnets. It runs as a kernel module and is distributed in hypervisors across all transport nodes, including Edge nodes. The DR provides East/West routing capabilities for the NSX domain.

An SR, also referred to as a services component, is instantiated when a service is enabled that cannot be distributed on a logical router. These services include connectivity to the external physical network or North/South routing, stateful NAT, Edge firewall.

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.

Transport zones Transport zones define the span of a virtual network (segment) across hosts or clusters. Transport zones dictate which ESXi hosts and which virtual machines can participate in the use of a particular network.

Transport node Each hypervisor that is prepared for NSX-T and has an NDVS component installed is an NSX-T transport node that is equipped with a tunnel endpoint (TEP). The TEPs are configured with IP addresses, and the physical network infrastructure provides IP connectivity either over Layer 2 or Layer 3. An NSX-T Edge node can also be a transport node that is used to provide routing services. When an Edge node or ESXi host contains an N-DVS component, it is considered a transport node.

NSX-T Edge node Edge nodes are service appliances with pools of capacity, dedicated to running network services that cannot be distributed to the hypervisors. Edge nodes can be viewed as empty containers

when they are first deployed. Centralized services like North/South routing or Stateful NAT which require the SR component of logical routers to run on the Edge node. The Edge node is also a transport node just like compute nodes in NSX-T. Similar to a compute node, it can connect to more than one transport zone. The Edge node typically connects to one for overlay and other for North/South peering with external devices.

NSX-T Edge cluster

An Edge cluster is a group of Edge transport nodes that provides scale out, redundant, and high-throughput gateway functionality for logical networks. An NSX-T Edge cluster does not have a one-to-one relationship with a VxRail cluster. NSX-T Edge clusters can be distributed across multiple VxRail clusters.

Distributed firewall

The NSX-T firewall is delivered as part of a distributed platform that offers ubiquitous enforcement, scalability, line rate performance, multi-hypervisor support, and API-driven orchestration. NSX-T distributed firewall provides stateful protection of the workload at the vNIC level. DFW enforcement occurs in the hypervisor kernel, helping to deliver micro-segmentation. Uniform security policy model for on-premises and cloud deployment support multi-hypervisor (that is, ESXi and KVM) and multi-workload, with a level of granularity down to VM and container attributes.

Chapter 6 NSX-T WLD Design

This chapter presents the following topics:

Introduction

Starting with VCF version 4.0, NSX-T is ubiquitous throughout the SDDC solution. Both management and VI WLDs are based on NSX-T. The design for the Mgmt WLD and the VI WLD are similar. We will go through each design to ensure that any differences are identified.

Application Virtual Network (AVN)

The deployment of the Mgmt WLD includes installation of NSX components. It lays down the Application Virtual Network (AVN) for the vRealize Suite. It deploys the necessary NSX-T Edges and configures T0/T1 routers and configures dynamic routing to allow traffic from the AVNs to the external networks. The following sections describe the various components in the NSX-T design and where each component is used for AVN.

NSX-T transport zone design

A transport zone defines the span of the virtual network, as logical switches only extend to N-VDS on the transport nodes that are attached to the transport zone. Each ESXi host has an N-VDS component for the hosts to communicate or participate in a network. They must be joined to the transport zone. There are two types of transport zones:

- Overlay – Used for all Overlay traffic for the Host TEP communication
- VLAN – Used for VLAN backed segments. This includes the Edge VM communications.
- For the Mgmt WLD, only one VxRail cluster exists in standard architecture. All nodes in the VxRail cluster are added to an Overlay network. This network is used for AVN if the feature is enabled. VLAN backed transport zone can be used for any VLAN-backed segments that are created in NSX-T.

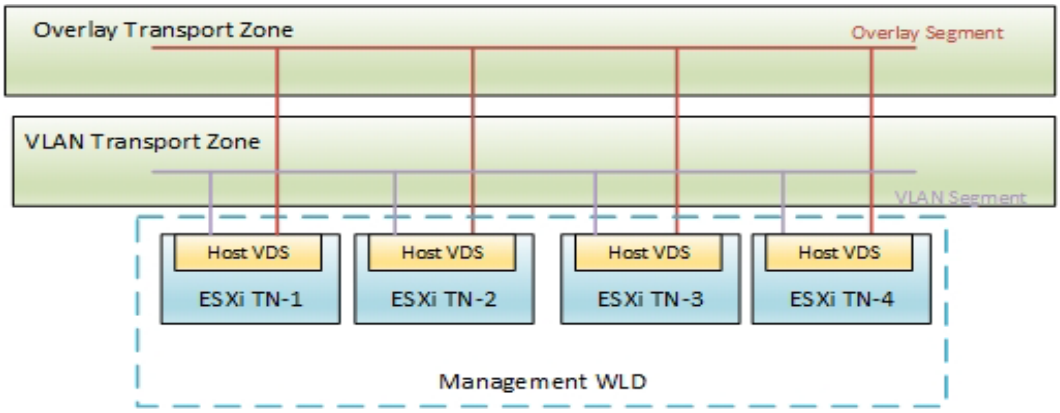

Figure 27. **Mgmt WLD Transport Zones**

For VI WLD transport zones, when the first VxRail cluster is added to the first VI WLD, SDDC Manager creates the Overlay and VLAN transport zones in the VI WLD. These transport zones are then used for that WLD or even another VI WLD if the same NSX-T instance is used. This is known as 1:Many or one NSX-T instance for multiple VI WLD. However, with VCF 4.0 it is now possible to create a new NSX-T instance for each VI WLD. This is known as the 1:1 NSX-T feature, one NSX-T instance for each VI WLD.

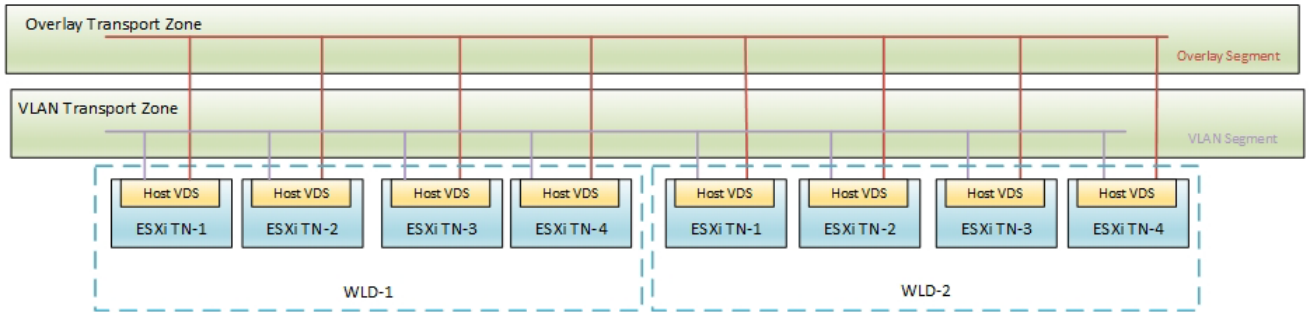

Figure 28. **VI WLD 1:Many NSX-T Transport Zones**

Note: When subsequent VxRail clusters are added to a WLD, or if a new WLD is created, all the nodes participate in the same Overlay Transport Zones. For each VxRail cluster, the same VLAN or a different VLAN can be used for the Host TEP traffic for the Overlay. We recommend using different VLANs as the size of VxRail clusters and the number .

NSX-T segments

Segments are used to connect VMs to Layer 2 networks, and they can be either VLAN or Overlay segments. For the Mgmt WLD, when the NSX-T edges and AVN networks are deployed from SDDC Manager, two segments are created for the AVN networks, Region A and xRegion, to be used for the vRealize Suite of components. Two segments are created to allow Edge node traffic to flow through the ESXi hosts to the physical network. Following is the complete list of segments that are created to support AVN in the virtual infrastructure of the SDDC.

Table 8. NSX-T Segments for Mgmt WLD

Segment	Transport Zone	VLAN (example)
Region A Segment (AVN Network)	Overlay	None
xRegion Segment (AVN Network)	Overlay	None
Edge-uplink01	VLAN (edge uplink TZ)	105
Edge-uplink02	VLAN (edge uplink TZ)	106

For the VI WLD, the edge automation is used to deploy the NSX-T Edges from SDDC manager; the edge uplink segments are created at this time. Segments for workloads are created as a Day-2 activity outside of SDDC Manager.

Uplink profile design

The uplink profile is a template that defines how an N-VDS or a vDS that exists in each transport node (either host or Edge VM) connects to the physical network. Starting with VCF version 4.0, vDS is used to back the host Overlay and VLAN transport zones for both the Mgmt and VI WLDs. The uplink profile specifies:

- Uplinks to be used by the transport node
- Teaming policy that is applied to those uplinks

- VLAN used for the profile
- MTU applied to the traffic

The following table shows the different uplink profiles that are used for the VCF on VxRail SDDC solution. This can be either the single VxRail vDS or the second dedicated NSX vDS when using only two uplinks.

Table 9. VxRail or NSX vDS with 2 uplinks - Mgmt and VI WLD Uplink Profiles

WLD Type	Profile	Default Teaming Policy	Active Uplinks	Transport VLAN (example)	Recommended MTU
Mgmt WLD	Host Overlay profile (deployed by Cloud Builder)	Load Balance Source	uplink-1,uplink-2	103	9000
Mgmt WLD	Edge uplink profile (Day-2 deployed by Edge cluster automation from SDDC Manager)	Load Balance Source	uplink-1,uplink-2	108	9000
VI WLD01	Host Overlay profile (deployed by SDDC Manager)	Load Balance Source	uplink-1,uplink-2	203	9000
VI WLD01	Edge uplink profile (Day 2 deployed by edge cluster automation from SDDC Manager)	Load Balance Source	uplink-1,uplink-2	208	9000

The following table shows the uplink profiles that are created when a dedicated NSX vDS is deployed with four uplinks.

Table 10. Second vDS with 4 uplinks - Mgmt and VI WLD Uplink Profiles

WLD Type	Profile	Default Teaming Policy	Active Uplinks	Transport VLAN (example)	Recommended MTU
Mgmt WLD	Host Overlay profile (deployed by Cloud Builder)	Load Balance Source	uplink-1,uplink-2, uplink-3,uplink-4	103	9000
Mgmt WLD	Edge uplink profile (Day-2 deployed by Edge cluster automation from SDDC Manager)	Load Balance Source	Any two of the four uplinks can be selected during deployment.	108	9000
VI WLD01	Host Overlay profile (deployed by SDDC Manager)	Load Balance Source	uplink-1,uplink-2, uplink-3,uplink-4	203	9000

WLD Type	Profile	Default Teaming Policy	Active Uplinks	Transport VLAN (example)	Recommended MTU
VI WLD01	Edge uplink profile (Day-2 deployed by edge cluster automation from SDDC Manager)	Load Balance Source	Any two of the four uplinks can be selected during deployment.	208	9000

Note: The Edge uplink profiles also include a Named Teaming policy which uses failover order for the uplink traffic. This allows traffic to be pinned to each physical network router for North/South traffic.

Each time a new VxRail cluster is added to an NSX-T VI WLD, a new host uplink profile is created to define the VLAN used for the host TEPs. The VLAN can be the same or different for each of the VxRail clusters.

For a single VxRail cluster VI WLD, two uplink profiles are required to complete the overall deployment of an NSX-T WLD, including the dynamic routing configuration. The host uplink profile is autogenerated when the VxRail cluster is added to the VI WLD. The edge uplink profile is created on Day 2 when adding an edge cluster. This can be done from SDDC Managers using the edge cluster automation feature.

Transport node profiles

A transport node as described earlier is either a host or an edge VM. The host transport uses a vDS for connectivity, whereas edge transport node uses the N-VDS. Each transport node can be added to one or more transport zones. Transport node profiles are used for host transport nodes. They contain the following information about the transport node that is vDS backed:

- Name
- Transport Zones for vDS participation – Overlay and VLAN TZ
- Uplink Profile
- IP Assignment type for the TEPs – DHCP or IP Pool
- Physical NIC Mapping – vmnics to uplinks.

The underlying vDS determines which pNICs are mapped in the transport node profile. If a system vDS is used for NSX-T, the pNIC can be mapped to the same pNICs used for external management, vMotion, or vSAN. If a dedicated vDS is used for NSX-T, the pNICs are selectable, and either two or four uplinks can be mapped.

The following table shows the settings that are applied to the Mgmt WLD and VI WLD with a VxRail vDS or the second NSX-T, or if a dedicated NSX vDS is used with only two uplinks.

Table 11. NSX-T transport node profiles with two uplinks

WLD Type	Transport Zones	Uplink Profile	IP Assignment	Physical NIC Mapping
Mgmt WLD	Host Overlay, VLAN	Mgmt WLD Host Uplink Profile	DHCP, IP Pool	pNIC1, pNIC2

WLD Type	Transport Zones	Uplink Profile	IP Assignment	Physical NIC Mapping
VI WLD01	Host Overlay	VI WLD Host Uplink Profile 01	DHCP, IP Pool	pNIC1, pNIC2

During the deployment of the Mgmt WLD by Cloud Builder, the following tasks are performed:

- A transport profile is created with the settings in the preceding table. When the management VxRail cluster is added to the NSX-T Mgmt WLD, the transport node profile is applied to the nodes in the VxRail cluster.
- The nodes are added to the transport zones,
- The TEPs are assigned an IP so that the hosts can communicate over the overlay network.

The following figure shows the Mgmt WLD node connectivity with single VxRail vDS with two uplinks that are used for the TEP traffic. The VMkernel interfaces that are used for the TEP traffic get their IPs assigned from a DHCP server or from an IP Pool. They communicate over the Host overlay VLAN defined before the deployment.

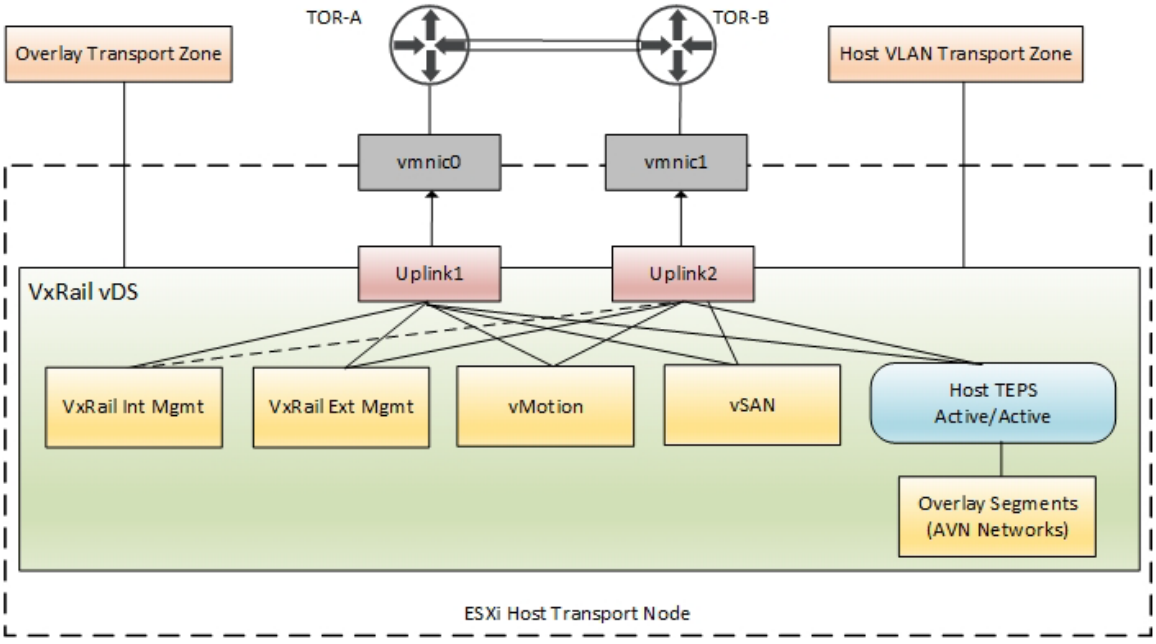

Figure 29. **Mgmt WLD transport node – single VxRail vDS (only two uplinks)**

Note: The preceding figure shows the AVN networks which are deployed when AVN is deployed from the SDDC Manager (Day-2).

The NSX-T VI WLD transport zone design is very similar to the Mgmt WLD. The two main differences are:

- No VLAN Transport Zone is added to the Transport Node profile during deployment.
- More VLAN transport zones can be added for the VI WLD nodes post deployment.

If NSX-T is deployed using a dedicated vDS, either two or four uplinks can be used, and the pNICs are selectable. The following table shows the transport node profile configuration with four uplinks.

Table 12. NSX-T transport node profiles with four uplinks

WLD Type	Transport Zones	Uplink Profile	IP Assignment	Physical NIC Mapping
Mgmt WLD	Host Overlay, VLAN	Mgmt WLD Host Uplink Profile	DHCP, IP Pool	User Selectable: pNIC1, pNIC2, pNIC3, pNIC4
VI WLD01	Host Overlay	VI WLD Host Uplink Profile 01	DHCP, IP Pool	User Selectable: pNIC1, pNIC2, pNIC3, pNIC4

The following figure shows a VI WLD node connectivity with a dedicated NSX vDS with two uplinks that are used for the TEP traffic. The VMkernel interfaces that are used for the TEP traffic get their IPs assigned from a DHCP server or from an IP Pool. They communicate over the Host overlay VLAN provided during the deployment.

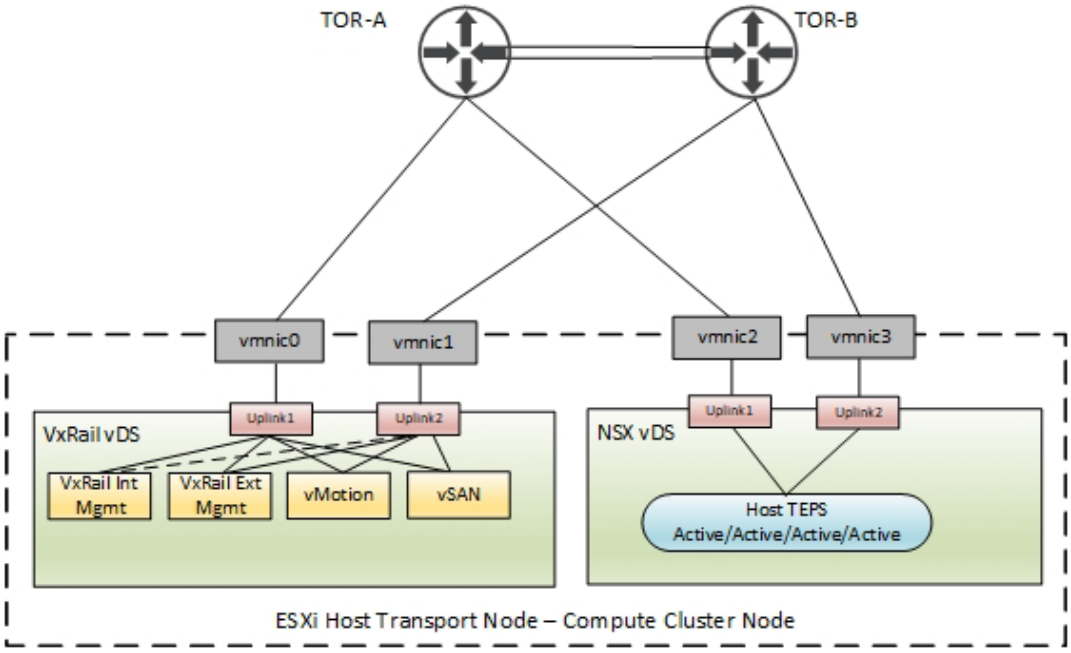

Figure 30. **VI WLD transport node – Second NSX vDS (can be two or four uplinks)**

NSX-T Edge node design

The Edge node design for the Mgmt WLD deployment with AVN deployed follows the VCF design. Starting with VCF 4.3, the edge nodes are deployed as a Day-2 activity from SDDC Manager. For the Mgmt WLD, two Edge node VMs are deployed in the Mgmt WLD VxRail cluster. The Edge nodes themselves have an N-VDS or NSX-T managed switch that is configured on them to provide the connectivity to external networks. The individual interfaces fp-eth0 and fp-eth1 on the N-VDS connect externally through a vDS using two different uplink port groups that are created in trunking mode. The vDS can be either a VxRail system vDS or a dedicated NSX vDS, depending on what network layout is required for the system and NSX traffic. Two TEPs are created on the edge N-VDS to provide East/West connectivity between the Edge nodes and the host transport nodes. This traffic is active/active using both uplinks which are defined in the uplink profile. The

management interface eth0 is connected to the vDS management port group. The following figure shows the connectivity for the Edge nodes running on the ESXi host in the Mgmt WLD cluster.

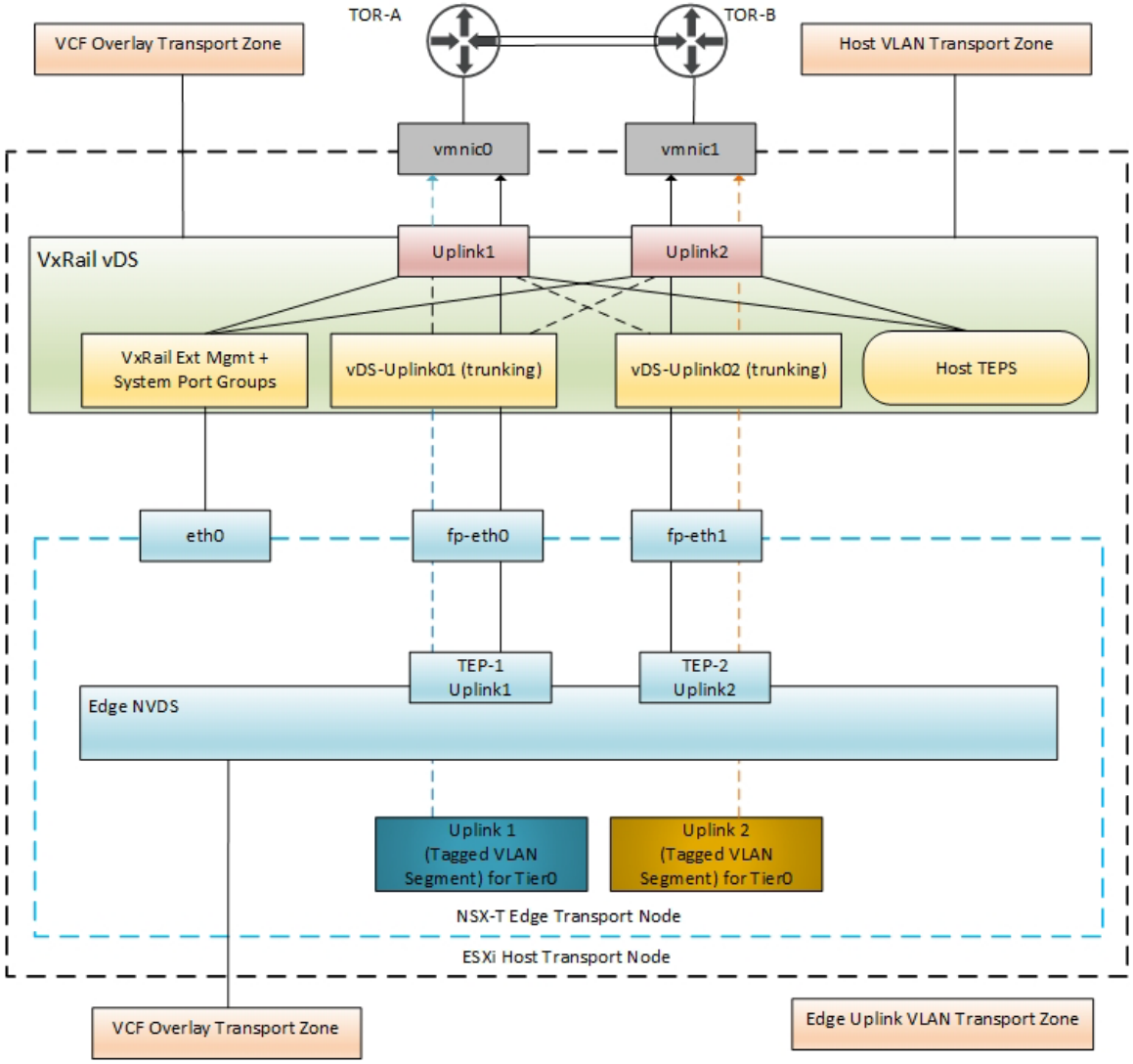

Figure 31. **Mgmt WLD - Edge node connectivity with single vDS**

Note: The uplink port groups used to connect the edge VM overlay interfaces are configured as trunk because the VLAN tagging is done by the N-VDS where the uplink profile defines the VLAN for the edge overlay.

The Edge node design for the VI WLD is very similar to the Mgmt WLD. If the Edge automation is used to deploy the edge cluster for a VI WLD, the same network configuration can be achieved. The following diagram shows the edge connectivity where the VxRail cluster was added to the VI WLD using a dedicated NSX-T vDS with two uplinks.

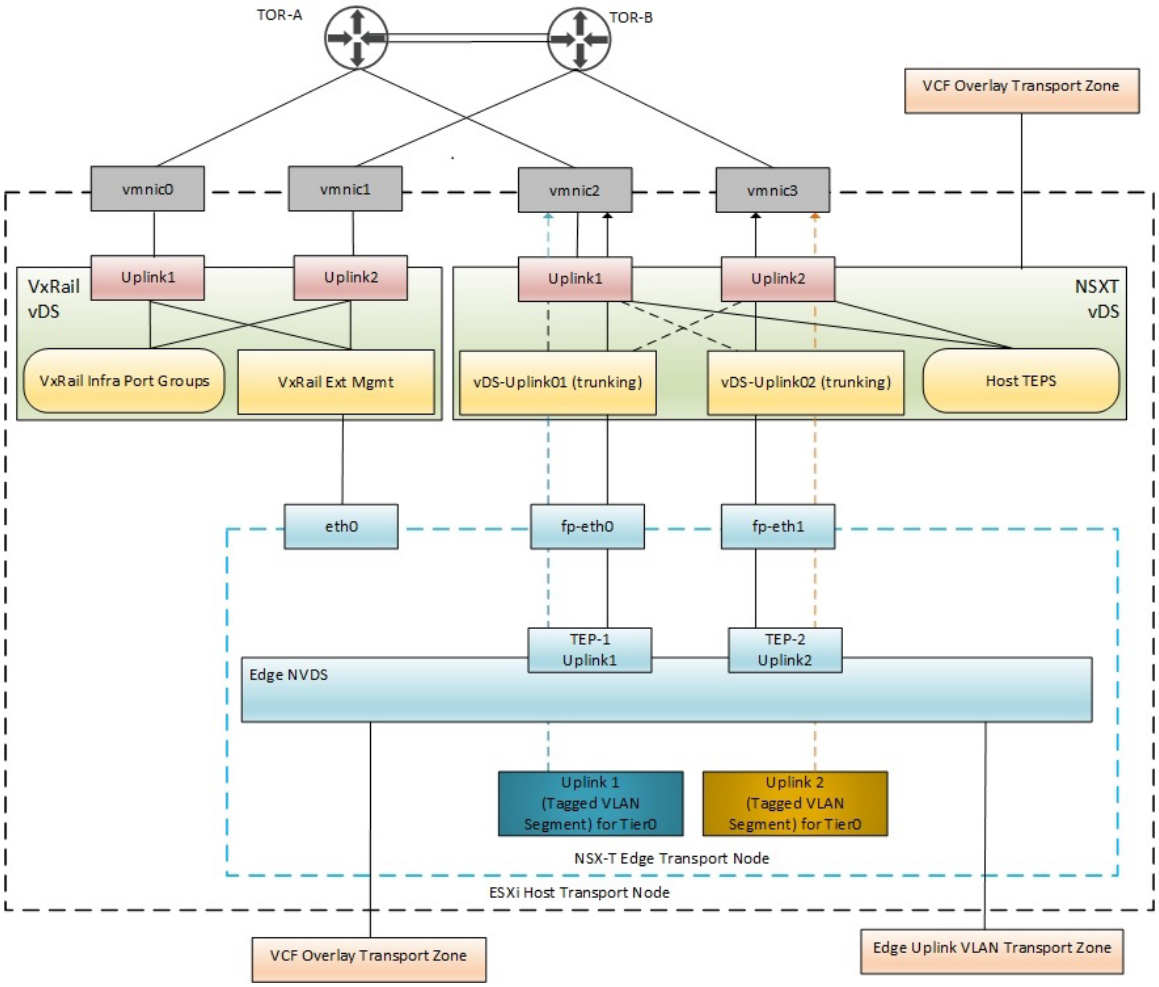

Figure 32. **VI WLD – Edge node connectivity with dedicated NSX vDS**

VCF on VxRail has a shared edge and compute cluster design. This means that the Edge node VMs TEP and uplink interfaces connect to the Host VDS for external connectivity and the same hosts can be used for user VMs that use the same host overlay.

North/South routing design

The NSX-T edge routing design is based on the VCF design that is located here Mgmt WLD Routing Design. A Tier-0 gateway is deployed in active/active mode with ECMP enabled to provide redundancy and better bandwidth utilization. Both uplinks are used. Two uplink VLANs are needed for North/South connectivity for the edge virtual machines in the Edge node cluster. The dedicated uplink profile that is created for the edge transport nodes defines named teaming policies. These policies are to be used in the edge uplink transport zone and the segments that are created for the Tier 0 gateway and use as a transit network to connect the Tier 0 interfaces. It is the named teaming policy that allows traffic from the Edge node to be pinned to an uplink network/VLAN connecting to the physical router. BGP provides dynamic routing between the physical environment and the virtual environment. eBGP is used between the Tier-0 Gateway and the physical TORs. An iBGP session is established between the T0 edge VMs SR components.

- VCF v4.3 and later support Static Routes if BGP is not a viable option for the customer.

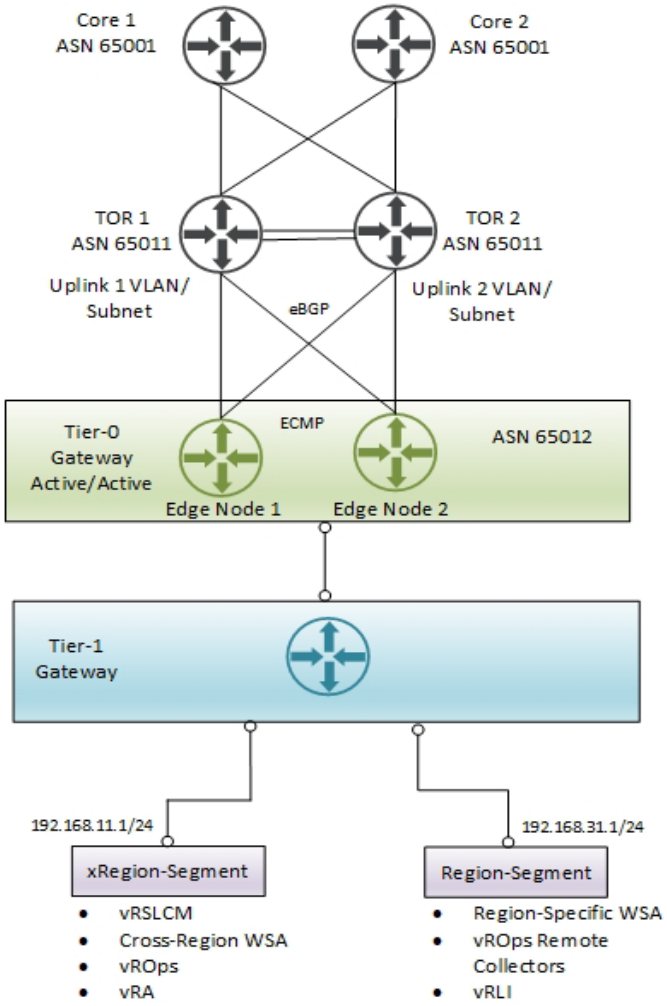

Figure 33. **Mgmt WLD Edge node North/South routing design**

NSX-T Mgmt WLD physical network requirements

The following NSX-T external network requirements must be met before deploying the Mgmt WLD:

- Minimum 1600 MTU for Geneve (Overlay) traffic, 9000 MTU is recommended.
- Host Overlay VLAN is created on the physical switches.
- Host Overlay VLAN added to the trunk ports connecting to the VxRail nodes.
- If TEP IP Pools are not used, DHCP is required to assign the Host TEPs IP.
- If DHCP is used for the Host TEPs, IP Helper is required on the switches if the DHCP server is in different L3 network.

For the Application Virtual Network (AVN) using an NSX-T overlay, the following additional requirements must be met before deployment:

- Layer 3 license requirement for peering with T0 Edges.
- BGP is configured for each router peering with a T0 Edge.
- Two Uplink VLANs for T0 Edge external connectivity to physical network.

- Edge Overlay VLAN is created on the physical switches.
- Uplink and Overlay VLANs added to the trunk ports connecting to VxRail nodes.

NSX-T VI WLD physical network requirements

The following NSX-T external network requirements must be met before deploying the Mgmt WLD:

- Minimum 1600 MTU for Geneve (Overlay) traffic, 9000 MTU is recommended.
- Host Overlay VLAN is created on the physical switches.
- If TEP IP Pools are not used, DHCP is required to assign the Host TEPs IP.
- If DHCP is used for the Host TEPs, IP Helper is required on the switches if the DHCP server is in different L3 network.

If NSX-T Edges are going to be deployed as per VCF design for the VI WLD Edge cluster, the following additional requirements must be met before deployment:

- Layer 3 license requirement for peering with T0 Edges.
- BGP is configured for each router peering with a T0 Edge.
- Two Uplink VLANs for T0 Edge external connectivity to physical network.
- Edge Overlay VLAN is created on the physical switches.

NSX-T deployment in Mgmt WLD

Cloud builder is used to deploy the NSX-T components in the Mgmt WLD VxRail. The following list highlights the major steps that are performed during the deployment process:

1. Deploy NSX-T Managers in Mgmt WLD VxRail cluster.
2. Create anti-affinity rules for the NSX-T Managers.
3. Set VIP for NSX-T Managers.
4. Add Mgmt WLD vCenter as a compute manager.
5. Assign NSX-T license.
6. Create overlay transport zone.
7. Create a VLAN transport zone.
8. Create a host uplink profile.
9. Create transport node profile.
10. Prepare the hosts in the VxRail cluster for NSX-T.

Starting with VCF 4.3, the deployment of AVN is a Day-2 activity performed from the SDDC Manager. The following tasks are also performed to deploy and configure the necessary components to provide the connectivity and routing for the NSX-T overlay backed networks for AVN.

1. Create Edge Uplink Profile.
2. Create Named Teaming Policy for Uplink traffic.

3. Create Trunked Uplink Port Groups on the vDS.
4. Deploy two Edge VMs.
5. Create anti-affinity rules.
6. Create Edge cluster.
7. Create Uplinks for T0.
8. Configure T0 and BGP.
9. Configure T1.
10. Verify BGP Peering with TORs.
11. Create AVN Segments (Region A and xRegion).

Once the Mgmt WLD has been deployed, it should contain the components that are shown in the following figure. The edges are only deployed as a Day-2 activity from the SDDC Manager.

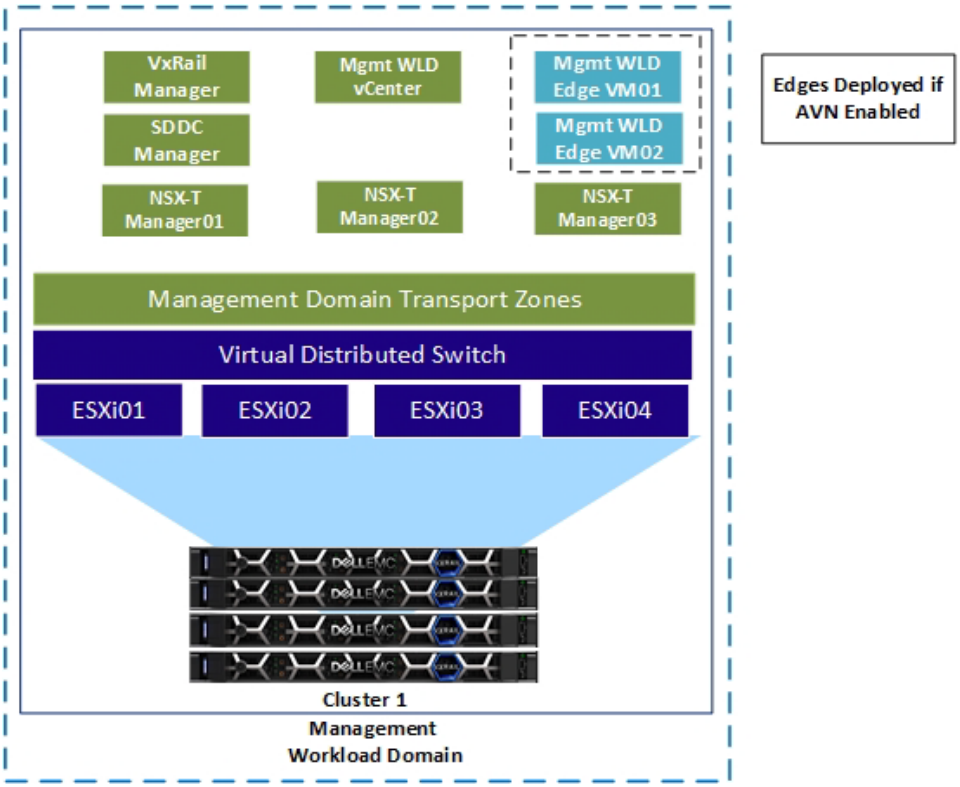

Figure 34. **Mgmt WLD after deployment**

NSX-T deployment in VI WLD

The NSX-T components are installed when the first VxRail cluster is added to the first NSX-T VI WLD. The SDDC Manager deploys the NSX-T Managers in the management workload domain, assigns an IP address to each NSX-T Manager virtual appliance and a virtual IP, and configures the VI WLD VxRail cluster to be used for NSX-T services. The following list highlights the major steps that are performed during the deployment process:

1. Deploy NSX-T Managers in Mgmt WLD VxRail cluster.

2. Create anti-affinity rules for the NSX-T Managers.
3. Set VIP for NSX-T Managers.
4. Add VI WLD vCenter as a Compute Manager.
5. Assign NSX-T license.
6. Create Overlay Transport zone.
7. Create an Uplink profile.
8. Create Transport Node Profile.
9. Prepare the hosts in the VxRail cluster for NSX-T.

Note: No additional NSX-T Managers are needed when a second NSX-T based VI WLD is added. However, starting with VCF 4.0, you can deploy a new NSX-T domain for each WLD if that is a requirement, 1:1 NSX-T for each VI WLD.

A new feature in VCF 4.0 allows NSX-T Edges to be deployed using automation for the VI WLDs using SDDC Manager. This allows the Edges to be automatically deployed in a consistent fashion per VCF guidance.

The following figure shows the components that are deployed in the MGMT VI WLD after a VI WLD has been deployed and two VxRail clusters have been added to the VI WLD. It shows the two NSX-T Edges in the first VxRail cluster of the VI WLD that can be deployed using Edge automation feature in SDDC Manager.

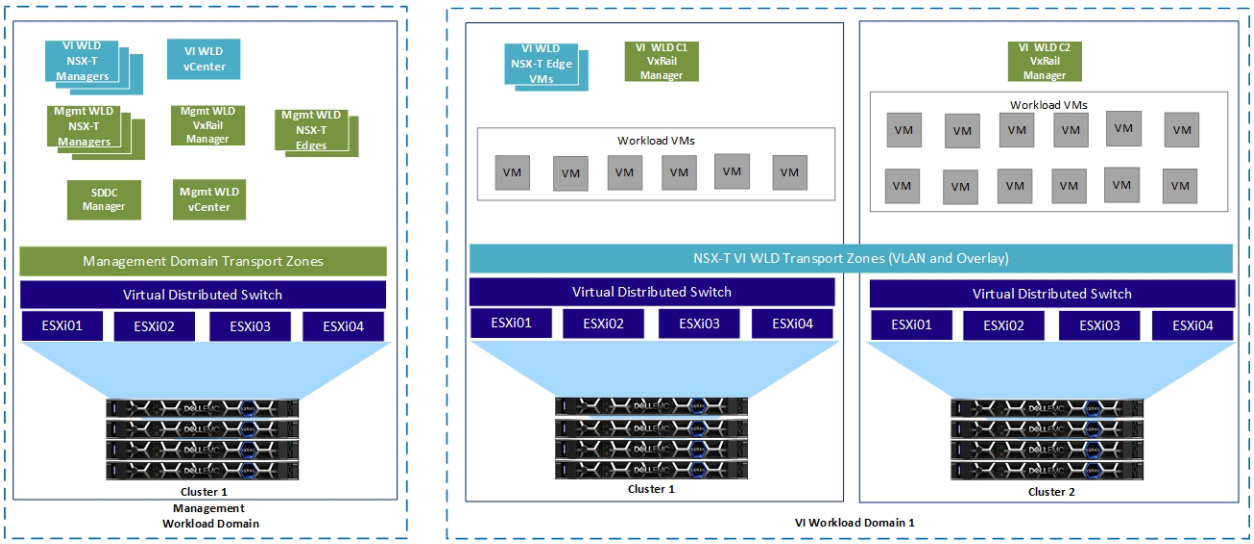

Figure 35. **NSX-T VI WLD VxRail Cluster Design**

Chapter 7 Enabling VCF with Tanzu Features on Workload Domains

This chapter presents the following topics:

Introduction

With VCF 4.0, you can enable a VI WLD for VCF with Tanzu. This enablement is known as Workload Management, where you can deploy and operate the compute, networking, and storage infrastructure required by VCF with Tanzu. VCF with Tanzu transforms vSphere to a platform for running Kubernetes workloads natively on the hypervisor layer. When enabled on a VxRail cluster, VCF with Tanzu provides the capability to run Kubernetes workloads directly on ESXi hosts and to create upstream Kubernetes clusters within dedicated resource pools. The workload management is enabled on the VI WLD through the Solutions deployment option that is found in SDDC Manager UI.

Prerequisites

The following prerequisites must be met before starting the workload management:

- Licensing: Within a WLD, all hosts within the selected VxRail clusters must have the proper vSphere for Kubernetes licensing to support Workload Management.
- Workload Domain: A VI WLD deployed as Workload Management ready must be available.
- NSX-T Edge Cluster: At least one NSX-T Edge cluster must be deployed from SDDC Manager and available.
- IP Addresses
  - Define a subnet for pod networking (nonroutable), minimum of a /22 subnet.
  - Define a subnet for Service IP addresses (nonroutable), minimum of a /24 subnet.
  - Define a subnet for Ingress (routable), minimum of a /27subnet.
  - Define a subnet for Egress (routable), minimum of a /27 subnet.

VCF with Tanzu detailed design

To learn more about the Kubernetes for vSphere detailed design, see the following VCF documentation: Detailed Design of Developer Ready Infrastructure for VMware Cloud Foundation.

Chapter 8 Physical Network Design Considerations

This chapter presents the following topics:

Introduction

The VCF on VxRail network design offers flexibility to allow for different topologies and different network hardware vendors. This enables you to use your existing network infrastructure or potentially add new hardware to an existing data center network infrastructure. Typically, data center network design has been shifting away from classical 3-tier network topologies using primarily Layer 2 fabric to the newer Leaf and Spine Layer 3 fabric architectures. When deciding whether to use Layer 2 or Layer 3, consider the following:

- NSX-T ECMP Edge devices establish Layer 3 routing adjacency with the first upstream Layer 3 device to provide equal cost routing for management and workload traffic.
- The investment that you have today in your current physical network infrastructure.
- The advantages and disadvantages for both Layer 2 and Layer 3 designs

The following section describes both designs and highlights the main advantages and disadvantages of each design.

Traditional 3-tier (access/core/aggregation)

The traditional 3-tier design is based on a Layer 2 fabric, as shown in the following figure:

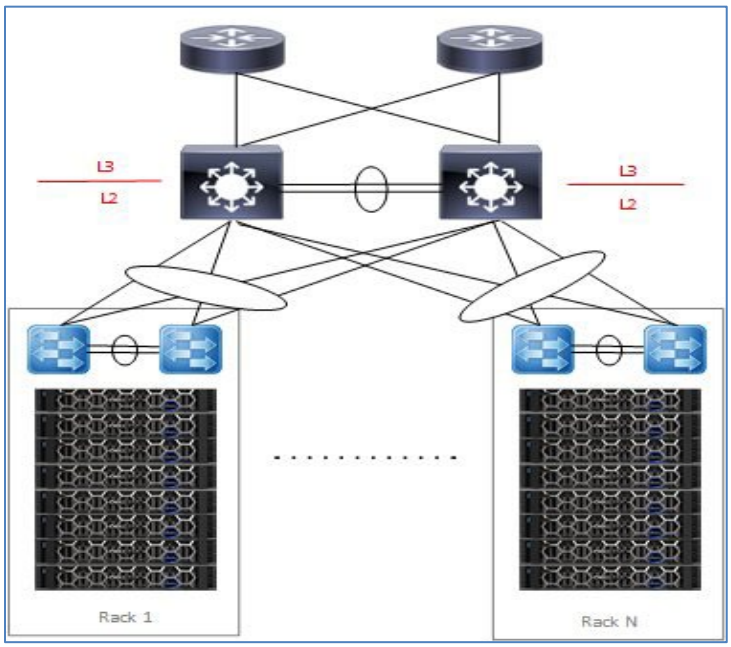

Figure 36. **Traditional 3-Tier Layer 2 Fabric Design**

It has the following characteristics:

- VLANs carried throughout the fabric –increases the size of the broadcast domain beyond racks if multiple racks are needed for the infrastructure and VxRail clusters span racks.
- The aggregation layer devices of each pod are the demarcation line between L2 and L3 network domains.
- Default Gateway – HSRP/VRRP at the aggregation layer
- The NSX-T T0 Gateway peers with the routers at the aggregation layer.

Advantages:

- VLANs can span racks which can be useful for VxRail system VLANs like vSAN/vMotion and node discovery.
- Layer 2 design might be considered less complex to implement.

Disadvantages:

- Large VxRail clusters spanning racks will create large broadcast domains.
- Interoperability issues between different switch vendors can introduce spanning tree issues in large fabrics.
- The NSX-T T0 gateways for each WLD needs to peer at the aggregation layer. For large-scale deployments with multiple WLDs, the configuration becomes complex.
- The size of such a deployment is limited because the fabric elements have to share a limited number of VLANs 4094. With NSX, the number of VLANs could be reduced so this might not be an issue.

Leaf and Spine Layer 3 fabric

The Layer 3 Leaf and Spine design is becoming the more adopted design for newer, more modern data center fabrics depicted in the following figure:

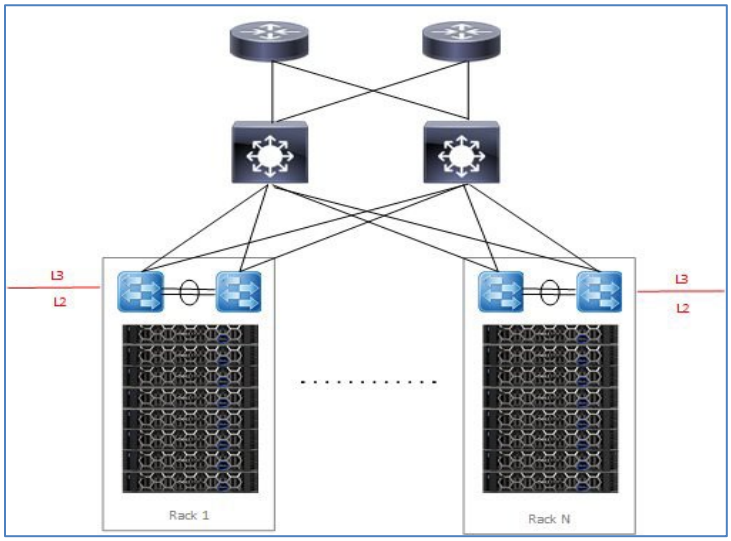

Figure 37. **Leaf and Spine Layer 3 Design**

It has the following characteristics:

- L3 is terminated at the leaf, thus all the VLANs originating from ESXi hosts terminate on leaf.
- The same VLANs can be reused for each rack.
- The leaf switches provide default gateway functionality.
- The NSX-T T0 Gateway for the WLD peers with the leaf switches in one rack.

Advantages:

- Vendor agnostic - Multiple network hardware vendors can be used in the design.

- Reduced VLAN span across racks, thus smaller broadcast domains.
- East/West for an NSX domain can be confined within a rack with intra-rack routing at the leaf.
- East/West across NSX domains or cross-rack is routed through the spine.
- NSX-T Tier0 peering is simplified by peering the WLDs with the leaf switches in the rack.

Disadvantages:

- The Layer 2 VLANs cannot span racks. VxRail clusters that span racks require a solution to allow VxRail system traffic to span racks using hardware VTEPs.
- The Layer 3 configuration might be more complex to implement.

Multirack design considerations

It might be desirable to span WLD VxRail clusters across racks to avoid a single point of failure within one rack. The management VMs running on the Mgmt WLD VxRail cluster and any management VMs running on the VI WLD require VxRail nodes to reside on the same L2 management network. This ensures that the VMs can be migrated between racks and maintain the same IP address. For a Layer 3 Leaf-Spine fabric, this is a problem as the VLANs are terminated at the leaf switches in each rack.

VxRail cluster across Racks

VxRail clusters deployed across racks require a network design that allows a single (or multiple) VxRail clusters to span between racks. This particular solution uses a Dell PowerSwitch hardware VTEP to provide an L2 overlay network to extend L2 segments over an L3 underlay network for VxRail node discovery, vSAN, vMotion, management, and VM/App L2 network connectivity between racks. The following diagram is an example of a multirack solution using hardware VTEP with VXLAN BGP EVPN. The advantage of VXLAN BGP EVPN over static VXLAN configuration is that each VTEP is automatically learned as a member of a virtual network from the EVPN routes received from the remote VTEP.

For more information about Dell Network solutions for VxRail, see the Dell VxRail Network Planning Guide.

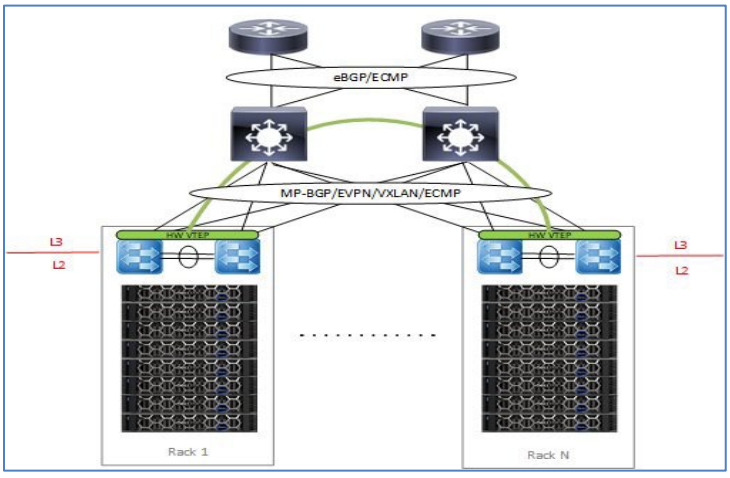

Figure 38. **MultiRack VxRail cluster with hardware VTEP**

VxRail physical network interfaces

VxRail can be deployed with either 2x10/2x25 GbE, 4x10 GbE, or 4x25 GbE predefined profiles. Starting from VxRail version 7.0.130, custom profiles can be used. VxRail needs the necessary network hardware to support the initial deployment. The following table illustrates various different physical network connectivity options with a single system vDS and a dedicated NSX vDS with and without NIC level redundancy. Note for this table, a standard wiring config can be used for connectivity: odd-numbered uplinks cabled to Fabric A, even-numbered uplinks cabled to Fabric B.

Notes:

- 100 Gbps PCIe adapters are supported using custom profiles.
- For the dedicated NSX-T vDS deployed by VCF, the vmnic to uplink mapping is in lexicographic order so this must be considered during the design phase.

				VxRail vDS				NSX-T vDS				
Option	Dedicated VDS for NSX-T	Uplinks per vDS	NIC Redundancy	Uplink 1	Uplink 2	Uplink 3	Uplink 4	Uplink 1	Uplink 2	Uplink 3	Uplink 4	
A	No	2	No	NDC-1	NDC-2							
B	No	2	Yes	NDC-1	PCI1-2							
C	No	4	No	NDC-1	NDC-2	NDC-3	NDC-4					
D	No	4	Yes	NDC-1	PCI1-2	NDC-2	PCI1-1					
E	Yes	2	No	NDC-1	NDC-2			NDC-3	NDC-4			
F	Yes	2	No	NDC-1	NDC-2			PCI1-1	PCI1-2			
G	Yes	2	Yes	NDC-1	PCI1-2			NDC-2	PCI1-1			
H	Yes	4/2	No	NDC-1	NDC-2	NDC-3	NDC-4	PCI1-1	PCI1-2			
I	Yes	4/2	Yes	NDC-1	PCI1-2	NDC-2	PCI1-1	PCI1-3	PCI1-4			
J	Yes	2/4	No	NDC-1	NDC-2			PCI1-1	PCI1-2	PCI1-3	PCI1-4	
K	Yes	4	No	NDC-1	NDC-2	NDC-3	NDC-4	PCI1-1	PCI1-2	PCI1-3	PCI1-4	
L	Yes	4	Yes	NDC-1	PCI1-2	NDC-2	PCI1-1	NDC-3	PCI1-4	PCI2-14	PCI2-2	

Table 13. Physical network connectivity options

The following diagrams illustrate some of the different host connectivity options from the preceding table for the different VxRail deployment types for either the Mgmt WLD or a VI WLD. For the Mgmt WLD and the VI WLD, the Edge overlay and the Edge Uplink networks will be deployed when NSX-T Edges are deployed using Edge automation in SDDC Manager. There are too many different options to cover in this section. The following is a selection of the most common options from Table 13.

Note: The PCIe card placements in the following diagrams are for illustration purposes only and might not match the configuration of the physical server. See the official VxRail documentation for riser and PCIe placement.

Single VxRail vDS connectivity options

This section illustrates the physical host network connectivity options for different VxRail profiles and connectivity options when only using the single VxRail vDS.

10 GbE connectivity options

This diagram illustrates option **A** in Table 13. The VxRail deployed with 2x10 predefined network profile on the 4-port NDC. The remaining two ports are unused, and can be used for other purposes if required.

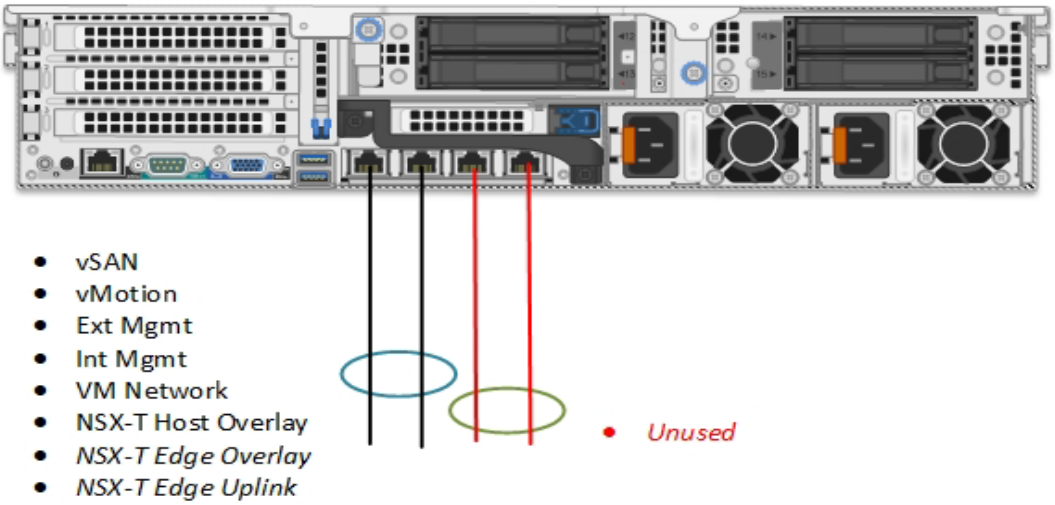

Figure 39. **Single VxRail vDS - 2x10 predefined network profile**

The next diagram refers to option **C** in Table 13. The VxRail is deployed with a 4x10 predefined network profile. This places vSAN and vMotion onto their own dedicated physical NICs on the NDC and NSX-T traffic will use vmnic0 and vmnic1 shared with management traffic. More PCI card can be installed and used for other traffic if that is required.

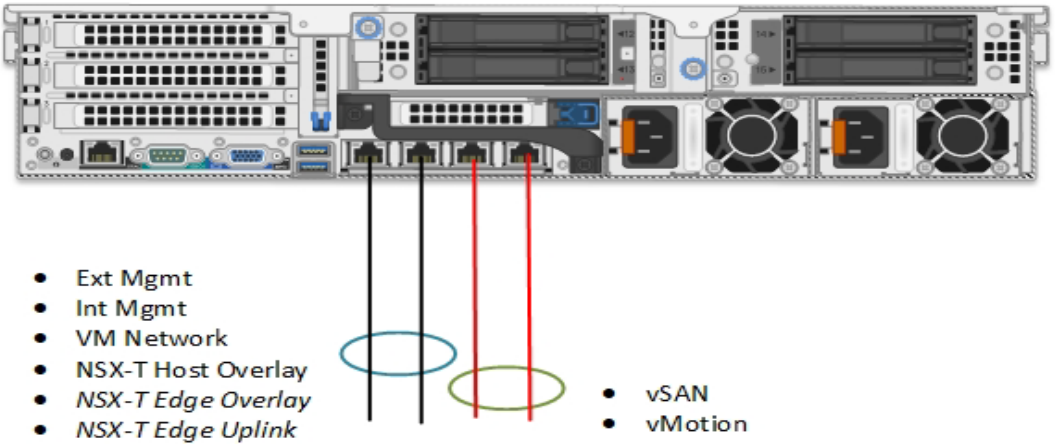

Figure 40. **Single VxRail vDS - 4x10 predefined network profile**

The final 10 GbE option in this next diagram will provide NIC level redundancy. This can be achieved using an NDC and PCIe in conjunction with using a custom profile to deploy the VxRail vDS. This is option D in Table 13.

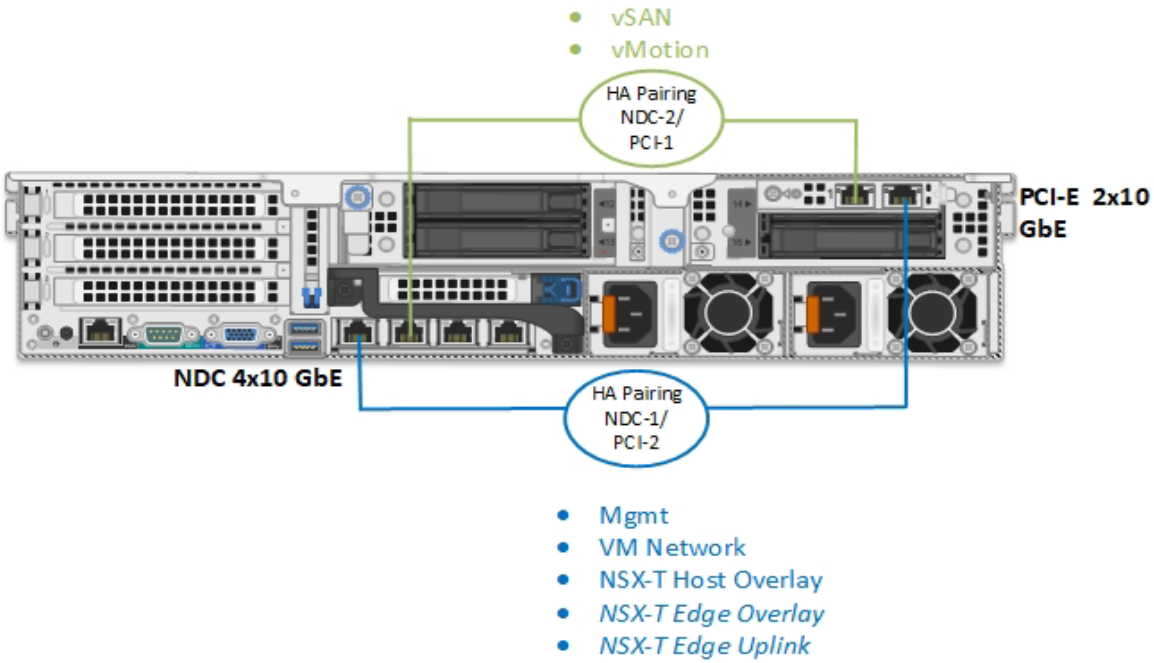

Figure 41. **Single VxRail vDS – 4x10 custom profile and NIC level redundancy**

25 GbE connectivity options

The first option aligns with option **A** in Table 13. A single VxRail vDS using 2x25 network profile for the VxRail using the 25GbE NDC. The VxRail system traffic uses the two ports of the NDC along with NSX-T traffic.

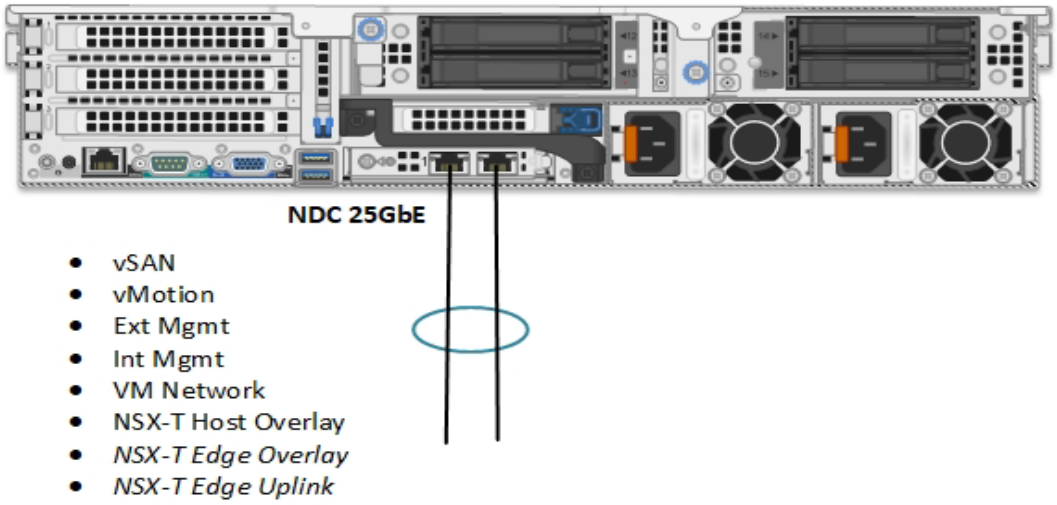

Figure 42. **Single VxRail vDS - 2x25 network profile**

As with the previous option, additional PCIe cards can be added to the node for other traffic, for example, backup, replication, and so on.

The second option aligns with option **D** in Table 13. A single VxRail vDS using a custom network profile which provides NIC level redundancy for the VxRail system traffic and the NSX-T TEP and Edge uplink traffic. A standard physical cabling configuration can be achieved with the logical network configuration described in section VxRail vDS custom profiles.

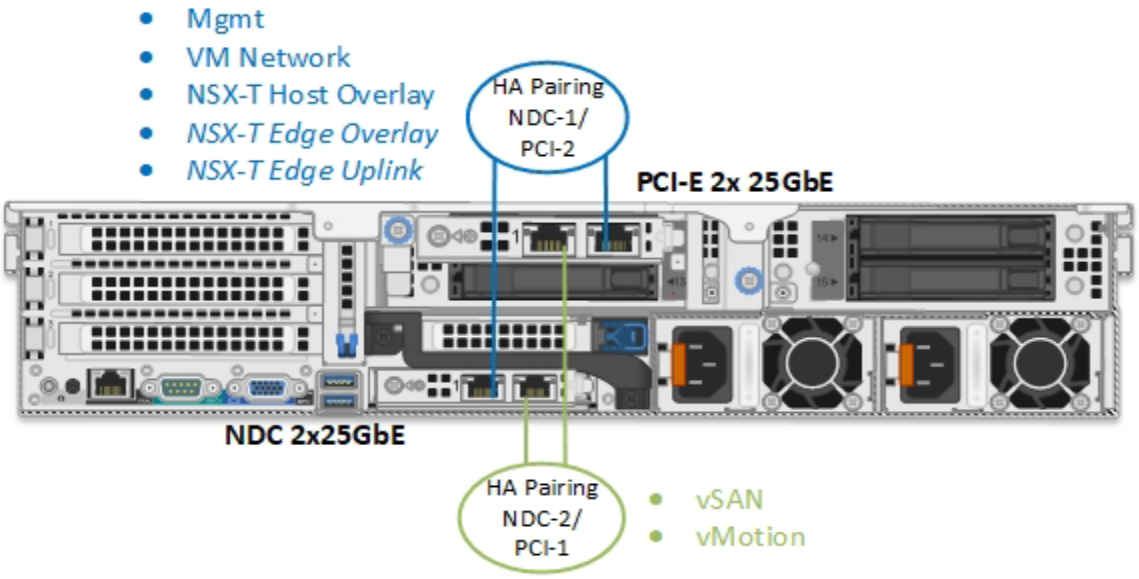

Figure 43. **Single VxRail vDS - 4x25 custom network profile**

NSX-T vDS connectivity options

This section illustrates the physical host network connectivity options for different VxRail profiles and connectivity options when only using a dedicated vDS for NSX traffic. The VxRail vDS is only used for system traffic.

10 GbE Connectivity Options

The following diagram illustrates a VxRail deployed with 2x10 predefined network profile on the 4-port NDC which aligns to option **E** in Table 13. The remaining two ports are used for the second vDS for the NSX traffic.

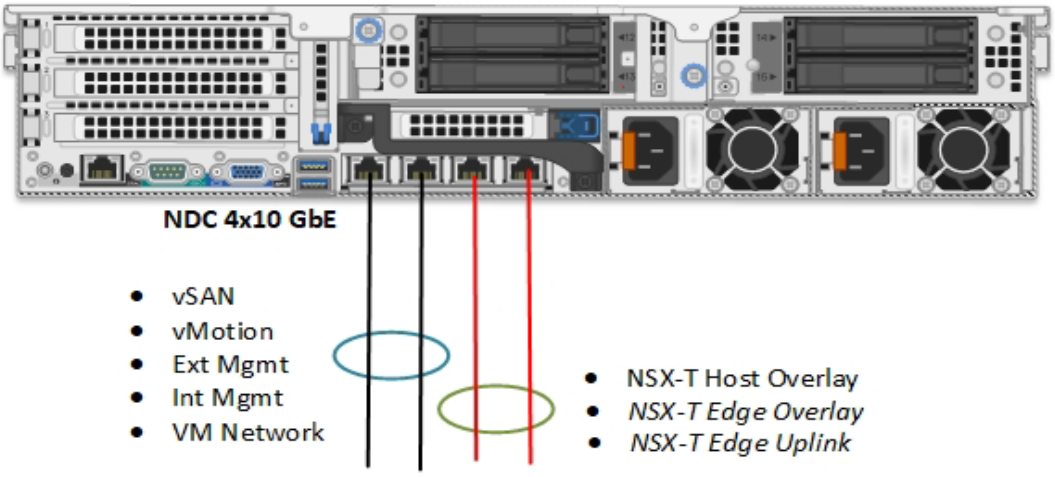

Figure 44. **VxRail vDS and NSX vDS using two ports each on 4x10 NDC**

The second option is VxRail deployed with a 4x10 predefined network profile consuming all four ports of the NDC. This places vSAN and vMotion onto their own dedicated physical NICs on the NDC. The NSX-T traffic uses a dedicated vDS and uplinks connecting to pNICs on the PCI-E 10 GbE, this aligns to option **H** in Table 13.

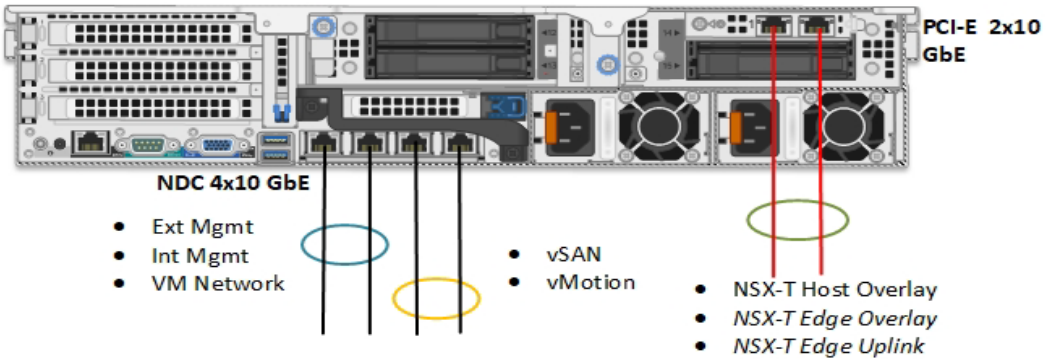

Figure 45. **VxRail vDS using four ports of NDC and NSX vDS using PCI-E**

The final option for a 10GbE network environment we want to illustrate provides NIC level redundancy across the NDC and PCIe for both system traffic on the VxRail vDS and NSX-T traffic on the dedicated NSX-T vDS. The VxRail is deployed with the custom profile option using one port from NDC and one from PCIe. Similarly, the NSX-T vDS will use a port from each NIC. This aligns to option **G** in Table 13.

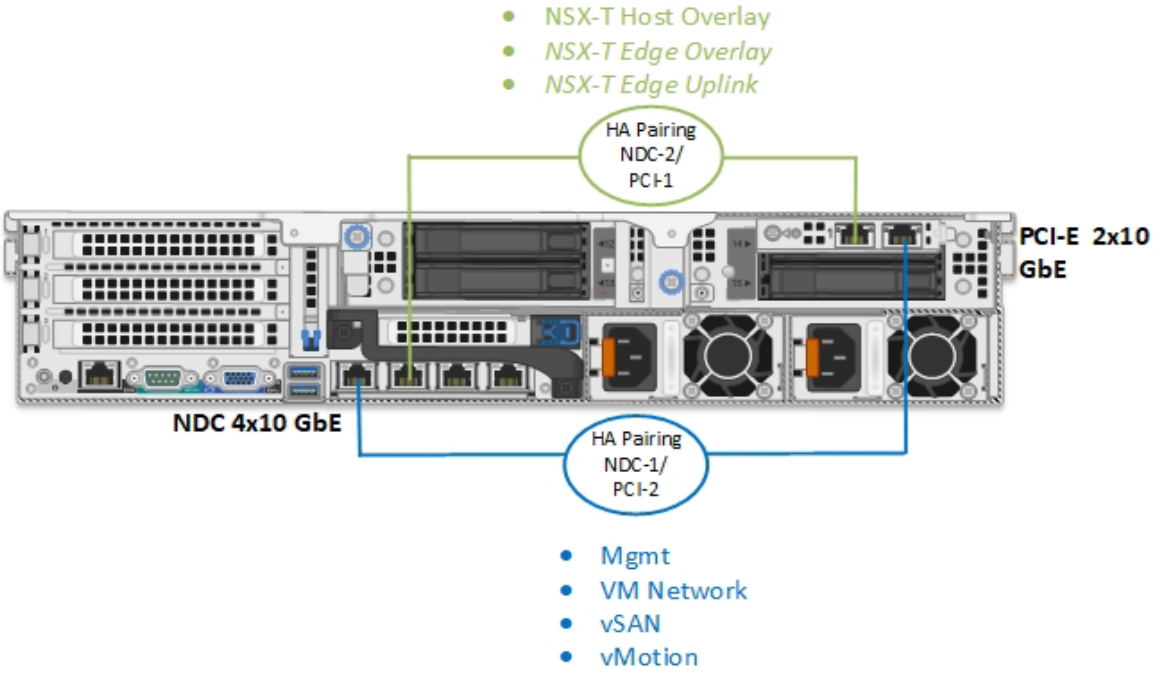

Figure 46. **VxRail vDS and NSXT vDS with NIC level redundancy**

25 GbE Connectivity Options

The first 25 GbE option shown in the following diagram uses the 2-port 25 GbE NDC for the VxRail vDS. A dedicated vDS is created for the NSX-T traffic using the two ports of the PCI-E card, option **F** in Table 13.

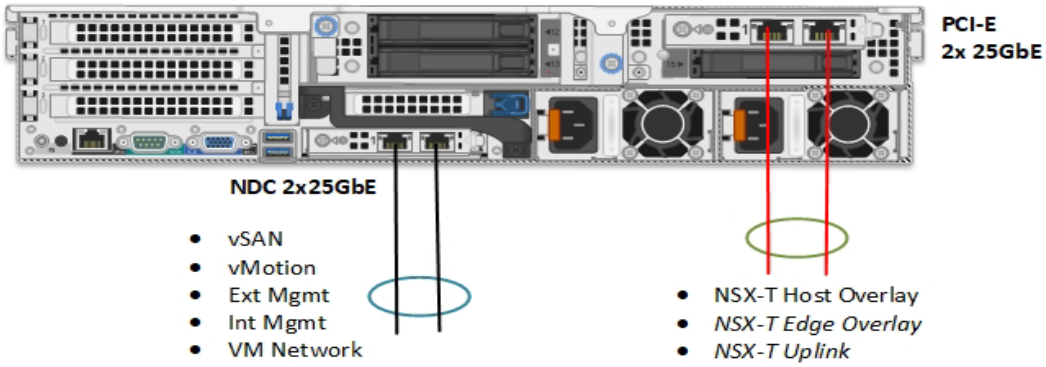

Figure 47. **VxRail vDS using 25 GbE NDC and NSX vDS using 25 GbE PCI-E**

As with the previous option, additional PCIe cards can be added to the node for other traffic, for example, backup, replication, and so on.

The second option requires a total of six 25GbE ports. The VxRail is deployed with the 4x25 custom profile option as previously discussed using the two port NDC. The two port PCIe and the second vDS for NSX-T traffic require an additional 2x25 GbE card. This aligns to option **I** in Table 13.

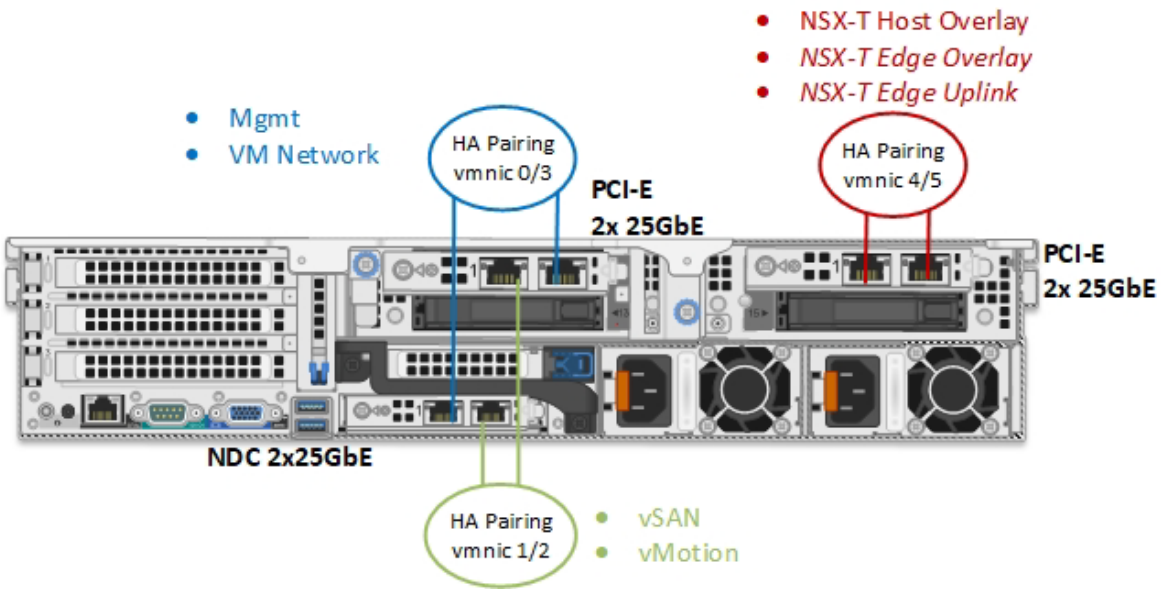

Figure 48. **VxRail vDS using 2x25 GbE NDC and PCI-E, NSX vDS using 25 GbE PCI-E**

The following last option provides full NIC level redundancy for both VxRail system traffic and also NSX-T traffic using two NICs NDC and PCIe connected to the TOR switches. The VxRail is deployed with the 2x25 custom profile using a port from NDC and the PCIe. The dedicated vDS for NSX-T traffic uses the remaining free pNIC on each NIC with one interface from each NIC connected to each TOR.

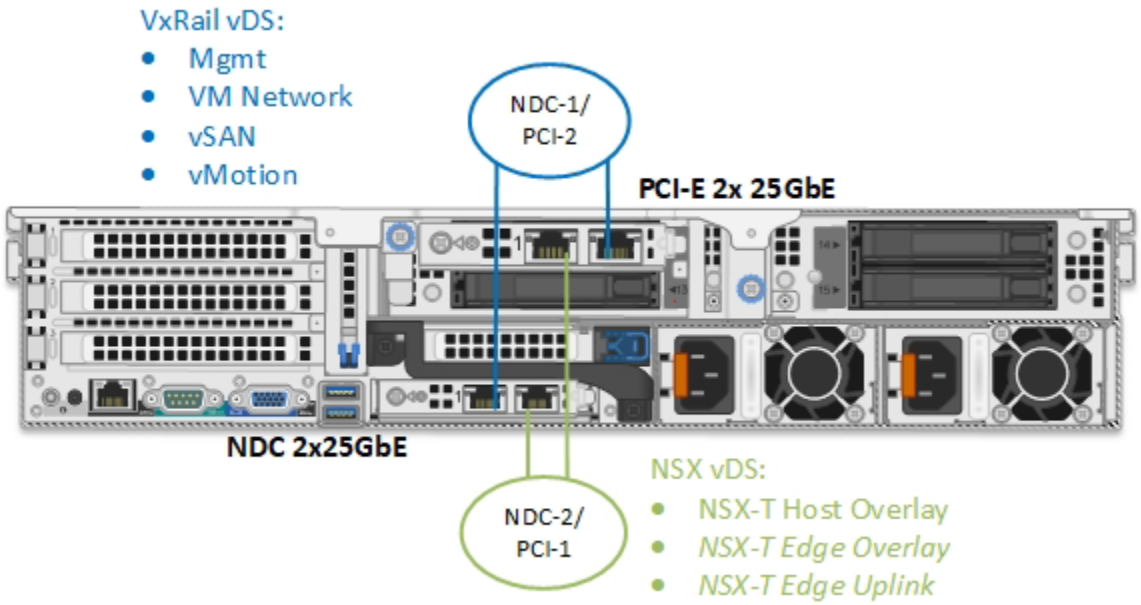

Figure 49. **25 GbE NIC Level redundancy for system and NSXT traffic**

There are additional options that support up to eight pNICs for both 10 GbE and 25 GbE networks that are not illustrated here, but are included in Table 13 for reference.

Chapter 9 Storage Options

This chapter presents the following topics:

Introduction

VMware Cloud Foundation on VxRail has flexible storage options that include vSAN or Fibre Channel storage (introduced in VCF 4.3.1) as primary storage option for a VI WLD. The Management WLD primary VxRail cluster still requires vSAN storage as its primary storage. iSCSI, NFS storage, and vSAN HCI Mesh can be used as supplemental storage for both the Management and VI WLDs. The following table shows the different storage options available.

Storage Option	Mgmt WLD VxRail Cluster	VI WLD VxRail Cluster
vSAN	Yes	Yes
vSAN HCI Mesh	No	Supplemental Storage only
FC Storage (VMFS)	Supplemental Storage only	Yes
iSCSI Storage	Supplemental Storage only	Supplemental Storage only
NFS Storage	Supplemental Storage only	Supplemental Storage only

Table 14. Storage Options

Note: In a consolidated architecture, a second or additional VxRail cluster in the Management WLD would have the same support as a VI WLD VxRail cluster.

vSAN

vSAN is deployed, scaled, and lifecycle-managed with SDDC Manager and VxRail HCI System software automation. Workload domains with VxRail vSAN clusters can be configured quickly and can be ready to use without having to make complex changes within VxRail hardware. vSAN Storage Policy Based Management (SPBM) allows for storage characteristics such as primary failures to tolerate and disk striping. Storage settings can be changed in software at a VM or object-level nondisruptively with just a few clicks. Compare this with traditional array-based storage which often requires hardware changes and updates across a whole LUN or volume which can be time consuming and often risky.

vSAN HCI Mesh

Introduced in VCF version 4.2 vSAN HCI Mesh supports the sharing of spare vSAN storage capacity between VxRail clusters in a VI workload domain. It can only be used as a secondary storage option and either vSAN or FC storage need to be used for primary storage. It enables an alternative method to alleviate decreasing storage capacity in a VI workload domain. This can be

useful for environments that are not compute-constrained in a VI workload domain in order to avoid adding nodes which increase both compute and storage resources.

Prerequisites The following prerequisites must be met before configuring vSAN HCI Mesh:

- All VxRail clusters participating in an HCI mesh topology must be managed by a single vCenter instance.
- All VxRail clusters participating in an HCI mesh topology located under a single data center instance.
- Data center network configured to enable connectivity between server VxRail cluster and client VxRail cluster.
- The client VxRail cluster can mount no more than 5 remote vSAN datastores from server VxRail clusters.
- The server VxRail cluster's vSAN datastore is not mounted by more than five client VxRail clusters.

Feature Support The configuration of the client and server VxRail clusters in a vSAN HCI mesh is performed in the VxRail cluster level using the vClient. SDDC manager detects the presence of a vSAN HCI Mesh in a VI workload domain and alerts the user through SDDC Manager. All existing VCF on VxRail workflows are compatible with vSAN HCI mesh. There are constraints built into SDDC Manager to prevent the removal of a dependent VI workload domain or dependent VxRail cluster if there is a sharing of a vSAN datastore in effect.

FC Storage

VCF 4.3.1 introduced the ability to deploy a VI WLD with external storage as primary storage with no VxRail managed vSAN. This allows support for a VxRail dynamic node cluster with three nodes or more to be added to a VI WLD. These nodes contain no internal storage disks normally required for vSAN.

Requirements The following requirements must be met before deploying a VxRail dynamic node cluster with FC storage as primary storage.

Supported storage arrays:

- PowerStore
- PowerMax
- UnityXT

Supported FC HBA:

- Emulex LPE 35002 Dual Port 32 Gb HBA
- Emulex LPE 31002 Dual Port 16 Gb HBA
- QLogic 2772 Dual Port 32 Gb HBA
- QLogic 2692 Dual Port 16 Gb HBA

Storage Configuration required:

- Zoning of the Vxrail nodes to the Storage system
- Creation and Masking of the LUN to the VxRail nodes
- 900GB of free space on the Volume
- Formatting the LUN with a Virtual Machine File system (VMFS)
- If multiped datastore are discovered during VxRail Day 1 workflow the largest one will be selected as the primary datastore for the VxRail systems
- If multiped datastore are discovered with the same size a random one will be select during VxRail Day 1 workflow.

Note: VCF 4.4 requires Workflow Optimization for 14/15G VI WLD deployments.

Chapter 10 Multisite Design Considerations

This chapter presents the following topics:

Introduction

The VCF on VxRail solution natively supports two different multisite options depending on the distance and latency between the sites and the type of protection needed for the workloads. Multi availability zones are offered by VxRail vSAN stretched clusters for the Mgmt WLD and VI WLDs between the two availability zones. This is typically only for sites within the same Metro area due to the latency requirements for stretched vSAN. New to VCF 4.2 NSX-T federation is now supported which enables support for dual-region VCF instances which can be located at much greater distances as there is no stretched cluster requirement.

Multi-AZ (VxRail vSAN stretched-cluster)

All WLDs can be stretched across two availability zones. Availability zones can be located in either the same data center but in different racks or server rooms, or in two different data centers in two different geographic locations. They are typically in the same Metro area. The VxRail vSAN stretched cluster configuration combines both standard VxRail procedures and automated steps that are performed by using a script from dev center that can be copied and run from SDDC Manager. The vSAN Witness is manually deployed and configured, and the SDDC Manager automates the configuration of the VxRail vSAN stretched cluster.

The following general requirements apply to a VCF on VxRail vSAN stretched cluster deployments.

- Witness is deployed at a third site using the same vSphere version that is used in the VCF on VxRail release.
- All VxRail vSAN stretched cluster configurations must be balanced with the same number of hosts in AZ1 and AZ2.
- A minimum of four nodes is required at each site for the management WLD.
- A minimum of three nodes is required at each site for a VI WLD.

Note: The VI WLD VxRail clusters can only be stretched if the Mgmt WLD VxRail cluster is first stretched.

The following network requirements apply for the Mgmt WLD and the VI WLD VxRail clusters that must be stretched across the AZs as per the VVD design:

- Stretched Layer 2 for the external management traffic.
- Routed L3 for vSAN between data node sites
- 5 millisecond RTT between data node sites
- Layer 3 vSAN between each data nodes site and Witness site
- 200 millisecond RTT between data node sites and the Witness site
- DHCP is required for the Host TEP networks at AZ1 and AZ2.
- Stretched Layer 2 for Edge TEP and Uplink networks for Edge nodes

Note: For the VI WLD, it might be possible to use a different edge design where the uplink and edge TEP networks do not need to be stretched. Consult with VMware before deciding on the design if not following the VCF guidance.

You cannot stretch a VxRail cluster in the following conditions:

- If a VxRail cluster uses IP Pool for the NSX-T Host Overlay Network TEPs
- If remote vSAN datastores are mounted on any VxRail cluster

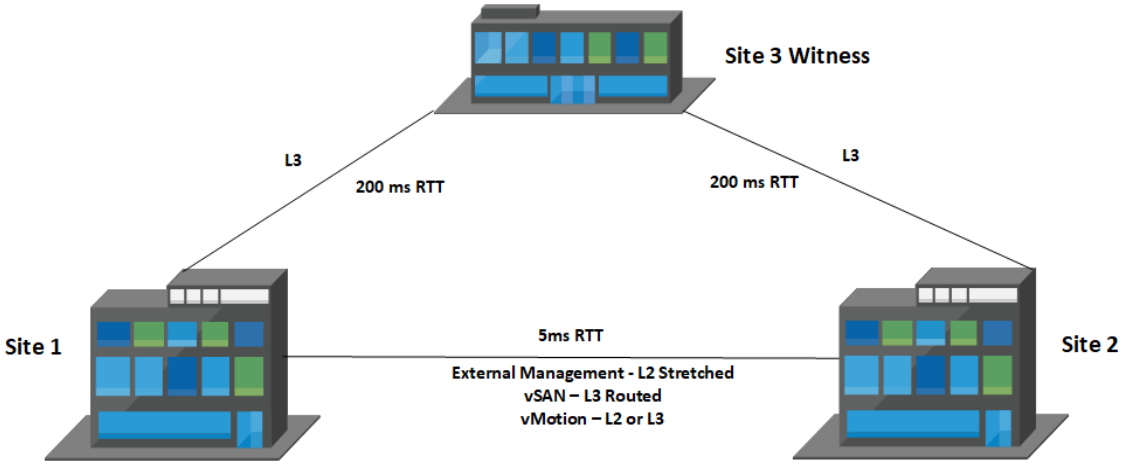

Figure 50. **VxRail vSAN stretched cluster Network Requirements**

The following section contains more detail about the requirements for the network requirements between sites for each type of WLD.

Multi-AZ Connectivity Requirements

The following table shows the supported connectivity for the data nodes sites for the different traffic types between sites.

Table 15. Site Connectivity and MTU

Traffic Type	Connectivity Options	Minimum MTU	Maximum MTU	Default Configuration
External Management	L2 Stretched	1500	9000	1500
vSAN	L3 Routed	1500	9000	1500
vMotion	L3 Routed/ L2 Stretched	1500	9000	1500
Host TEP	L3 Routed	1600	9000	9000
Witness vSAN	L3 Routed to Witness Site	1500	9000	1500
Mgmt WLD- Edge TEP (AVN Enabled)	L2 Stretched	1600	9000	9000
Mgmt WLD - Edge Uplink 01 (AVN Enabled)	L2 Stretched	1500	9000	9000
Mgmt WLD - Edge Uplink 02 (AVN Enabled)	L2 Stretched	1500	9000	9000
VI WLD -Edge TEP	L2 Stretched	1500	9000	User Input

Traffic Type	Connectivity Options	Minimum MTU	Maximum MTU	Default Configuration
VI WLD - Edge Uplink 01	L2 Stretched	1500	9000	User Input
VI WLD - Edge Uplink 02	L2 Stretched	1500	9000	User Input

Increasing the vSAN traffic MTU to improve performance requires the MTU for the witness traffic to the witness site to also use an MTU of 9000. This might cause an issue if the routed traffic needs to pass through firewalls or use VPNs for site-to-site connectivity. Witness traffic separation is one option to work around this issue, but is not yet officially supported for VCF on VxRail.

Note: Witness Traffic Separation (WTS) is not officially supported but if there is a requirement to use WTS, the configuration can be supported through the RPQ process. The VxRail vSAN stretched cluster with Witness Traffic Separation procedure requires a manual procedure to configure WTS interfaces and creating static routes, it also has an impact to Day 2 node expansion procedure.

The vSAN traffic can only be extended using Layer 3 routed networks between sites. The vMotion traffic can be stretched Layer 2 or extended using Layer 3 routed networks, Layer 3 is recommended. The external management traffic must be stretched Layer 2 only. This ensures that the management VMs do not need re-IP when they are restarted on AZ2 if AZ1 fails. The Geneve overlay network can either use the same or different VLANs for each AZ. The same VLAN can be used at each site non-stretched, or a different VLAN can be used at each site allowing the traffic to route between sites. The management WLD sample VLAN and sample IP subnets are shown in the following table.

Table 16. Mgmt WLD Sample VLAN and IP Subnets

Traffic Type	AZ1	AZ2	Sample VLAN	Sample IP Range
External Management	✓	✓	1611 (stretched)	172.16.11.0/24
VxRail Discovery	✓	✓	3939	N/A
vSAN	✓	×	1612	172.16.12.0/24
vMotion	✓	×	1613	172.16.13.0/24
Host TEP	✓	×	1614	172.16.14.0/24
Edge TEP	✓	✓	2711 (stretched)	172.27.11.0/24
Edge Uplink 01	✓	✓	2712 (stretched)	172.27.12.0/24
Edge Uplink 02	✓	✓	2713 (stretched)	172.27.13.0/24
vSAN	×	✓	1621	172.16.21.0/24
vMotion	×	✓	1622	172.16.22.0/24
Host TEP	×	✓	1623	172.16.23.0/24

The VVD requirements for the VI WLD are the same as the Mgmt WLD. If edge nodes are deployed, the edge TEP and uplink networks must be stretched Layer 2 between sites. However, it might be possible to implement a different design if this does not meet the requirements. VMware must be consulted in the design phase of the project for alternative designs.

Table 17. VI WLD Sample VLAN and IP Subnets

Traffic Type	AZ1	AZ2	Sample VLAN	Sample IP Range
External Management	✓	✓	1631 (stretched)	172.16.31.0/24
VxRail Discovery	✓	✓	3939	N/A
vSAN	✓	×	1632	172.16.32.0/24
vMotion	✓	×	1633	172.16.33.0/24
Host TEP	✓	×	1634	172.16.34.0/24
Edge TEP	✓	✓	2731 (stretched)	172.27.31.0/24
Edge Uplink 01	✓	✓	2732 (stretched)	172.27.32.0/24
Edge Uplink 02	✓	✓	2733 (stretched)	172.27.33.0/24
vSAN	×	✓	1641	172.16.41.0/24
vMotion	×	✓	1642	172.16.42.0/24
Host TEP	×	✓	1643	172.16.43.0/24

The following diagram illustrates the VLAN requirements for the Mgmt and first WLD for a VCF multi-AZ VxRail vSAN stretched cluster deployment.

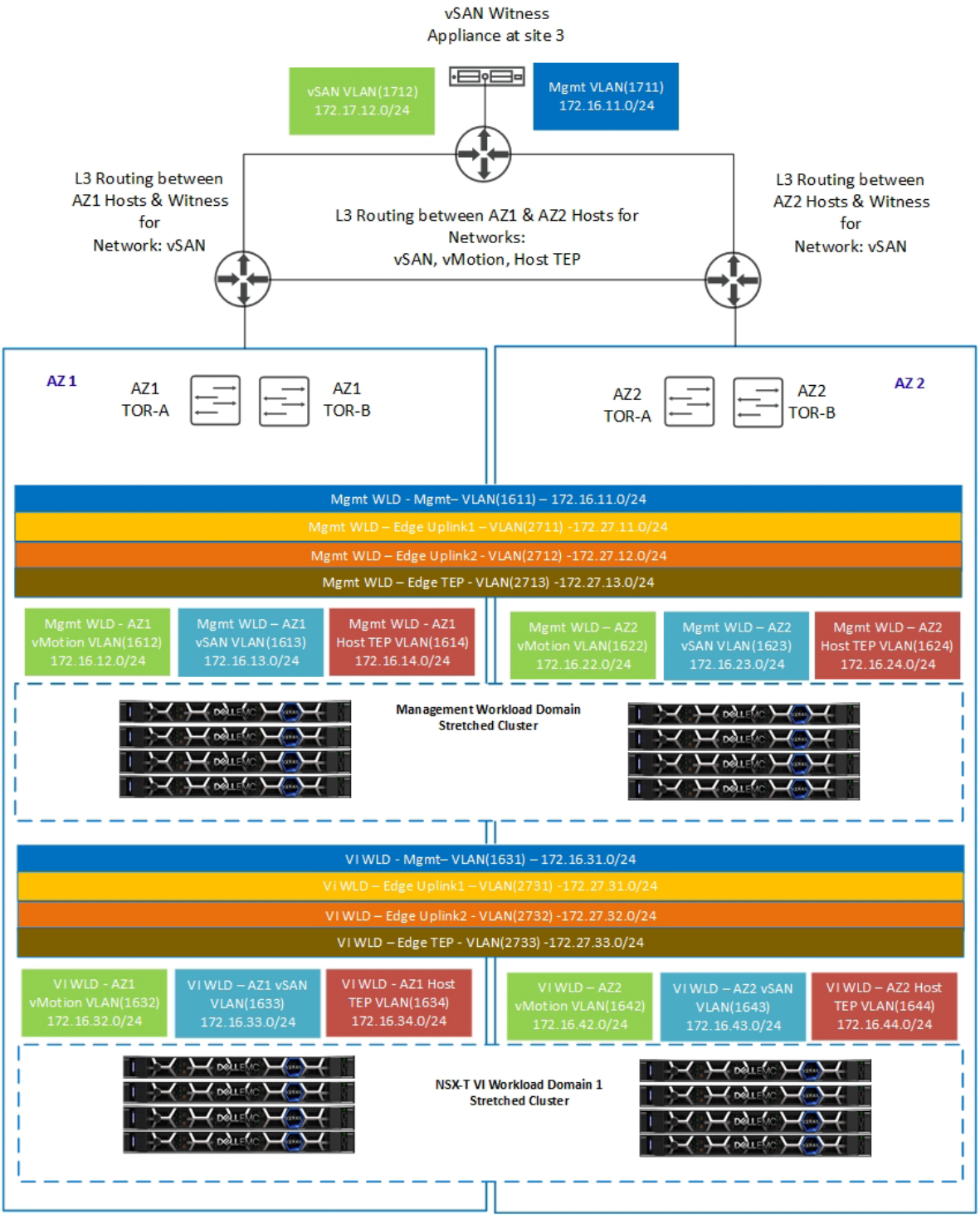

Figure 51. **VLAN and Network requirements for multi-AZ (VxRail vSAN stretched cluster)**

Multi-AZ component placement

During the VxRail vSAN stretched cluster configuration, the management VMs are configured to run on the first AZ by default. This is achieved using Host/VM groups and affinity rules that keep these VMs running on the hosts in AZ1 during normal operation. The following diagram shows where the management and NSX VMs are placed after the stretched configuration is complete for the Mgmt WLD and the first VxRail cluster of an NSX-T VI WLD.

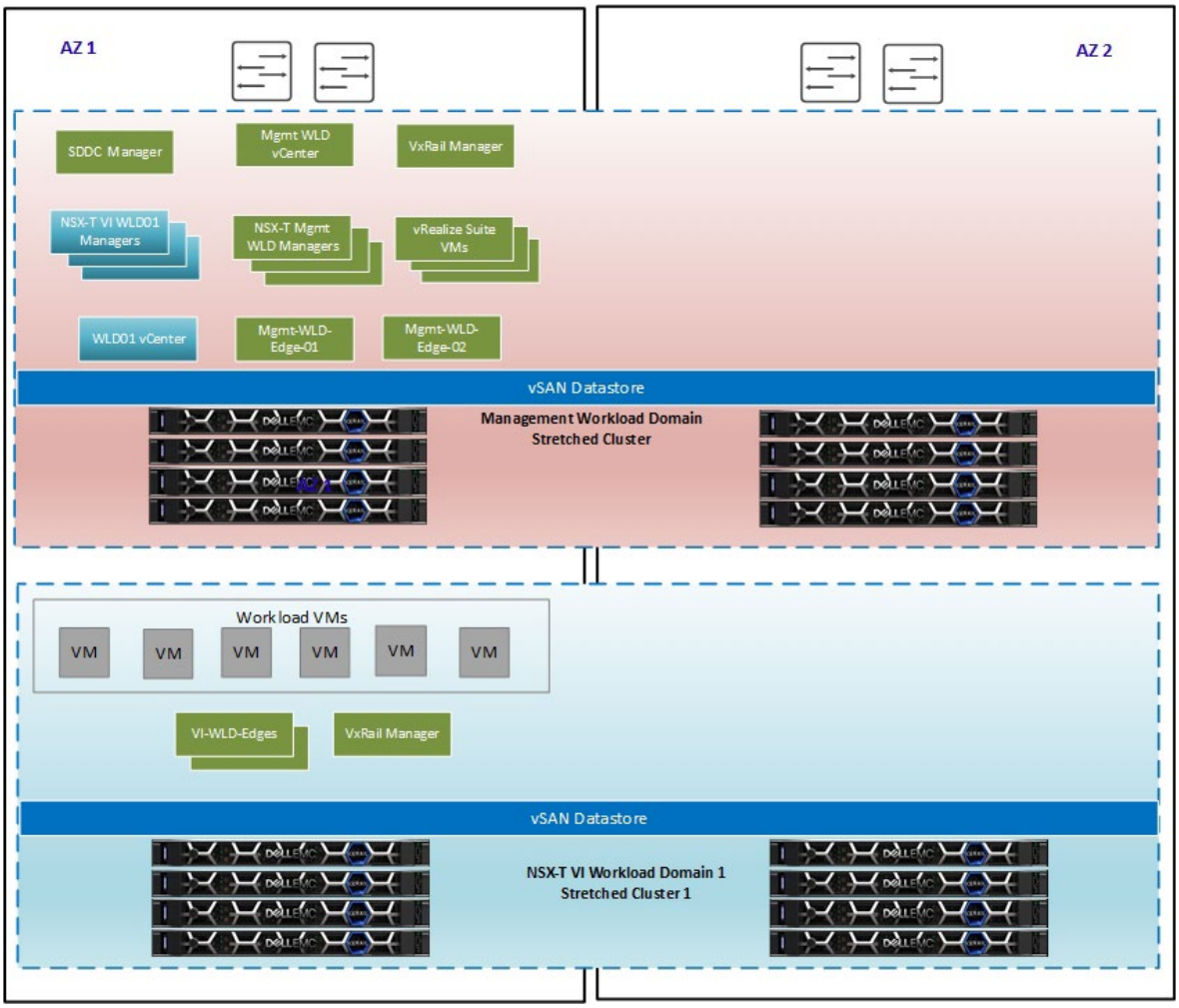

Figure 52. **Multi-AZ Component Layout**

Multi-AZ Supported Topologies

In this next section, we cover some of the different deployment options for a multi-AZ deployment. The management WLD VxRail cluster must always be stretched but the VI WLD VxRail clusters can either be local, stretched, or remote. The VI WLDs can use a shared NSX-T instance (1:Many), or they can use a dedicated NSX-T instance for each VI WLD (1:1). This first diagram shows a standard multi-AZ VxRail vSAN stretched cluster deployment with a stretched Mgmt WLD and one stretched VI WLD with one VxRail cluster and a single NSX-T instance.

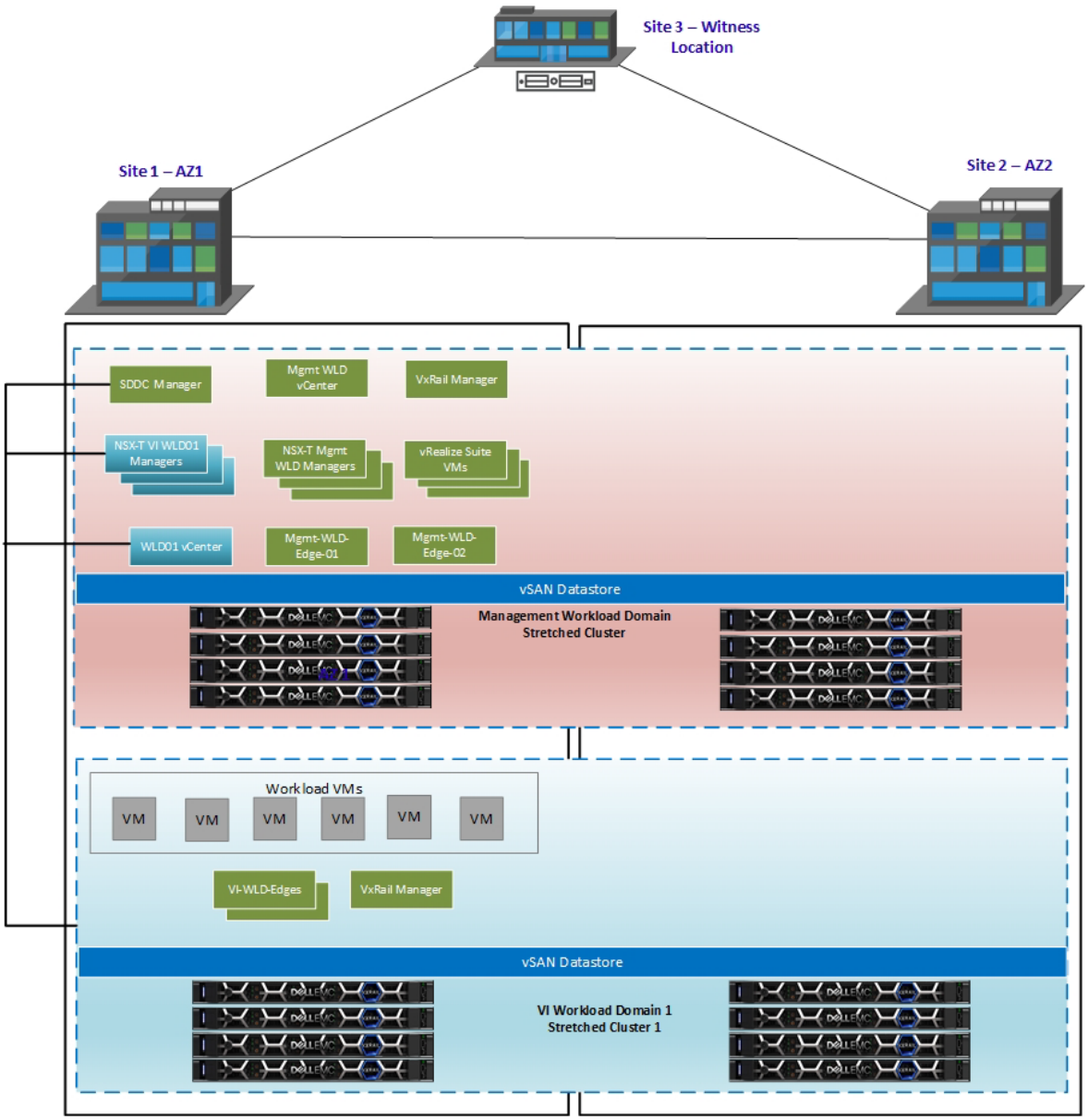

Figure 53. **Mgmt and single VI WLD stretched**

In the next diagram, we have a stretched management WLD and two VI WLDs stretched but using a single NSX-T instance for the two VI WLDs. A single NSX-T edge cluster is used for both VI WLD VxRail clusters.

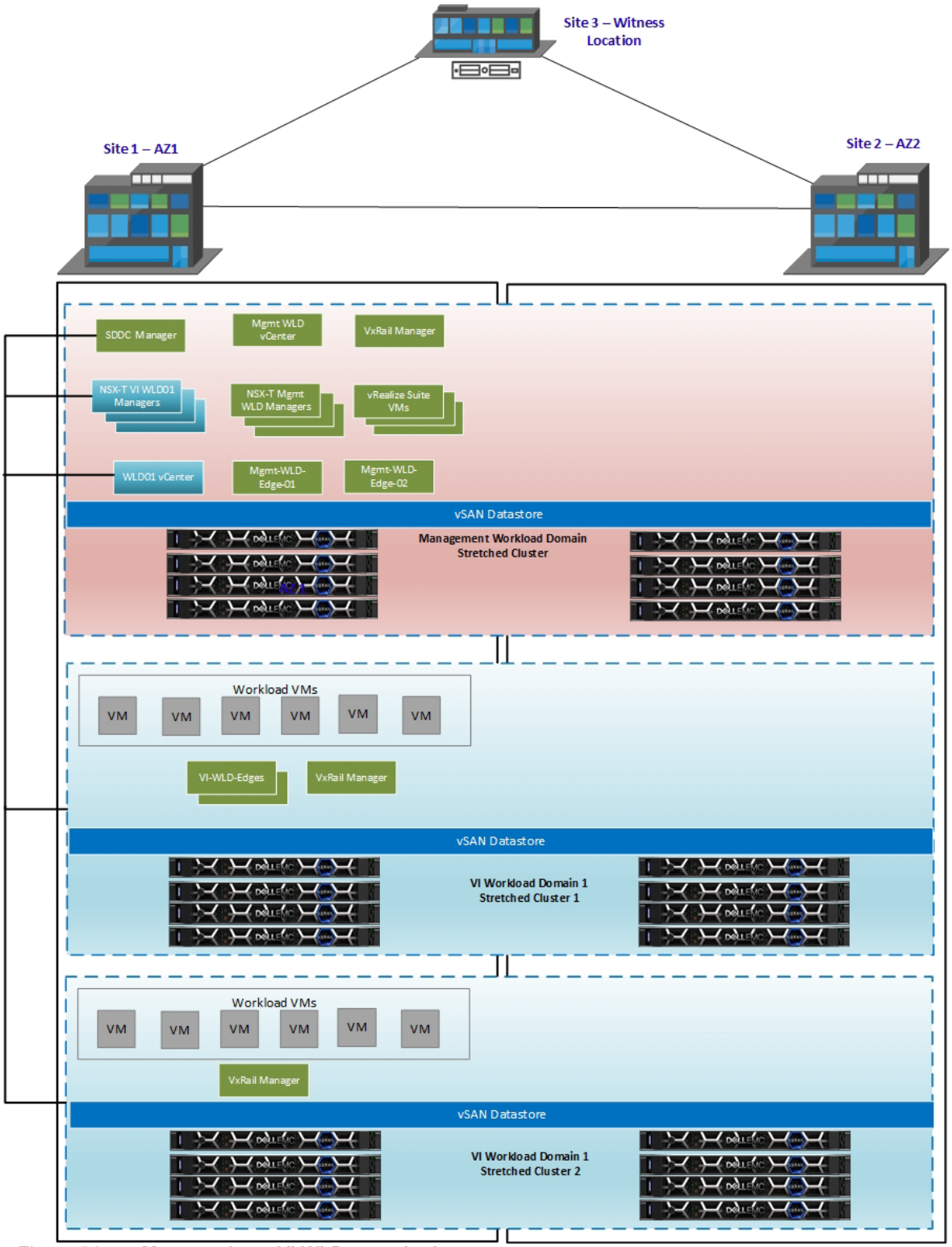

Figure 54. **Mgmt and two VI WLD stretched**

The next diagram illustrates the concept of mixing local and stretched clusters in dual AZ. In this scenario, we have a stretched management WLD and two VI WLDs with a single NSX-T instance.

The first VI WLD has one stretched VxRail cluster and the second VI WLD has two clusters, one is deployed at site 1 and the second VxRail cluster is deployed at site 2. A dedicated NSX-T edge cluster is deployed on the cluster at site 2.

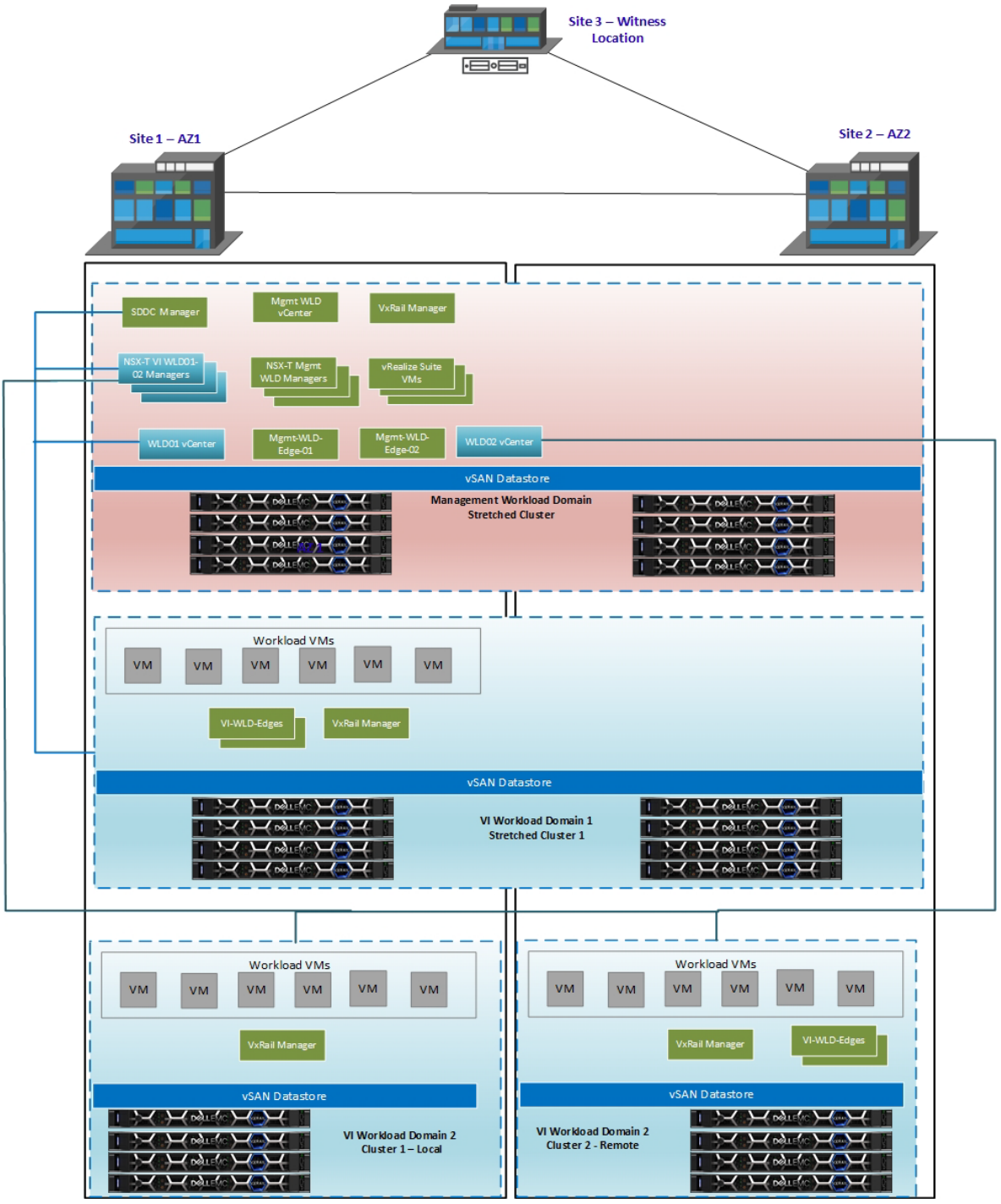

Figure 55. **Mgmt and VI WLD01 stretched, non-stretched in VI WLD02**

The final topology illustrated in the next diagram is similar to the previous design. This time, we have a second NSX-T instance that is deployed to manage the network virtualization for WLD02. This is considered a 1:1 NSX-T design where each WLD has a dedicated NSX-T instance. We also have dedicated edges for both VxRail clusters at each site in WLD02. This prevents traffic hair pinning between sites and keeps traffic local to the site.

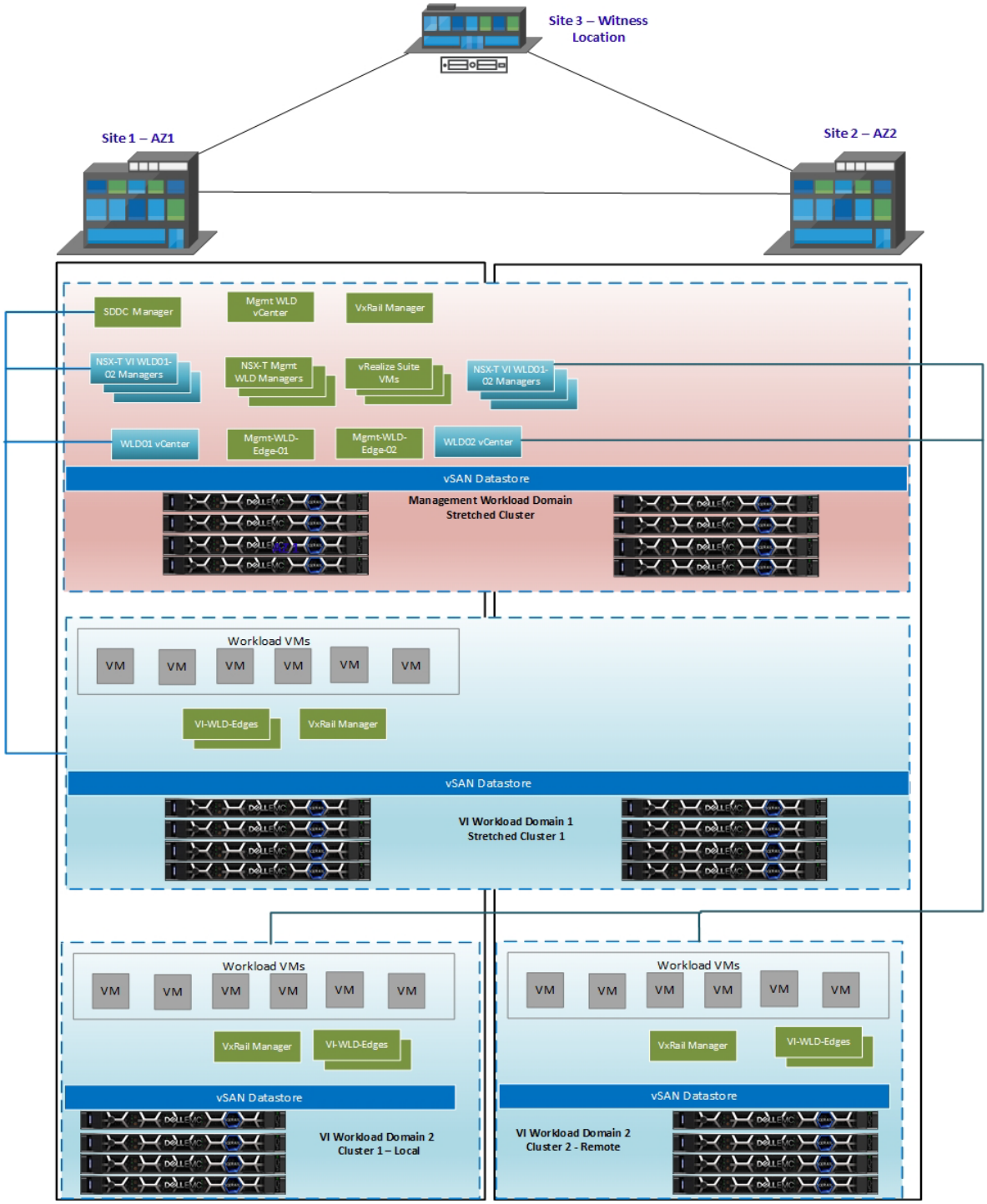

Figure 56. **Mgmt and VI WLD01 stretched, non-stretched VI WLD02 with 1:1 NSX-T**

Management WLD multi-AZ – VxRail vSAN stretched-cluster routing design

As previously mentioned with AVN overlay networks deployed, the Edge nodes are deployed and configured to enable the management components in the vRealize Suite to use this network. In the case of multi-AZ, the North/South routing that occurs through AZ1 would need to failover to AZ2 if there is a full site failure. This is achieved by adding the AZ2 TOR switches as BGP neighbors to the Tier 0 gateway so that traffic from the Tier1 can now flow through the TORs at either site. Using both BGP local preference and Path prepend configured on the Tier 0 gateway to steer the traffic out of AZ1 in normal operating conditions requires manual Day 2 configuration. This configuration is outlined in the VCF documentation NSX-T Data Center Configuration for Availability Zone 2.

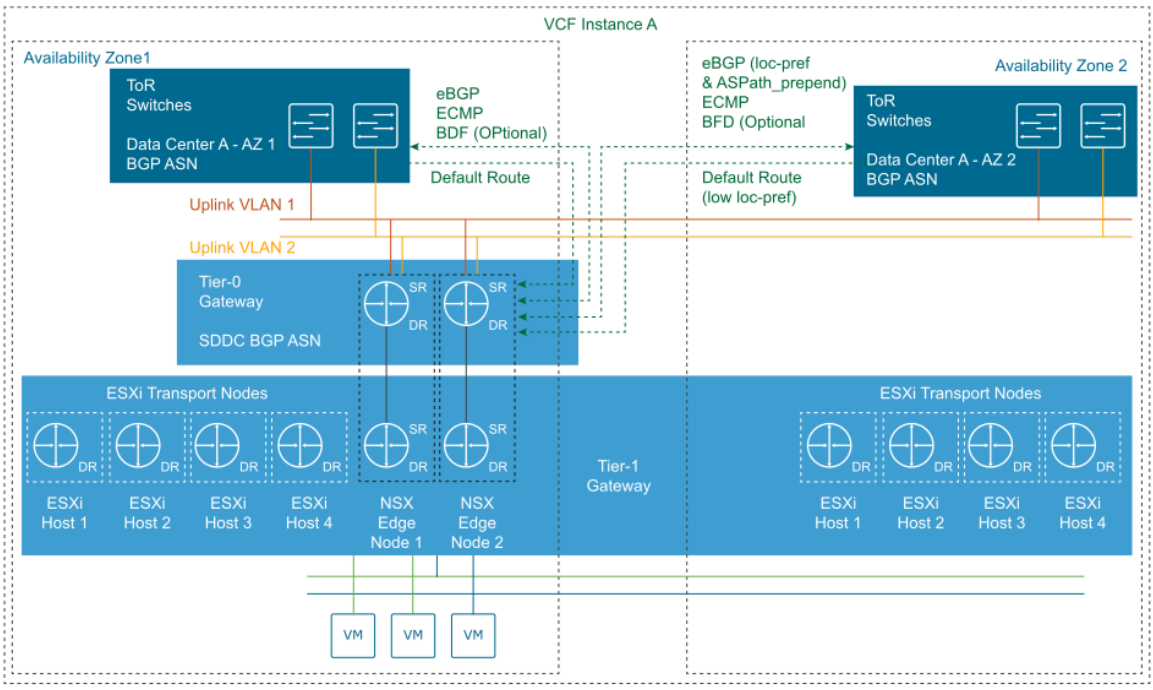

Figure 57. **Multi AZ – Mgmt WLD VCF routing design**

Multisite (Dual Region)

Starting at VCF 4.2 NSX-T federation is now supported which is the foundation to support a multisite dual region deployment. This allows two separate VCF instances in Datacenters at two different locations in large distance regions to be connected to provide centralized management, consistent networking and security policy configuration with enforcement and synchronized operational state. With NSX-T Federation, VCF can leverage stretched networks and unified security policies across multi-region VCF deployments providing workload mobility and simplified disaster recovery. The deployment and configuration are done manually following prescriptive guidance in VMware VVD documentation.

NSX-T Global Manager

The NSX-T Global Manager is part of multi-region deployments where NSX-T Federation is required. NSX-T Global Manager is a central component deployed as a cluster for availability and can connect multiple NSX-T Local Manager instances under a single global management plane. NSX-T Global Manager provides the user interface and the RESTful API for creating, configuring, and monitoring NSX-T global objects, such as global virtual network segments, and global Tier-0 and Tier-1 gateways.

Connected NSX-T Local Manager instances create the global objects on the underlying software-defined network that you define from NSX-T Global Manager. An NSX-T Local Manager instance in an individual region directly communicates with NSX-T Local Manager instances in other regions to synchronize configuration and state needed to implement a global policy.

NSX-T Federation Requirements

Following are some additional requirements that you must consider for an NSX-T federation deployment.

- There must be a maximum round-trip time of 150 milliseconds between the following nodes:
  - Global Manager and Local Manager
  - Local Manager and remote Local Manager
- A Remote Tunnel Endpoint (RTEP) VLAN is required at each site, this is used for intersite communications.
- The management WLD must be sized accordingly to allow for the additional Global Manager clusters that will be deployed if NSX-T federation is implemented.
- The Global Manager and Local Manager appliances must all have NSX-T Data Center 3.1.0 or later installed. All appliances in an NSX-T Federation environment must have the same version installed.
- The required ports must be open to allow communication between the Global Manager and Local Managers. See VMware Ports and Protocols at NSX-T Federation Ports.

Dual Region Component Placement

An NSX-T Global Manager cluster is deployed in the management WLD at region A and region B. The cluster in the second region acts as a standby and becomes active if the first region cluster fails or is lost. A cluster consists of three manager VMs. Each NSX-T domain that needs to be federated require an NSX-T Global manager cluster deployed in the management workload at each region. The following diagram shows a dual region deployment with a single NSX-T VI WLD. A Global manager cluster is deployed at each location for the Mgmt WLD and the VI WLD NSX-T domains.

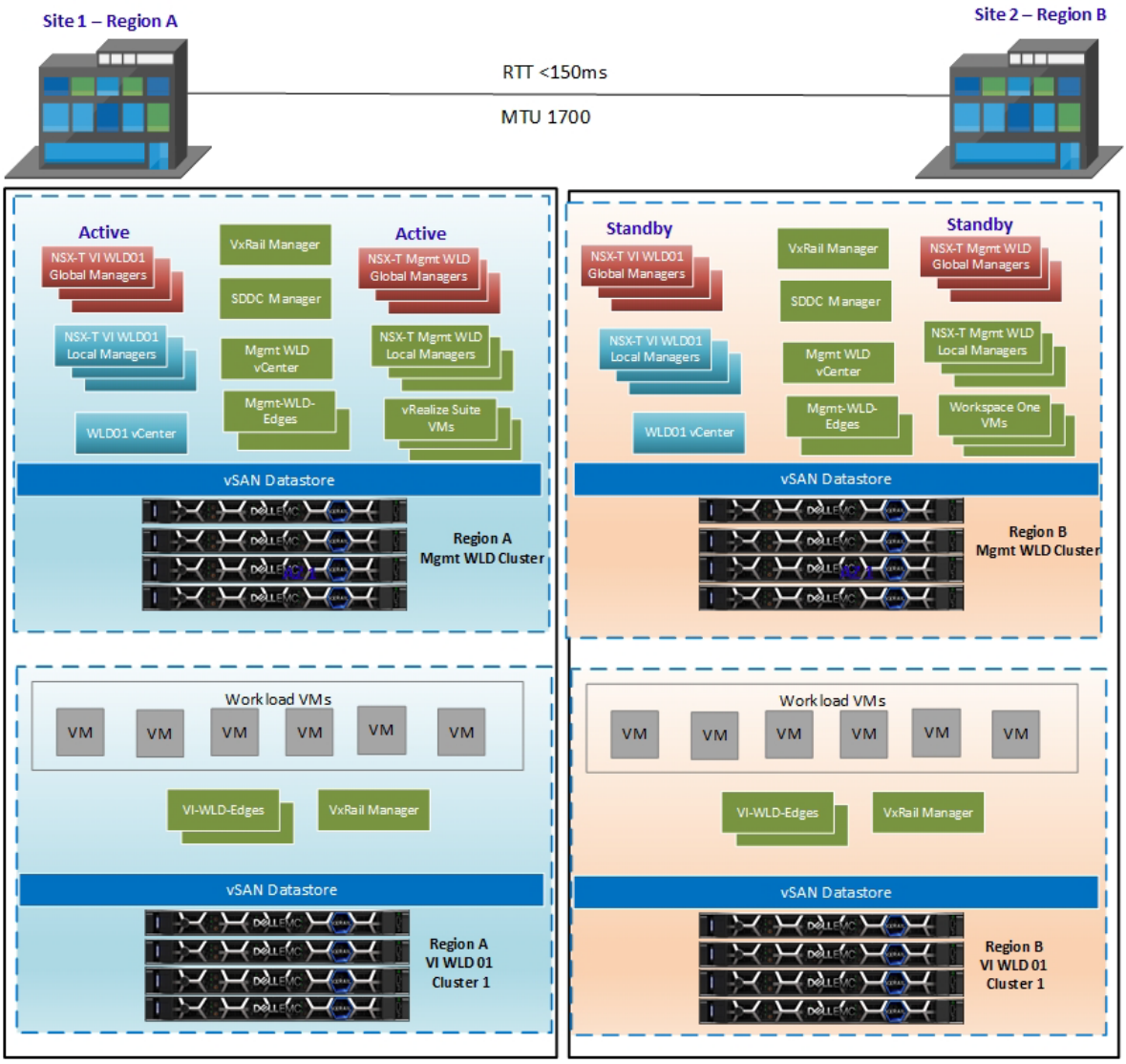

Figure 58. **Multisite – Dual Region NSX-T Global Manager placement**

Inter-region connectivity

In a dual region deployment, each region has its own NSX-T Edge cluster. In each region, the edge nodes and clusters are deployed with the same design but with region-specific settings such as IP addressing, VLAN IDs, and names. Each edge cluster is managed by the NSX-T Local Manager instance for that region and WLD. After a VCF deployment of the Mgmt WLD, all NSX-T network components will be local to the Mgmt WLD NSX-T instance. As part of the NSX-T federation deployment, the network components are configured to span both regions. For more detail for the deployment of NSX-T federation, see the VCF documentation Deploy NSX-T Federation for the Management Domain in the Dual-Region SDDC.

Region-to-region workload traffic traverses the inter-region overlay tunnels which terminate on the RTEPs on the NSX-T Edge nodes. To support this inter-region communication, you must provision additional RTEP VLANs for the edge nodes. If the region also contains multiple availability zones, this network must be stretched across all availability zones in Region A.

Multiregion routing design

The VVD routing design uses region preference for North/South traffic and does not use local-egress. All segments have a preferred and failover region for network traffic ingress and egress for that segment. This eliminates the complexities of trying to prevent asymmetrical routing, and control of local-ingress at the physical network layer. For full detail of the North/South routing design, see the VVD documentation NSX-T Routing for a Multi-Region SDDC for the Management Domain.

LCM Considerations

The NSX-T global managers are deployed manually outside of VCF. The life cycle of these components needs to be done outside of SDDC Manager as SDDC Manager has no awareness of the Global managers. The upgrade of the NSX-T Global managers must be done using the upgrade coordinator available on the Global manager appliance. The following should be considered when planning an upgrade of VCF when NSX-T federation has been deployed.

- Before the upgrade of any WLD, the impact of any version upgrades should be evaluated regarding the need to upgrade NSX-T Global Manager.
- Use NSX-T Upgrade Coordinator to perform life cycle management on the NSX-T Global Manager appliances.
- Before the upgrade of the NSX-T Global Manager, the impact of any version change should be evaluated against the existing NSX-T Local Manager nodes and WLDs.

Multi VCF instance SSO considerations

With VxRail version 4.7.300 or later, you can join an existing SSO domain during first run. This allows for two VCF on VxRail management WLDs to join the same SSO domain. This must be configured during the deployment of the second VCF instance. This allows for a single-pane-of-glass view for the management and WLD vCenters at each site. Following are important factors to consider:

- The vCenters for each VCF instance that participates in the same SSO Domain are connected using Enhanced Link Mode (ELM).
- Maximum number of WLDs is reduced by half.
- Total of 15 WLDs shared across the 2 VCF instances. This limitation is due to the maximum number of vCenters that can be connected with ELM.
- Replication configuration should be in a closed-loop design.
- Manual configuration is required to point Site2 back to Site1.

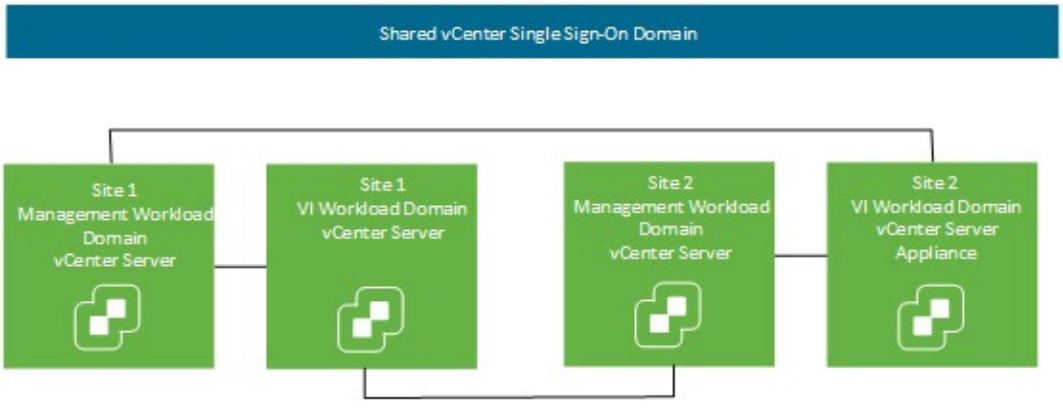

Figure 59. **Shared SSO Domain topology for two VCF Instances**

Future upgrade considerations

There are some factors that must be considered when it comes to upgrading a VCF multi-instance shared SSO domain deployment. The system administrator must use caution when upgrading VCF instances that are part of the same SSO. The following guidelines must be considered before an upgrade of the VCF instances:

- Keep all VCF instances in the same SSO at the same VCF on VxRail version.
- Upgrades should be performed on each VCF on VxRail system in sequential order.
- Ensure that all VCF instances in the same SSO are at N or N-1 versions.
- Do not upgrade a VCF instance that would result in having a participating VCF instance at an N-2 version.
- The compatibility rules in VCF LCM do not extend to external VCF instances.

There are no safeguards that would prevent you from upgrading one VCF instance that would break compatibility between the components participating in the shared SSO domain.

Chapter 11 Operations Management Architecture

This chapter presents the following topics:

Introduction

For the VCF on VxRail solution, there are several different components that can be deployed to support centralized monitoring and logging of the solutions within the SDDC. The vRealize Lifecycle Manager VM is deployed from SDDC Manager and used to deploy the vRealize Suite of components. They are described in more detail in this section.

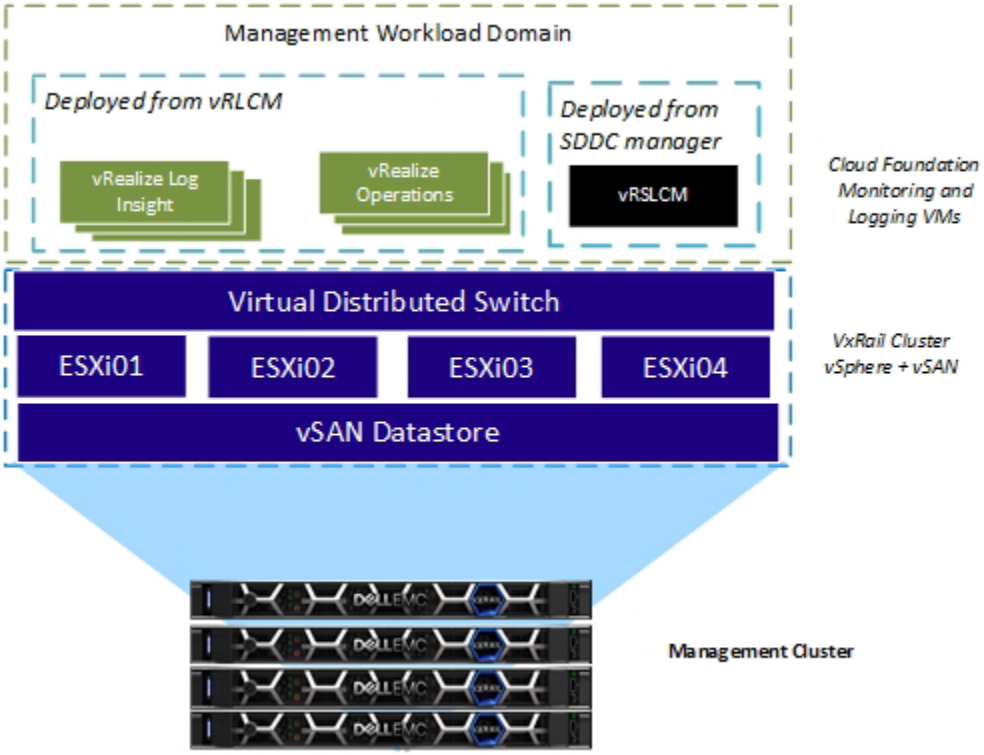

Figure 60. **Monitoring and Logging Operations**

VxRail vCenter UI

The VxRail vCenter HTML 5 plug-in provides a rich set of features to monitor the health of the logical and physical components of the VxRail cluster. A link-and-launch feature provides a dashboard to view the physical layout of each VxRail appliance and displays the status of the physical hardware components. The VxRail Manager is fully integrated with the vCenter Events and Alarms. An underlying VxRail issue is raised as an event or an alarm to inform the user of such an issue.

Intelligent Logging an Analytics

The Intelligent Logging and Analytics for VMware Cloud Foundation-validated solution provides information on the use of a log analysis tool that delivers highly scalable log management with intuitive and actionable dashboards, sophisticated analytics, and broad third-party extensibility. The solution provides deep operational visibility and fast troubleshooting across physical, virtual, and cloud environments. For more detail about the design for logging with vRealize Log Insight as the core component, see Detailed Design of Intelligent Logging and Analytics for VMware Cloud Foundation The deployment of vRealize Log Insight must be done through vRealize Lifecycle

Manager following the VVS deployment guideline. See Implementation of Intelligent Logging and Analytics for VMware Cloud Foundation.

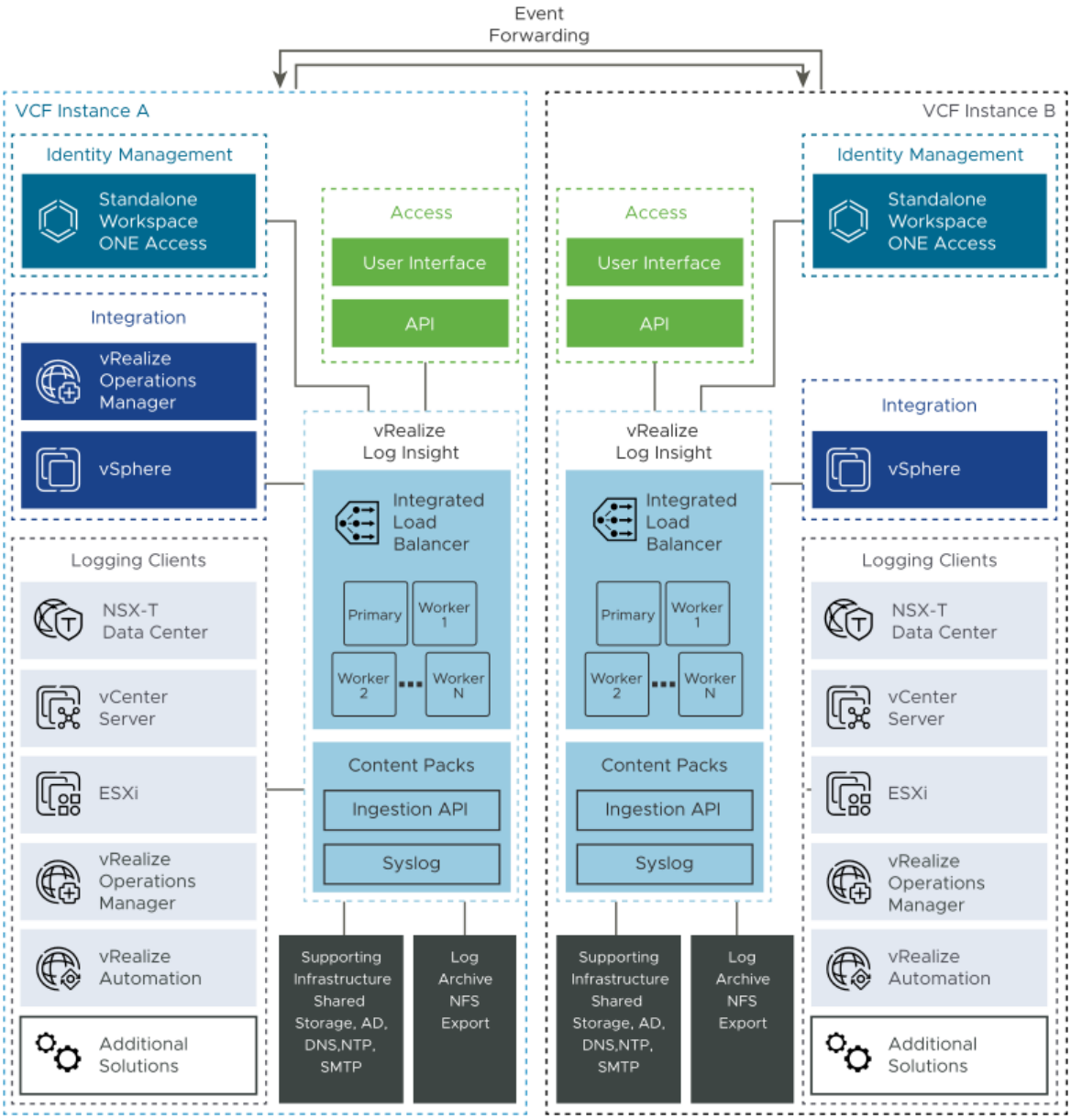

Figure 61. **Intelligent Logging and Analytics for VCF**

Intelligent Operations Management

The *Intelligent Operations Management for VMware Cloud Foundation* validated solution provides a centralized monitoring and alerting to the solution. It provides the virtual infrastructure or cloud admin through a single interface to review and act on events and alerts allowing them to deliver proactive management of system failures. For additional details see the following documentation: Intelligent Operations Management for VMware Cloud Foundation.

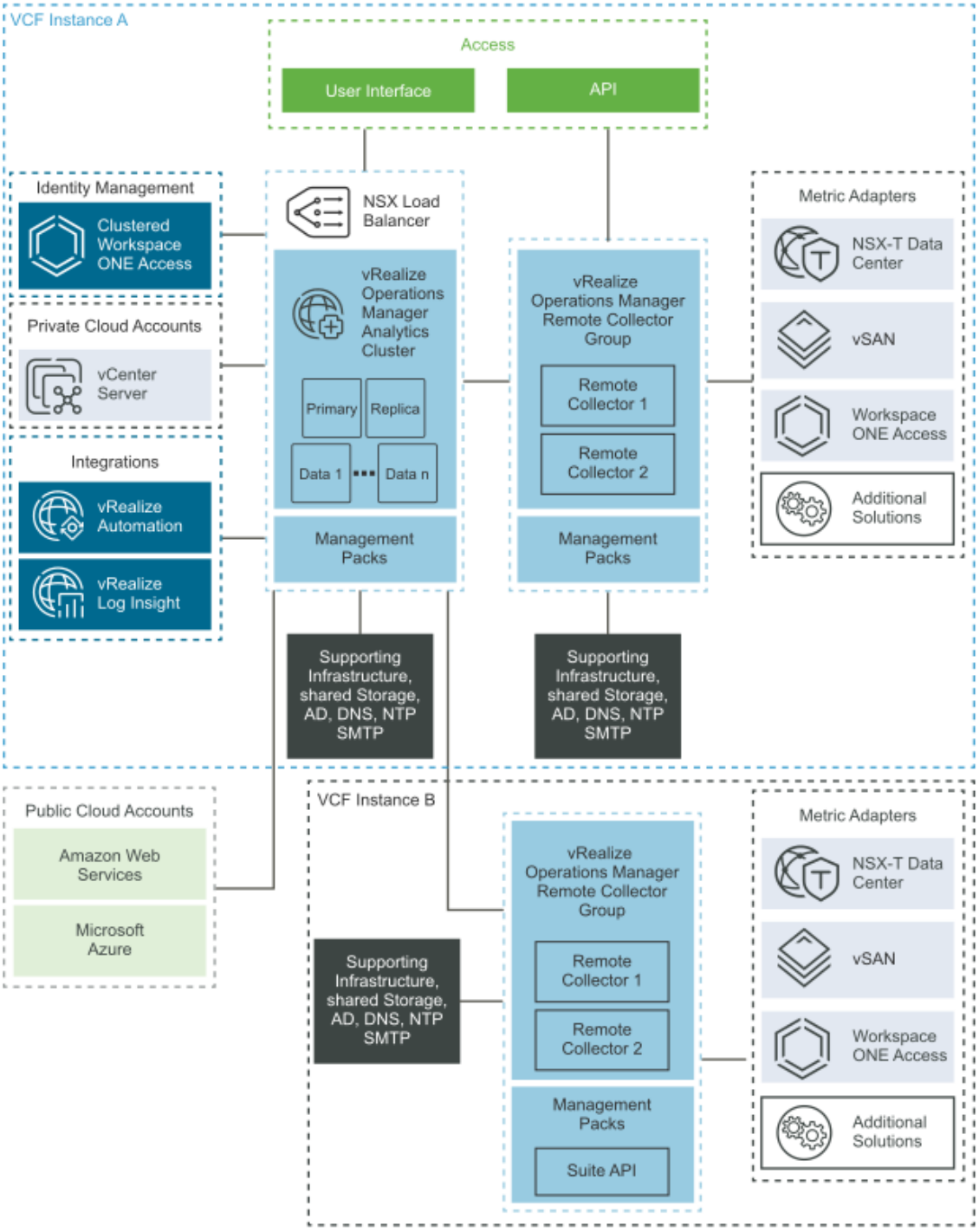

Figure 62. **Intelligent Operations Management for VMware Cloud Foundation**

The deployment of vRealize Operations must be done through vRealize Lifecycle Manager following the VVS deployment guideline. See Implementation of Intelligent Operations Management for VMware Cloud Foundation.

Chapter 12 Lifecycle Management

This chapter presents the following topics:

Introduction

One of the major benefits of VCF on VxRail is the complete end-to-end life cycle of the entire hardware and software stack. This makes operating the data center fundamentally simpler by bringing the ease-of-built in life-cycle automation for the entire cloud infrastructure stack including hardware. The SDDC Manager orchestrates the end-to-end life-cycle process and is fully integrated with VxRail Manager for each VxRail cluster. The VxRail hardware and software life cycles are orchestrated from the SDDC Manager. The underlying hardware, firmware, vSphere ESXi, and vSAN upgrade process for each VxRail cluster is managed by VxRail Manager.

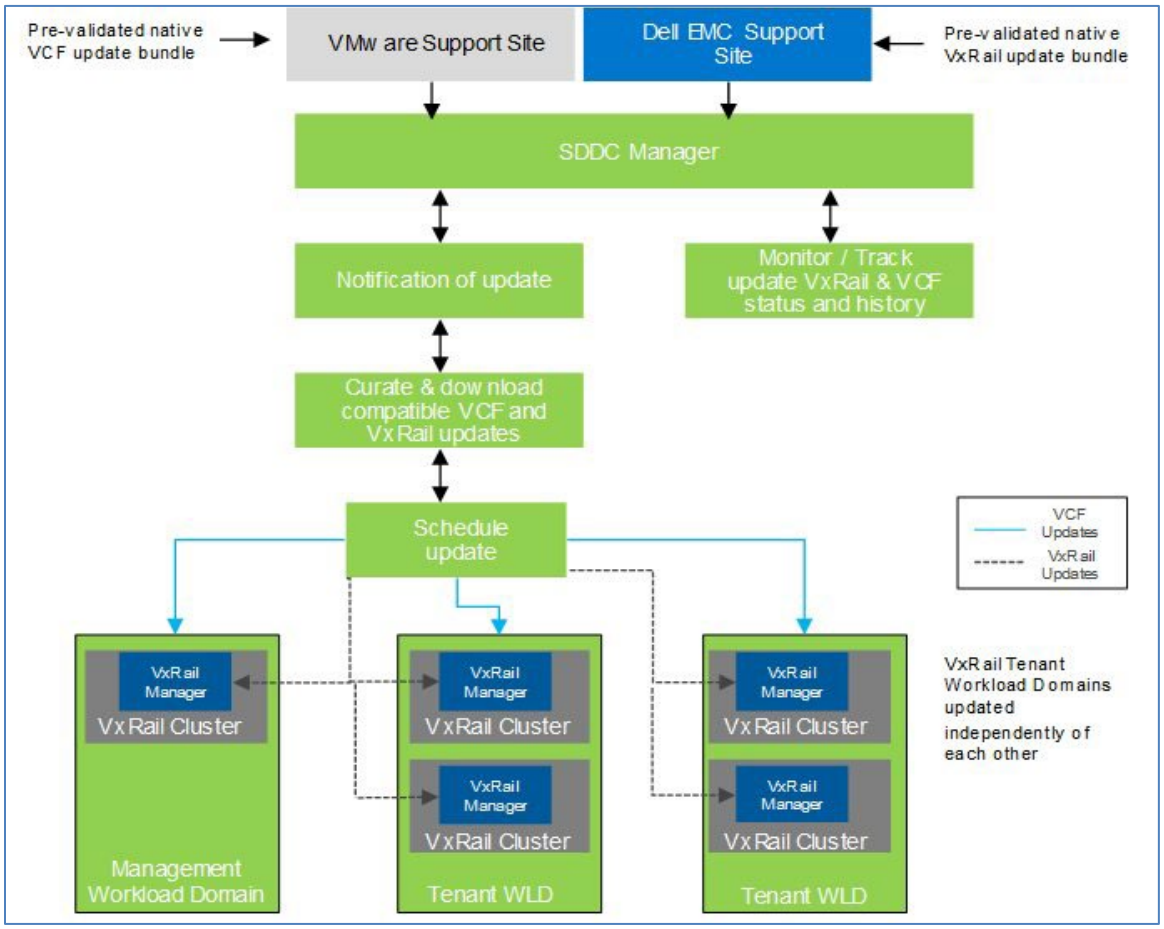

Figure 63. **VCF on VxRail LCM Components**

Credentials for a My VMware account **and** a Dell Support account must be provided for the LCM process to download the appropriate upgrade bundles. VMware and Dell Technologies validate updates and distribute them using native VCF and Dell VxRail upgrade bundles. Upon notification of the available update, the upgrade bundle must be manually downloaded and staged to SDDC Manager before starting the upgrade.

Note: The Mgmt WLD must be upgraded first. Upgrades cannot be applied to VxRail VI WLD before they are applied to the Mgmt WLD.

vRealize Suite Lifecycle Manager

The VMware vRealize Suite Lifecycle Manager automates the LCM of the vRealize Suite. It must be deployed before any vRealize Log Insight, vRealize Operations, or vRealize Automation components can be deployed. The vRealize Suite Lifecycle Manager contains the functional elements that collaborate to orchestrate the LCM operations of the vRealize Suite environment. The vRLCM bundle must be downloaded using SDDC Manager from the VCF bundle repository. Once the bundle is downloaded, the vRLCM can be installed from the SDDC Manager vRealize Suite tab. If AVN was enabled, the vRLCM VM is deployed onto the xRegion NSX-T segment. If AVN was not enabled, the vRLCM VM must be deployed onto VLAN backed network using the procedure in the following VMware KB article: https://kb.vmware.com/s/article/80864.

Chapter 13 Cloud Management Architecture

This chapter presents the following topic:

Private Cloud Automation for VMware Cloud Foundation

The Private Cloud Automation for VCF-validated solution provides information on the use of vRealize Automation for cloud automation services with the VMware Cloud Foundation platform. The solution can be extended to support public cloud automation and covers recommended operational practices and considerations, where applicable. For more details about the design of Private Cloud Automation, see Detailed Design for Private Cloud Automation for VMware Cloud Foundation. For details about the deployment of vRealize Automation following VVS guidance, see Implementation of Private Cloud Automation for VMware Cloud Foundation.

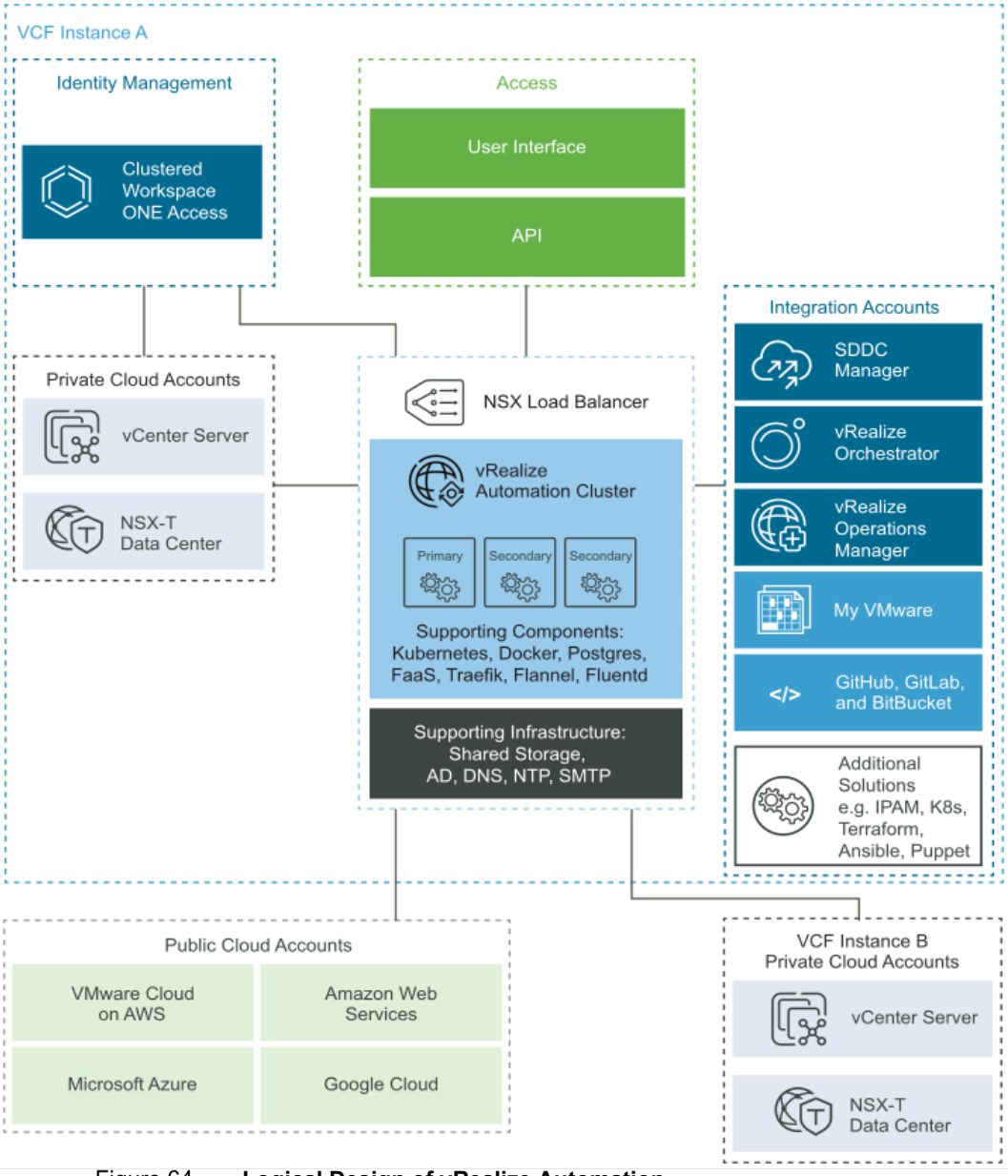

Figure 64. **Logical Design of vRealize Automation**

Before deploying the vRealize Automation, the vRealize Lifecycle Manger must be deployed from SDDC Manager. It is used to deploy and life cycle the vRealize Suite components.