DELL EMC

Architecture Guide

VMware Cloud Foundation on VxRail 3.9.1 Architecture Guide

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.

March 2020

H17731.2

Revisions

Date	Description
April 2019	Initial release
September 2019	Update to support VMware Cloud Foundation 3.8 and NSX-T.
March 2020	Update to support VMware Cloud Foundation 3.9.1 and PKS, DR guidance removed.

Table of contents

1 Executive Summary

1.1 VMware Cloud Foundation on VxRail

VMware Cloud Foundation (VCF) on VxRail is a Dell EMC 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) in 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 that leverages native VxRail hardware and software capabilities and other VxRail-unique integrations (such as vCenter plugins and Dell EMC 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.

1.2 Document purpose

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

1.3 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.

2 Architecture Overview

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 for SDC and SDDC Manager for SDDC LCM.

By virtualizing all of 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 of the SDDC SW and HW components of the stack. These processes are now streamlined using a common management toolset in SDDC Manager in conjunction 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 micro-segmentation, 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.

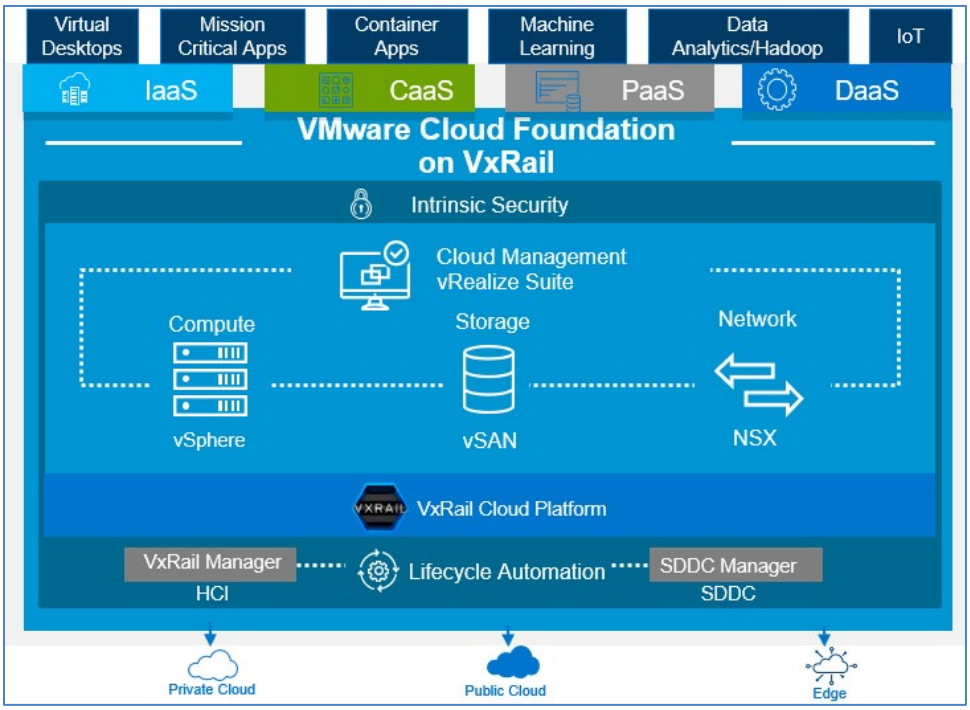

Figure 1 **Architecture Overview**

2.1 VxRail Manager

VCF on VxRail uses VxRail Manager to deploy and configure vSphere clusters that are powered by vSAN. It is also used to execute 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. By 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.

2.2 SDDC Manager

SDDC Manager orchestrates the deployment, configuration, and (LCM) of vCenter, NSX, and vRealize Suite 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 stretched cluster configuration for a dual-availability zone (AZ) WLD.

2.3 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. Users can create multiple virtual networks with diverse requirements, leveraging a combination of the services that are offered by NSX. These services include micro-segmentation 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.

2.4 vRealize operations

The vRealize Operation management components allow centralized monitoring of and logging data about the other solutions in the SDDC. The physical infrastructure, virtual infrastructure, and tenant workloads are monitored in real-time, collecting information for intelligent and dynamic operational management.

2.5 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.

2.6 Cloud management

The Cloud Management platform (CMP) is the main consumption portal for the SDDC. You use vRealize Automation to author, administer, and consume VM 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 execute requests to instantiate specific services.

3 Workload Domain Architecture

A WLD consists of one or more Dell EMC 14G 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 workloads 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 four 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
- A Horizon WLD
- A PKS WLD

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

3.1 Management workload domain

The VCF Management WLD cluster requires a minimum of four hosts on which the infrastructure components used to instantiate and manage the private cloud infrastructure run. For VCF on VxRail, the Management WLD cannot be used to host business workloads. This Management WLD is created during initial system install (or bring-up) using the VCF Cloud Builder tool.

In the Management WLD cluster, vSphere runs with a dedicated vCenter server and a pair of PSCs in the same SSO domain. This cluster, backed by vSAN storage, hosts the SDDC Manager and VxRail Manager VMs, NSX-V, and vRealize Log Insight for Management domain logging. Other components such as vRealize Operations and vRealize Automation are optional. If a Horizon WLD is deployed, the management components will also be deployed in the Mgmt WLD. Because the management cluster contains critical infrastructure, consider implementing a basic level of hardware redundancy for this cluster. The management cluster must have a minimum of four hosts to provide vSAN FTT=1 during maintenance operations.

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

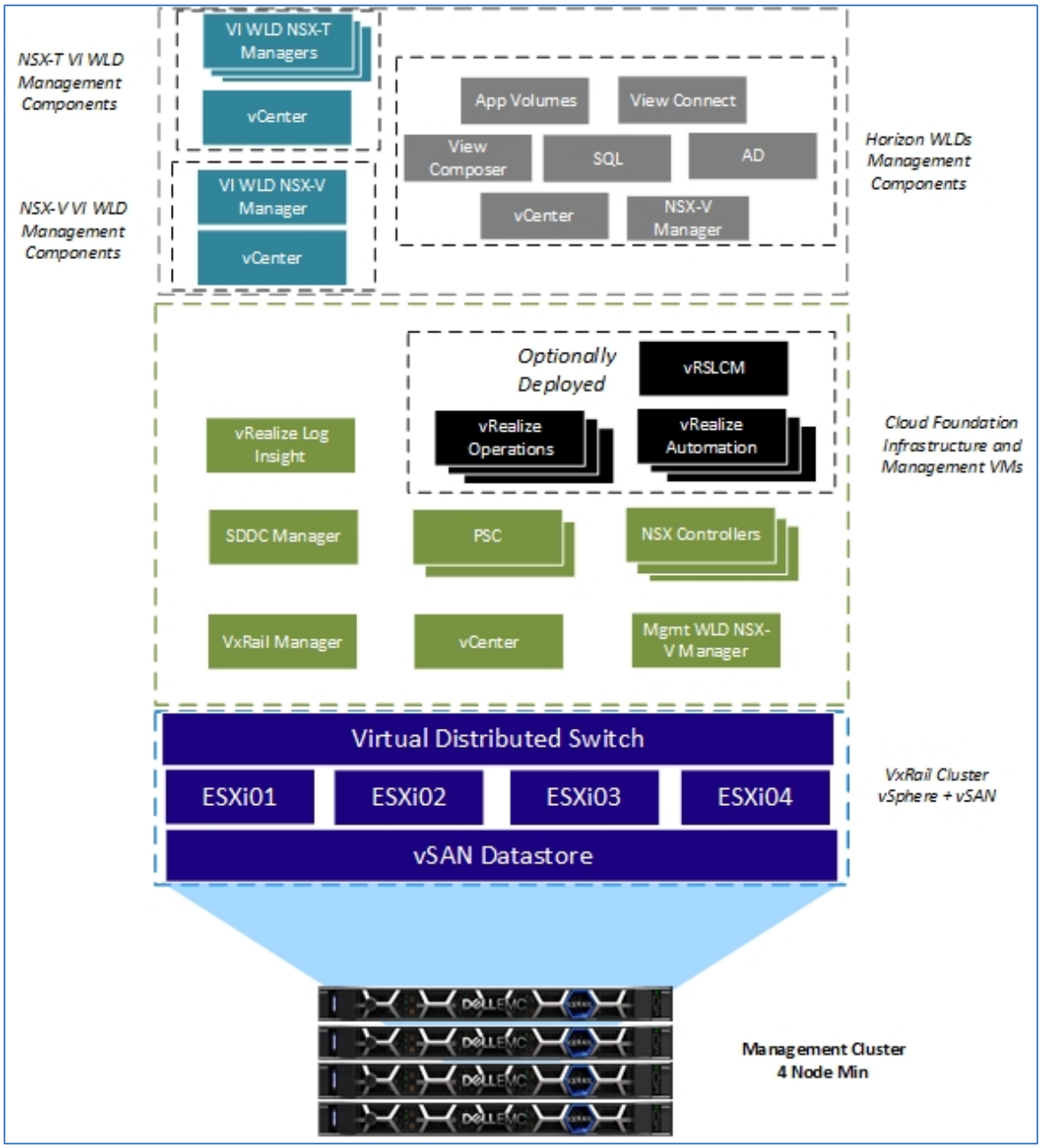

Figure 2 **Management Domain Components**

3.1.1 vCenter design

The management domain vCenter is deployed with an external PSC by using the embedded VxRail cluster deployment process. This vCenter and PSC is then configured as an external vCenter and PSC from the vCenter UI. 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 and PSC components in the solution.

The SDDC Manager deploys the second PSC during the VCF Cloud Builder bring-up process. A replication is established with the first PSC that was deployed when the VxRail cluster was first deployed.

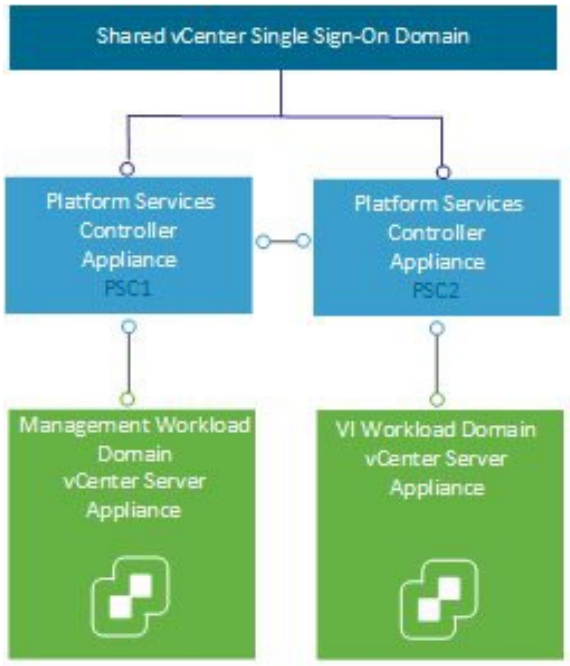

Figure 3 **vCenter Design**

3.2 VxRail VI workload domain

VxRail VI WLD can be either Virtual Infrastructure (VI) WLDs, Horizon domains, or PKS WLDs.

3.2.1 VI WLD

The VI WLD can consist of one or more VxRail clusters. The VxRail cluster is the building block for the VxRail VI WLD. The first cluster of each VI WLD must have four hosts, but subsequent clusters can start with three hosts. The VI WLD can be either an NSX-V based WLD or an NSX-T based WLD. This can be selected when adding the first cluster to the WLD. The vCenter and NSX-V or NSX-T Manager for each VI WLD are deployed into the Mgmt WLD. For an NSX-V based VI WLD the controllers are deployed to the first cluster in the VI WLD added by the SDDC Manager. Each new VI WLD requires an NSX-V Manager to be deployed in the Mgmt WLD and the three controllers deployed into the first cluster of the VI WLD.

For NSX-T based VI WLD, when the first 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 do not require additional NSX-T mangers but each VI WLD VI is added as a compute manager to NSX-T.

For both NSX-T based VI WLD and NSV-V based VI WLD, the first cluster can be considered a compute-and-edge cluster as it contains both NSX and compute components. NSX virtual routers can be deployed to this first cluster. The second and subsequent clusters in a VI WLD can be considered compute-only clusters as they do not need to host any NSX routing virtual machines.

Note: NSX-T is not yet supported for the Mgmt WLD. Only NSX-V is deployed during the automated deployment of the Management domain using Cloud Builder.

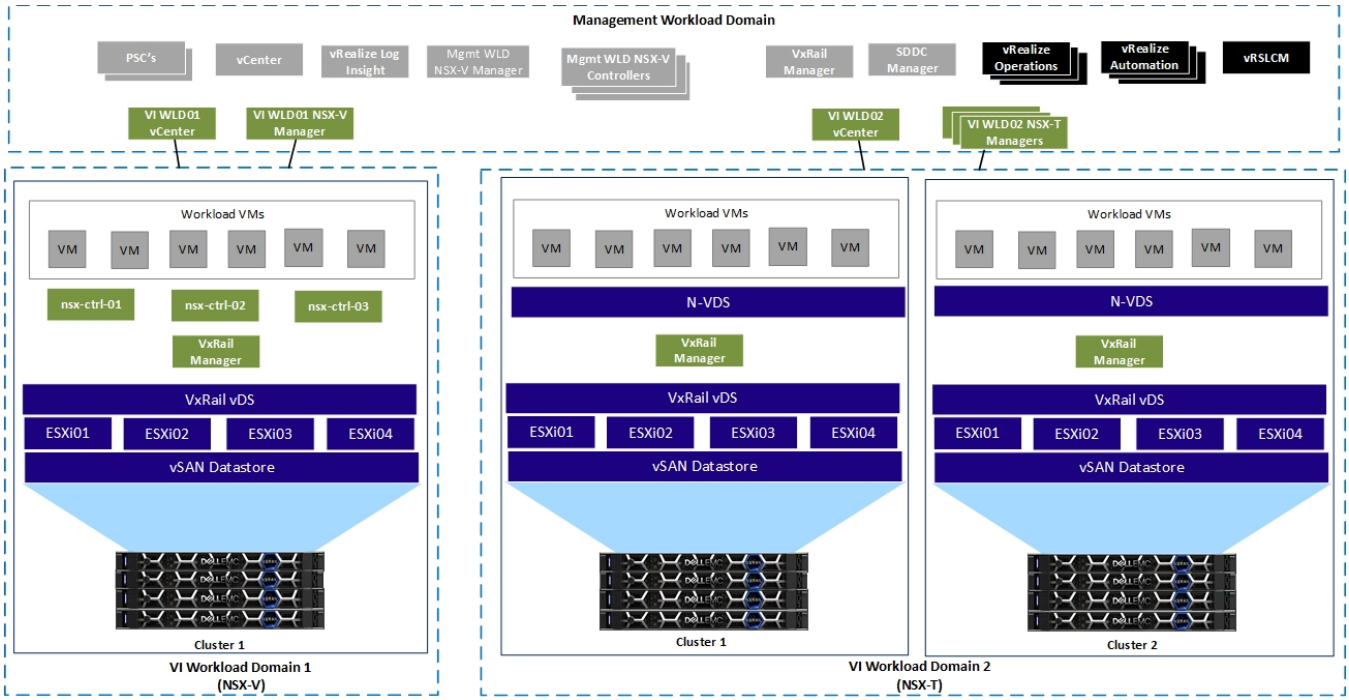

Figure 4 **VI WLD Component Layout with NSX-T and NSX-V VI WLDs**

3.2.1.1 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 Figure 4. 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.

3.2.2 Horizon VDI domain

A Horizon domain automates deployment of VMware Horizon components and supporting infrastructure to enable you to deliver virtual desktop infrastructure (VDI) and remote desktop session host (RDSH) desktops and applications. These desktops can be delivered as persistent, linked clones, or instant clones. The Horizon domain can include VMware app volumes for dynamic application mounting and User-Environment Manager for a persistent end-user experience.

The Horizon domain consumes one or more VI WLD, but requires additional Horizon desktop management components to be deployed as part of the Horizon workload creation process. The Horizon domain is decoupled from resource provisioning - one or more VI WLD must be created before deploying a Horizon domain. There are several prerequisites that must be completed before deploying a Horizon domain. They are documented in the Prerequisites for a Horizon Domain.

During the Horizon domain deployment, one to three connection servers and a corresponding load balancer is deployed. You can also choose the optional components that you want to deploy:

- Composer Server
- App Volumes
- User Environment Manager
- Unified Access Gateway

The Horizon domain is based on the Horizon reference architecture, which uses Pod Block architecture to enable you to scale as your use cases grow. For more information about the architecture and number of supported virtual machines, see the Horizon 7 Pod and Block section in the VMware Workspace ONE and VMware Horizon 7 Enterprise Edition On-premises Reference Architecture document.

3.3 PKS workload domains

VCF on VxRail provides a way to automate the deployment of a VMware PKS platform within a workload domain. This helps accelerate deployment and access of Kubernetes infrastructure in private, public, and hybrid cloud environments so that application development teams can start deploying container-based applications faster.

The PKS WLD deploys VMware Enterprise PKS which is built on upstream Kubernetes and delivered as an integrated solution. The integration within VCF features automated and centralized life cycle management and operations of the underlying WLD clusters. It includes integrated container networking and network security with NSX-T, is easily scalable, and includes automated deployment and configuration.

3.3.1 PKS workload domain architecture

The PKS WLD architecture is based from VVD-compliant best practices designs. Like the Horizon WLD, a PKS WLD is built on top off pre-provisioned VxRail VI WLD. However, PKS requires the WLD to be NSX-T based due to its tight integration with NSX-T. SDDC Manager fully automates the deployment of PKS components, including PCF Ops Manager, BOSH Director, PKS Control Plane, and optionally, the Harbor Enterprise Container Registry. These management components get deployed on the VxRail NSX-T VI WLD cluster. The NSX-T Edge VMs must be manually deployed and configured before starting the PKS WLD deployment. The following diagram illustrates the component layout for a VCF PKS WLD deployment.

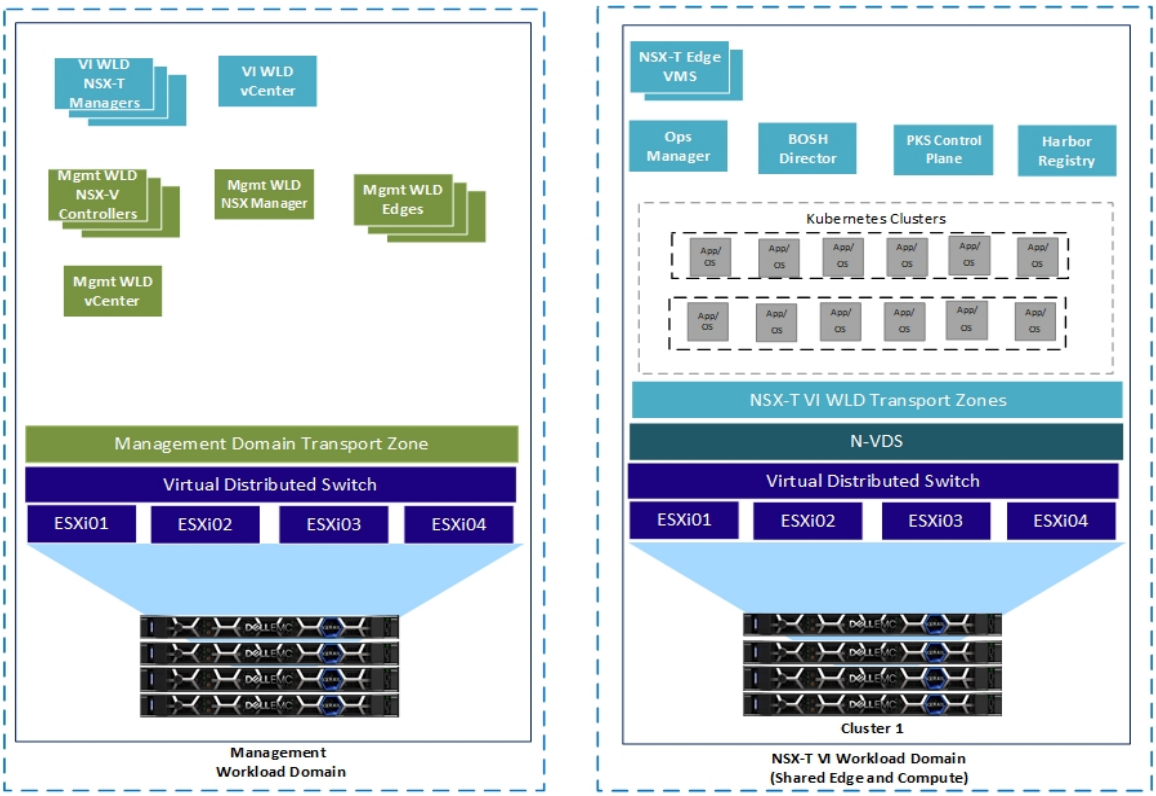

Figure 5 **PKS WLD component layout**

The vCenter and NSX-T Manager components for the VI WLD run on the VCF Management Domain. With the deployment of a PKS WLD, users can map PKS availability zones to Resource Pools into the VI workload domain first cluster.

3.3.2 Prerequisites for Deploying VMware Enterprise PKS

There are several prerequisites that must be met before starting the VCF PKS WLD deployment. An NSX-T VI WLD must first be created within VCF. The NSX-T Edge components must then be manually deployed and configured to allow the PKS components to use the network virtualization features and capabilities that are built into NSX-T, like North–South routing, NAT, and Load Balancing. Following is the list of prerequisites:

1. NSX-T VI WLD
2. NSX-T Edge VMs are deployed and configured.
3. IP/Hostnames for the PKS Management components
4. NSX-T segments created for PKS Management and Service networks
5. IP block for the Kubernetes Pods
6. IP block for the Kubernetes Nodes
7. Floating IP pool for the load balancers for each of Kubernetes clusters
8. CA-signed certificates for Operations Manager, Enterprise PKS control plane, and Harbor Registry
9. Certificate for NSX-T Super User
10. Resource Pools created in the VI WLD vCenter that will be mapped as PKS availability zones

For additional details on the prerequisites, see the VCF Admin guide PKS Prerequisites.

Lifecycle Management of the PKS components is accomplished using the native LCM capabilities that are built into the PKS platform. VxRail VI WLD infrastructure LCM is accomplished using SDDC Manager along with the integration that exists with VxRail Manager.

3.4 Physical workload domain 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. Note that a tenant WLD can consist of one or more clusters, this will be discussed later.

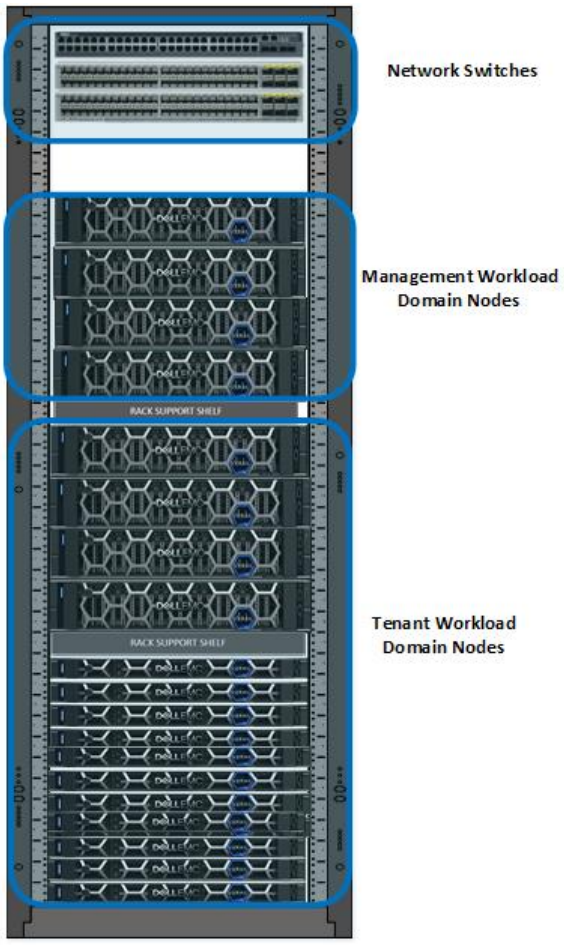

Figure 6 **Single Rack WLD Mapping**

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 Figure 7.

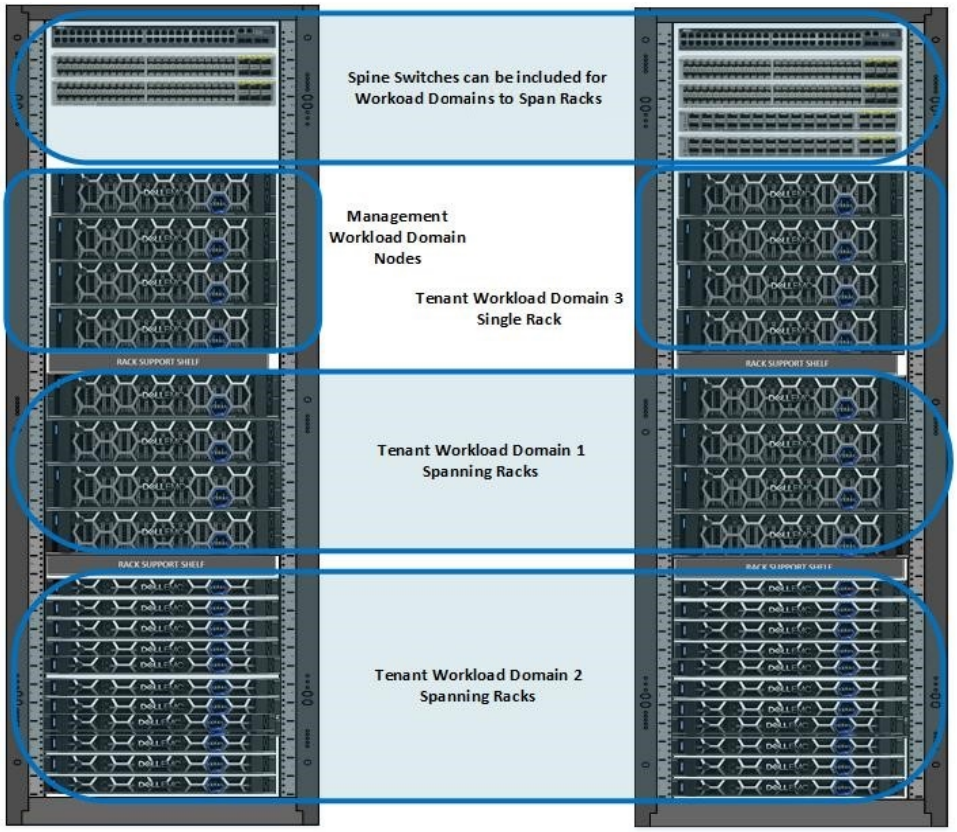

Figure 7 **WLDs Spanning Racks**

3.4.1 VxRail hardware options

Depending on the management workload and the tenant workload and application requirements, the right VxRail hardware platform must be selected. The VxRail HCI family provides the following offerings for all types of workloads.

E Series Nodes	G Series Nodes	P Series Nodes	V Series Nodes	S Series Nodes
Low profile	Compute dense	Performance optimized	VDI optimized	Storage dense
E560/F/N	G560/F/N	P570/F	V570/F	S570
1100 or 1600 W PSU 10 GbE or 25 GbE NVMe cache support	2000 W or 2400 W PSU 10 GbE Optane and NVMe cache Mixed-use SAS cache	1100 W or 1600 W PSU 20 capacity drives 10 GbE or 25 GbE support P580N 1600, 2000, or 2400W PSU 20 capacity drives 10 GbE or 25 GbE NVMe cache support	2000 W PSU Up to 3 GPUs 8 more capacity drives 10 GbE or 25 GbE support	1100 W PSU 10 GbE or 25 GbE support

Table 1. VxRail HCI offerings

See the VxRail sizing tool for guidance on VxRail hardware platforms.

4 VxRail virtual network architecture

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

4.1 Virtual distributed switches

The VxRail is the building block for each cluster, either Mgmt WLD or VxRail VI WLD. The VxRail virtual distributed switch (vDS) provides the virtual network layer for the system network services that are needed for the VMware Cloud Foundation solution and provides the underlying networks for NSX-V based WLDs. 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:

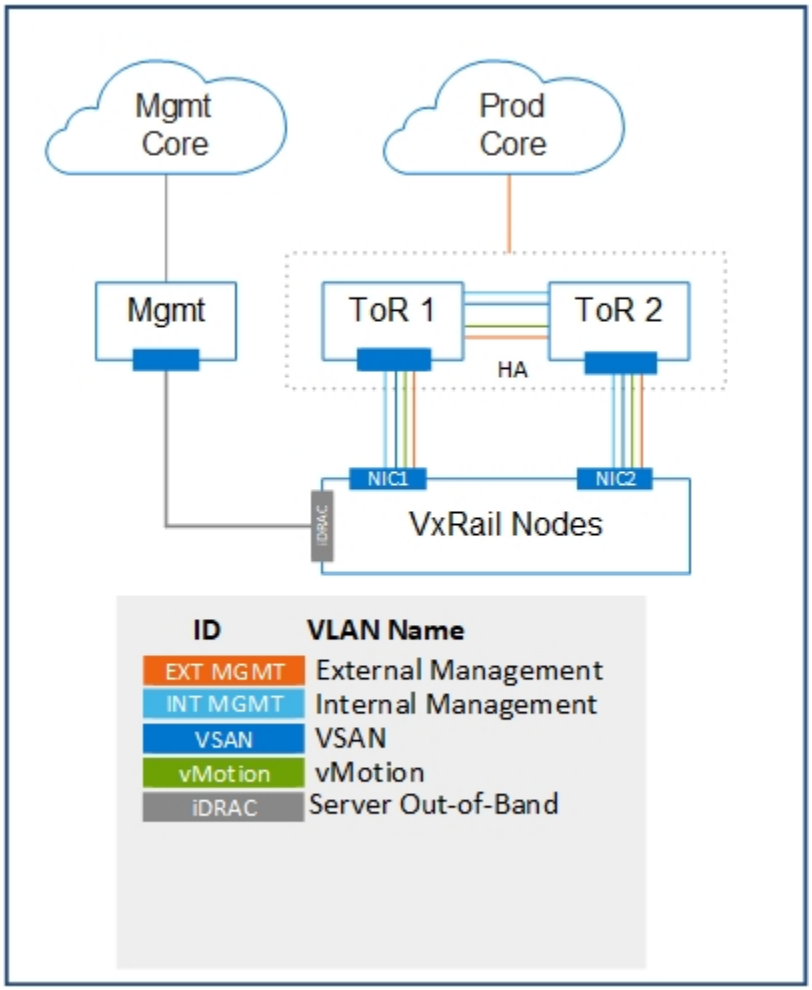

Figure 8 **VxRail Cluster VLANs**

VCF requires the following additional VLANs created and configured on the TOR switches connecting to VxRail nodes in the management WLD cluster before the VCF bring-up process is started using the VCF Cloud Builder tool.

Workload Domain	Network Virtualization Type	Network Traffic
Management WLD	NSV-V	VXLAN
Management WLD	NSV-V	ESG Uplink01
Management WLD	NSV-V	ESG Uplink02

Table 2. VCF VLANs for management WLD deployment

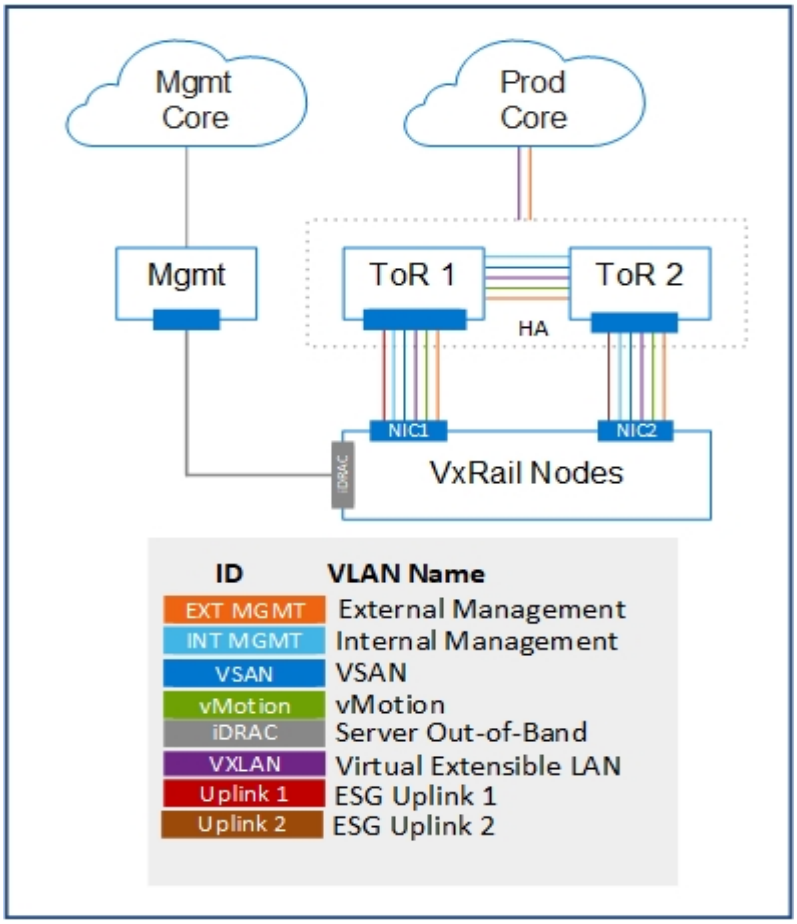

Figure 9 **VxRail Management WLD cluster VLANs**

The VMware Cloud Foundation requires the following additional VLANs created and configured on the TOR switches before deploying a VI WLD.

Workload Domain Type	Network Virtualization Type	Network Traffic
VI WLD	NSX-T	Host Overlay (Geneve)
VI WLD	NSX-T	Edge Node Uplink 1
VI WLD	NSX-T	Edge Node Uplink 2
VI WLD	NSX-T	Edge Overlay
VI WLD	NSX-V	VXLAN
VI WLD	NSX-V	ESG Uplink 1
VI WLD	NSX-V	ESG Uplink 2

Table 3. VCF VLANs for VI WLD deployment

Note: The Edge uplink deployment for both NSX-V and NSX-T based VI WLD is a manual process that must be performed after the VI WLD has been completed.

The following diagram illustrates the different port groups that are created on the Management workload domain VxRail vDS.

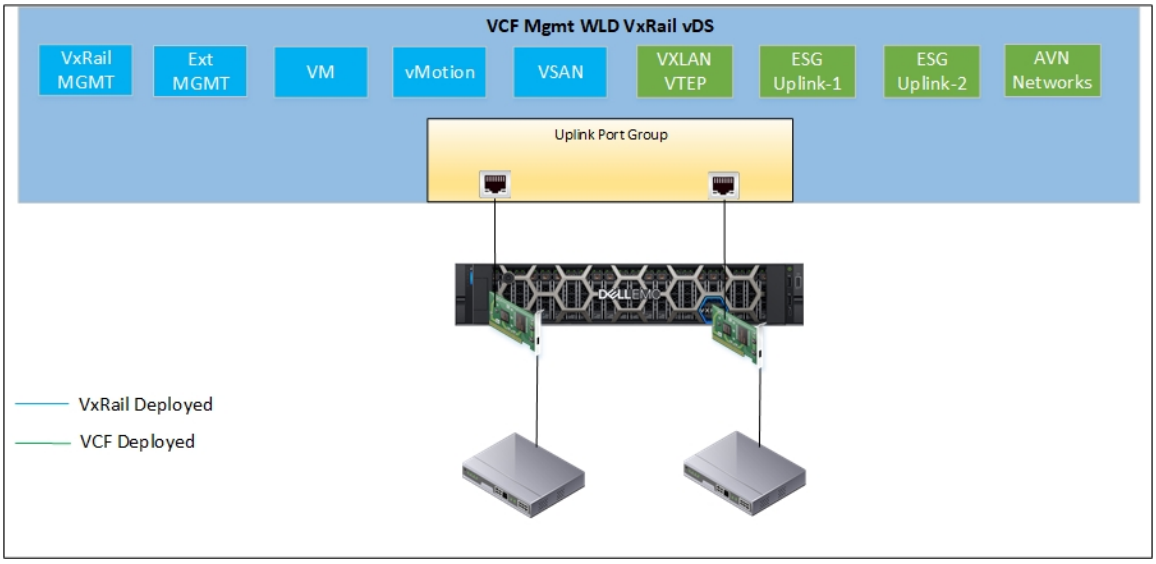

Figure 10 **VCF on VxRail management WLD vDS port groups**

4.2 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. The vSAN, vMotion, and external management (vSphere) network use load-based teaming policy. NSX-V VTEP does not support load-based teaming so this is also created with route-based on the originating virtual port. Finally, the NSX-V ESG port groups are created with just one active uplink that is pinned to each physical vmnic. The following table shows the teaming policies for each port group for a VxRail deployed with 2x10 or 2x25 GbE profile.

Port Group	Teaming Policy	VMNIC0	VMNIC1
VxRail Management	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
VXLAN VTEP (Only NSX-V)	Route based on the originating virtual port	Active	Active
ESG Uplink 1 (Only NSX-V)	Route based on the originating virtual port	Active	Unused
ESG Uplink 2 (Only NSX-V)	Route based on the originating virtual port	Unused	Active

Table 4. Port Group Teaming Policy for 2x10 or 2x25 GbE profile

You can also deploy a VxRail cluster with a 4x10 network profile for either a Mgmt

WLD or a VI WLD. This is not the VVD standard, but it is permitted for VCF on VxRail. The following table shows the teaming policy for each port group that is created with this profile.

Port Group	Teaming Policy	VMNIC0	VMNIC1	VMNIC2	VMNIC3
VxRail Management	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
VXLAN VTEP (Only NSX-V)	Route based on the originating virtual port	Standby	Standby	Active	Active
ESG Uplink 1 (Only NSX-V)	Route based on the originating virtual port	Active	Unused	Unused	Unused
ESG Uplink 2 (Only NSX-V)	Route based on the originating virtual port	Unused	Active	Unused	Unused

Table 5. Port Group Teaming Policy for 4x10 profile

Note: The NSX-V deployment uses vmnic2 and vmnic3 on the vDS for the VTEPs so VXLAN traffic is mixed with vSAN and vMotion. As a day 2 operation, you can move this traffic to vmnic0/vmnic1, if desired, to keep this traffic separate from vSAN.

5 Network Virtualization

The foundation of the Network virtualization layer for VCF on VxRail is provided by NSX-V or NSX-T. The Management domain supports NSX-V only, but the VI WLD domains can use either NSX-V or NSX-T. These solutions provide 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 does not provide. For multi-cloud connectivity and security, NSX-T should be deployed in the VI WLD as NSX-V has no multi-cloud support. NSX-T provides native support for Kubernetes, PKS, and Cloud Native applications.

5.1 NSX-V Components and Services

5.1.1 NSX Manager

The NSX Manager is responsible for the deployment of the NSX controller clusters and ESXi host preparation for NSX. The host preparation process installs various vSphere installation bundles (VIBs) to enable VXLAN, distributed routing, distributed firewall, and a user world agent for control plane communications. The NSX Manager is also responsible for the deployment and configuration of the NSX Edge services gateways and associated network services (load balancing, firewalling, NAT, and so on). It provides the single point of configuration and the REST API entry points for NSX in a vSphere environment.

The NSX Manager ensures security of the control plane communication of the NSX architecture. It creates self-signed certificates for the nodes of the controller cluster and ESXi hosts that are allowed to join the NSX domain. Each WLD has an NSX Manager as part of the VCF on VxRail solution.

5.1.2 NSX Controllers

The controller cluster in the NSX platform is the control plane component that manages the hypervisor switching and routing modules. The controller cluster consists of controller nodes that manage specific logical switches and includes three nodes that are clustered for scale-out and high-availability. The NSX controllers are required for each WLD, including management and any additional VxRail VI WLD.

5.1.3 NSX vSwitch

The vSwitch in NSX for vSphere is based on the VDS with additional components added to enable a rich set of services. The add-on NSX components include kernel modules that are distributed as VMware installation bundles (VIBs). These modules run within the hypervisor kernel, providing services including distributed routing, distributed firewall, and VXLAN to VLAN bridging. The NSX VDS abstracts the physical network, providing access-level switching in the hypervisor. This is central to network virtualization as it enables logical networks that are independent of physical constructs, such as VLANs.

The NSX vSwitch enables support for overlay networking with the use of the VXLAN protocol and centralized network configuration. Overlay networking with NSX provides the following capabilities:

- Creation of a flexible logical Layer 2 (L2) overlay over existing IP networks on existing physical infrastructure
- Agile provisioning of communication – both East–West and North–South – while maintaining isolation between tenants
- Application workloads and VMs that are agnostic of the overlay network, operating as if they were connected to a physical network.

- Massive scalability of hypervisors

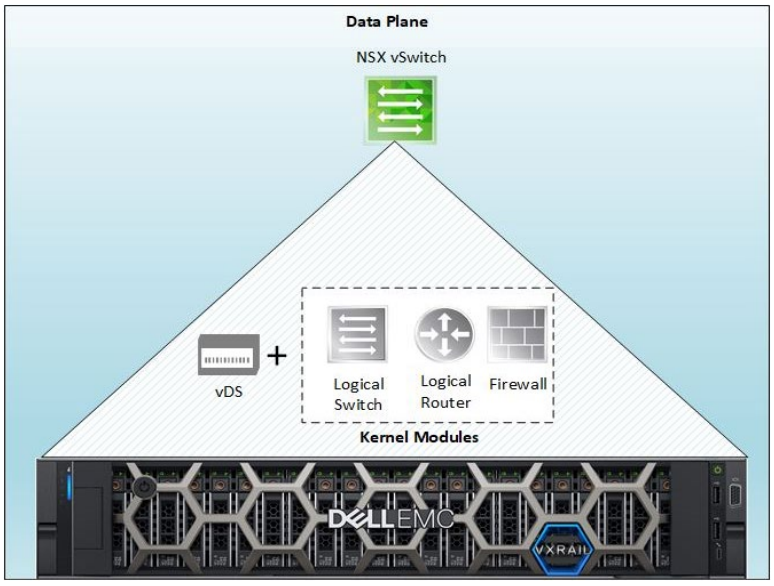

Figure 11 **NSX vSwitch Data Plane Components**

5.1.4 VXLAN and VTEPs

VXLAN is an overlay technology encapsulating the original Ethernet frames that are generated by workloads connected to the same logical Layer 2 segment or logical switch.

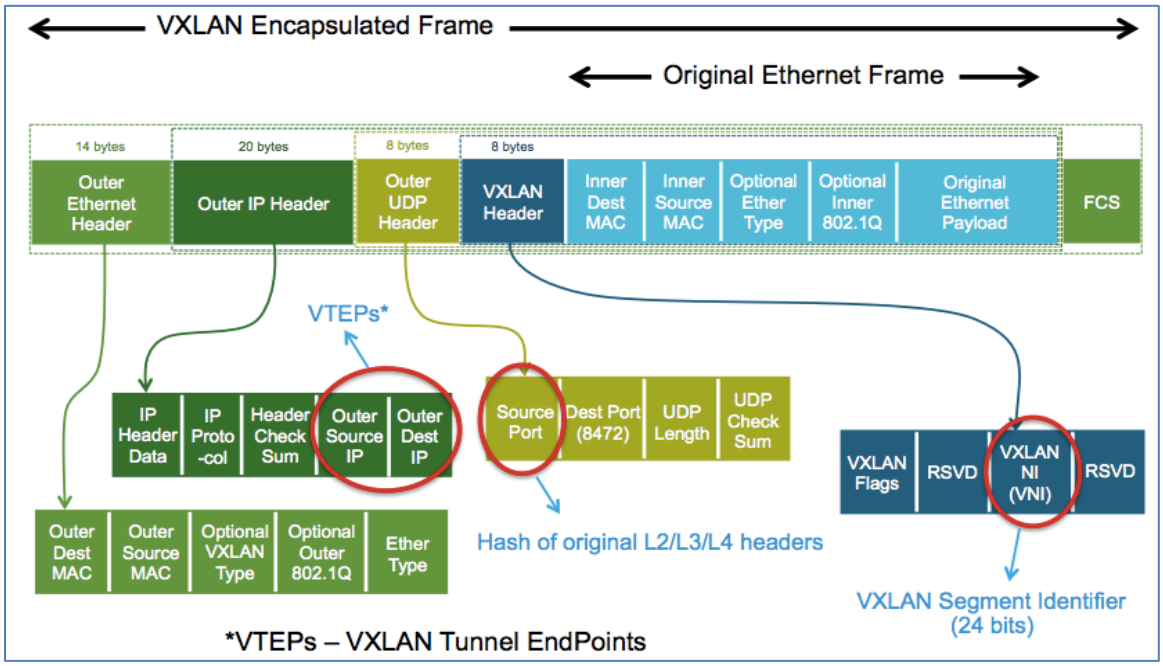

Figure 12 **VXLAN Encapsulation**

VXLAN is a L2 over L3 (L2oL3) encapsulation technology. The original Ethernet frame that is generated by a workload is encapsulated with external VXLAN, UDP, IP, and Ethernet headers to ensure it can be transported across the network infrastructure interconnecting the VXLAN endpoints.

Scaling beyond the 4094 VLAN limitation on traditional switches has been solved by leveraging a 24-bit identifier, named VXLAN Network Identifier (VNI), which is associated to each L2 segment created in logical space. This value is carried inside the VXLAN header and is normally associated to an IP subnet, similarly to what traditionally happens with VLANs. Intra-IP subnet communication occurs between devices that are connected to the same virtual network or logical switch.

VXLAN tunnel endpoints (VTEPs) are created within the vSphere distributed switch to which the ESXi hosts that are prepared for NSX for vSphere are connected. VTEPs are responsible for encapsulating VXLAN traffic as frames in UDP packets and for the corresponding decapsulation. VTEPs are essentially VMkernel ports with IP addresses and are used both to exchange packets with other VTEPs and to join IP multicast groups through Internet Group Membership Protocol (IGMP).

5.1.5 Logical switching

Logical switching enables extension of an L2 segment or IP subnet anywhere in the fabric independent of the physical network design. The logical switching capability in the NSX platform provides the ability to deploy isolated logical L2 networks with the same flexibility and agility that exists for virtual machines. Virtual and physical endpoints can connect to logical segments and establish connectivity independently from their physical location in the data center network.

5.1.6 Distributed logical routing

The NSX distributed logical router (DLR) provides an optimal data path for traffic within the virtual infrastructure, particularly East-West communications. It consists of a control plane component and a data plane component. The control virtual machine is the control plane component of the routing process, which provides communication between the NSX Manager and the NSX Controller cluster. NSX Manager sends logical interface information to the control virtual machine and the NSX Controller cluster. The control virtual machine sends routing updates to the NSX Controller cluster.

The data plane consists of kernel modules running on the hypervisor that provide high performance, low overhead first-hop routing.

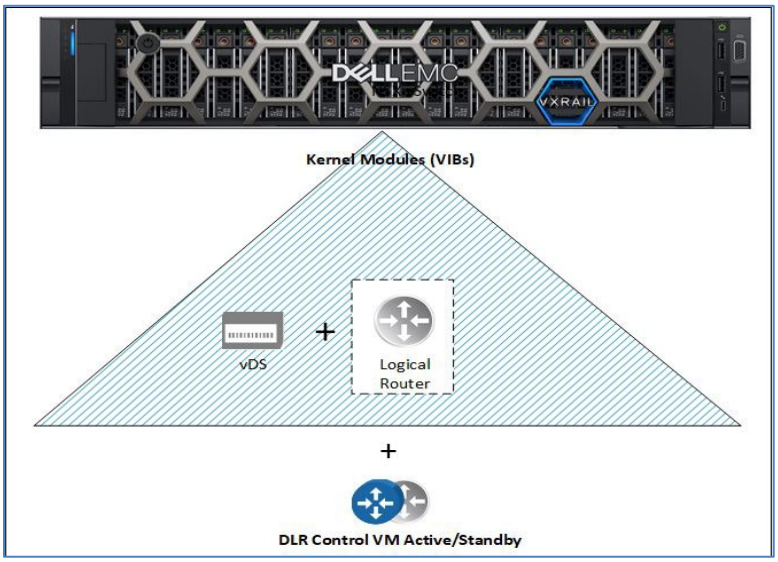

Figure 13 **DLR Components**

5.1.7 Edge services gateway (ESG)

The NSX Edge provides centralized on-ramp and off-ramp routing between the logical networks that are deployed in the NSX domain and the external physical network infrastructure. The NSX Edge supports various dynamic routing protocols (for example, OSPF, iBGP, eBGP) and can also leverage static routing. The routing capability supports two models, active-standby stateful services and ECMP. It also offers support for Layer 2, Layer 3, perimeter firewall, load balancing, and other services such as SSLVPN and DHCP-relay. Figure 12 shows how the Edge services gateways can be deployed in a pair using equal cost multi-path (ECMP) to load balance. The ESGs peer with an upstream physical router to allow traffic from the NSX domain out to the physical network and beyond to the Internet if necessary.

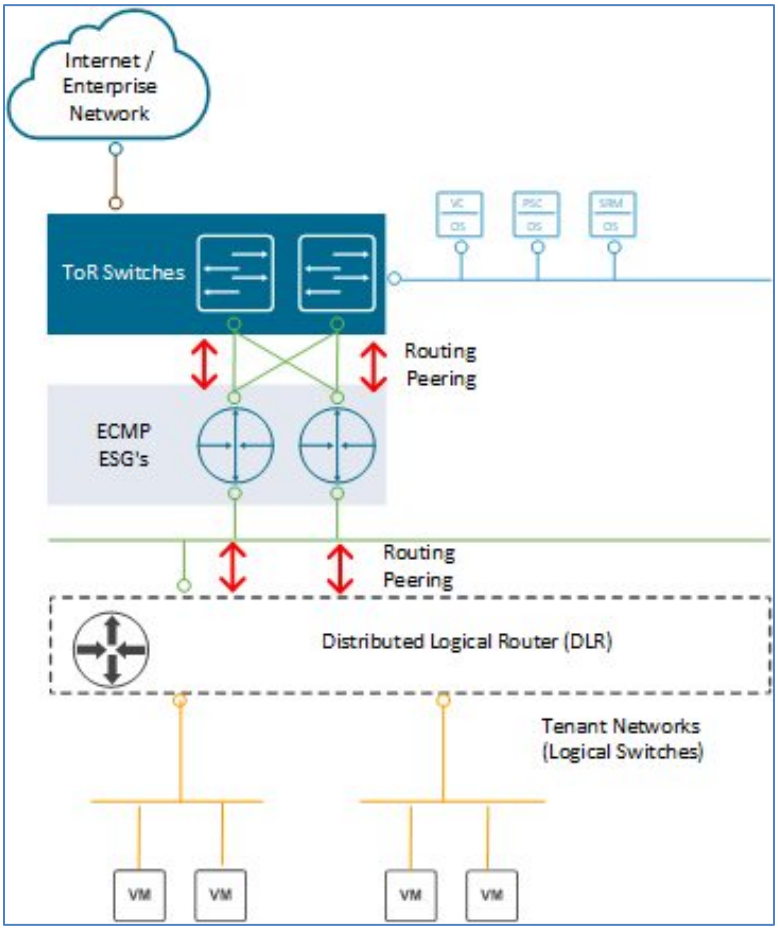

Figure 14 **Edge Services Gateway North–South Communication**

5.1.8 Distributed firewall (DFW)

The NSX DFW provides stateful firewall services to any workload in the NSX environment. DFW runs in the kernel space and provides near-line rate network traffic protection. The security enforcement implementation enables firewall rule enforcement in a highly scalable manner without creating bottlenecks on physical appliances. DFW is activated when the host preparation process is completed. If a VM does not require DFW service, it can be added to the exclusion list functionality.

By default, NSX Manager, NSX Controllers, and Edge services gateways are automatically excluded from DFW function. During deployment, VCF also adds the Management VMs to the DFW exclusion list.

5.2 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)

5.2.1 Management Plane

The management plane provides a single API entry point to the system. It is responsible for maintaining user configuration, handling user queries, and performing 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.

5.2.2 Control Plane

The control plane computes the runtime state of the system based on configuration from the management plane. It is also responsible for disseminating topology information that is reported by the data plane elements and pushing 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.

5.2.3 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.

5.3 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.

5.3.1 Segments (Logical Switch)

The segment previously known as logical switch is a Layer 2 construct similar to a VLAN backed network only 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.

5.3.2 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.

5.3.3 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.

5.3.4 Transport Node

Each hypervisor that is prepared for NSX-T and has an NDVS component installed is an NSX-T transport node 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.

5.3.5 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 will 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.

5.3.6 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 vSphere cluster. A vSphere cluster can run multiple NSX-T Edge clusters.

5.3.7 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.

6 NSX-V and NSX-T WLD Design

The initial deployment of the SDDC management or workload domain depends on the supporting physical network and underlying VxRail vSphere virtual network to establish basic network connectivity for domain management. It establishes the foundation for a future fully virtualized network with NSX. At this stage, network design considerations are applied to the domain to enable a fully virtualized network using NSX.

6.1 NSX-V based WLD

As documented in previous sections, there are two types of WLD domains pertaining to NSX: NSX-V based and NSX-T. The Mgmt WLD is only NSX-V based, but VI WLD domains can be either NSX-V or NSX-T.

6.1.1 NSX-V physical network requirements

NSX-V has the following external network requirements that must be met before the VCF on VxRail solution can be deployed.

- MTU 9000 for VXLAN traffic (for multi-site dual AZ ensure MTU across ISL)
- IGMP Snooping for VXLAN VLAN on the first hop switches
- IGMP querier is enabled on the connected router or Layer 3 switch.
- VLAN for VXLAN is created on the physical switches.
- DHCP is configured for VXLAN VLAN to assign the VTEPs IP.
- IP Helper on the switches if the DHCP server is in different L3 network
- Layer 3 license requirement for peering with ESGs
- BGP is configured for each router peering with an ESG.
- Two Uplink VLANs for ESGs in Mgmt WLD
- AVN subnets are routable to the Mgmt WLD management network.
- AVN networks are routable at the Core network to reach external services.

6.1.2 NSX-V deployment in management WLD

The deployment of the Mgmt WLD includes installation of NSX components and layers down the Application Virtual Network (AVN) for the vRealize Suite. It deploys the Edge services gateway (ESG) VMs and the universal logical router (uDLR) and configures dynamic routing to allow traffic from the AVNs to the external networks. The following steps are performed by the VCF Cloud Builder to deploy and configure NSX during the Mgmt WLD bring-up process:

1. Create two ESG uplink port groups on the vDS.
2. Deploy NSX Manager in Mgmt WLD cluster.
3. Register NSX Manager with PSC01.
4. Register NSX Manager with Mgmt WLD VC.
5. Install license for NSX.
6. Set NSX Manager as the Primary NSX Manager.
7. Deploy three NSX controllers to Mgmt WLD cluster.
8. Create anti-affinity rules for controllers.
9. Create VXLAN segment ID range.
10. Create a VXLAN Multicast IP range.
11. Create a universal transport zone.
12. Add cluster to transport zone.
13. Install NSX VIBs (host prep).
14. Create VXLAN port group.
15. Create Host VTEPs.

16. Deploy 2x ESGs.
17. Create anti-affinity rules for the ESCGs.
18. Deploy uDLR.
19. Create anti-affinity rules for the uDLR Control VMs.
20. Configure Dynamic Routing.
21. Create AVN Networks for vRealize Suite.

Figure 15 shows the NSX-V components that are deployed after the VCF bring-up process is completed.

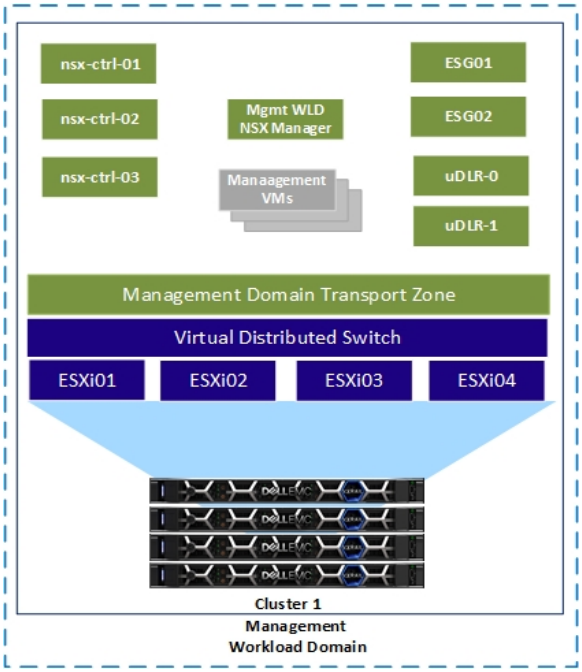

Figure 15 **Management WLD after initial deployment**

6.1.3 NSX-V deployment in the VI WLD

Although the deployment of NSX-V in the WLD is becoming less common as NSX-T becomes the standard network virtualization platform, we still want to cover the deployment of an NSX-V VI WLD. SDDC Manager is used to deploy the NSX-V WLD. The fundamental difference is that NSX-V Edge routing is not deployed or configured, and no Universal objects are created during the deployment. The following steps are performed by SDDC Manager to deploy and configure NSX-V during the VI WLD deployment process.

1. Deploy VI WLD NSX Manager in Mgmt WLD cluster.
2. Register NSX Manager with PSC01.
3. Register NSX Manager with VI WLD VC.
4. Install license for NSX.
5. Deploy three NSX controllers to VI WLD cluster.
6. Create anti-affinity rules for controllers.
7. Create VXLAN segment ID range.
8. Create a VXLAN Multicast IP range.
9. Create a global transport zone.
10. Add cluster to transport zone.
11. Install NSX VIBs (host prep).
12. Create VXLAN port groups.
13. Create Host VTEPs.

Figure 16 shows the NSX-V components that are deployed after an NSX-V WLD domain has been added and a second cluster has been added to the VI WLD. Note when a second cluster is added, only the preceding Steps 10, 11 and 13 are performed by SDDC Manager.

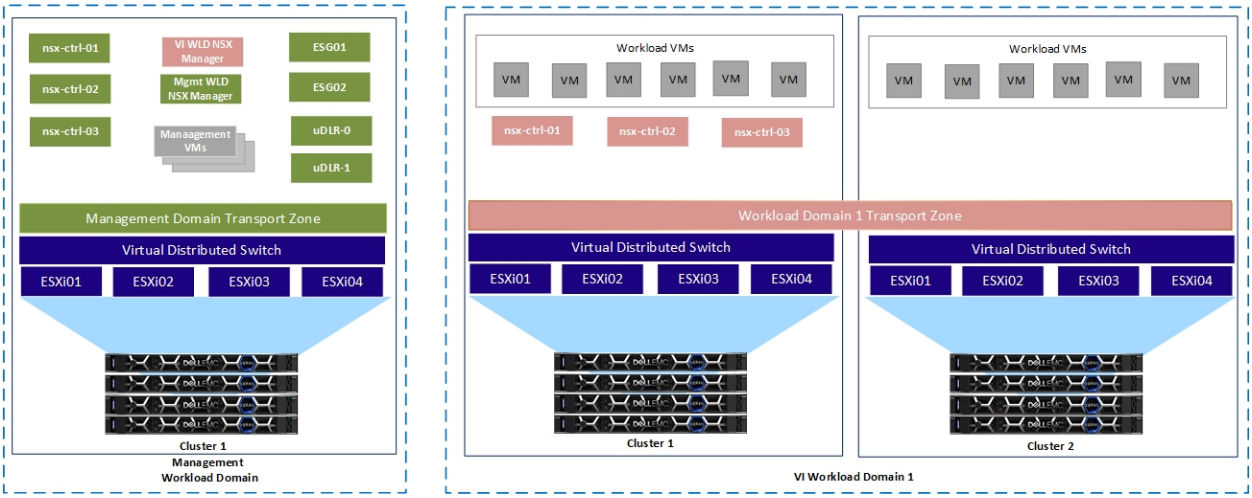

Figure 16 **Management WLD and NSX-V VI WLD with two clusters**

6.1.4 NSX-V Transport Zone Design

A transport zone controls which hosts a logical switch can reach and span one or more vSphere clusters. Transport zones dictate which clusters and, therefore, which VMs can participate in the use of a Layer 2 network.

6.1.4.1 Management WLD Transport Zone

The Mgmt WLD has a dedicated NSX-V Manager. It has its own transport zone created during the VCF bring-up process, the transport zone is created as a universal transport zone, universal objects can span sites. This is important for multi-region deployments that will be supported in a later release of the solution. Figure 17 shows the Mgmt WLD transport zone as a universal transport zone. Associated with this transport zone is the universal multicast IP range that will be used for the Universal Logical Switches or AVNs.

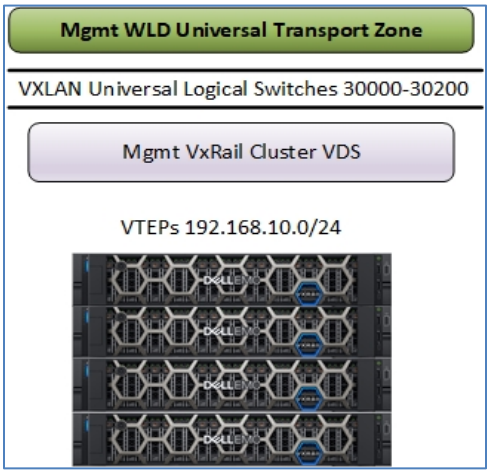

Figure 17 **Management WLD Universal Transport Zone**

6.1.4.2 VI WLD Transport Zone

When creating a VI WLD, the SDDC Manager creates the transport zone while the first cluster is added to the VI WLD. Subsequent clusters are added to the transport zone. This allows VMs in a WLD on the same logical switch to span clusters. This can be useful when designing three-tier applications using a flat Layer 2 but keeping different workloads on different clusters. Micro-segmentation provides security between the different tiers of the application. Figure 18 shows the transport zone configuration for a VI WLD with three clusters added. As each cluster is added, it is added to the transport zone.

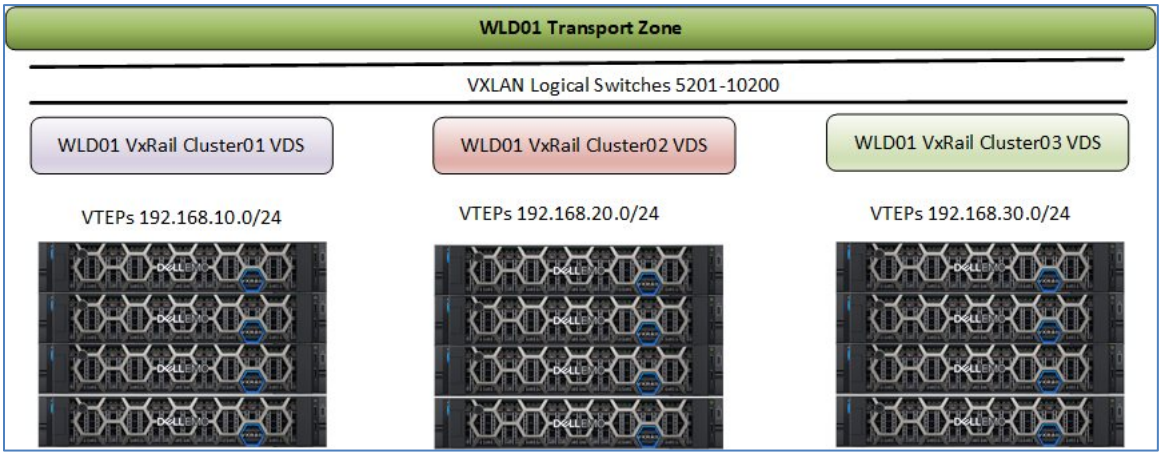

Figure 18 **NSX-V based VI WLD Transport Zone**

6.1.5 NSX-V Logical switch control plane replication mode

The control plane decouples connectivity in the logical space from the physical network infrastructure and handles the broadcast, unknown unicast, and multicast (BUM) traffic within the logical switches. The control plane is on top of the transport zone and is inherited by all logical switches that are created within it. VCF on VxRail uses the hybrid replication mode to send BUM traffic. This mode is an optimized version of the unicast mode where local traffic replication for the subnet is offloaded to the physical network. Hybrid mode requires IGMP snooping on the first-hop switch and access to an IGMP querier in each VTEP subnet or VLAN. Using VLANs for the management domain VXLAN VLAN that are different than the VLANs used for WLDs is recommended to completely isolate the management and tenant traffic in the environment. The hybrid replication mode operation is depicted in Figure 19, where a specific VTEP performs replication to the other local VTEPs. This VTEP uses L2 multicast to replicate BUM frames locally, while the unicast is used to send the traffic to a designate VTEP in a remote L3 segment. This performs the same L2 multicast replication in that segment.

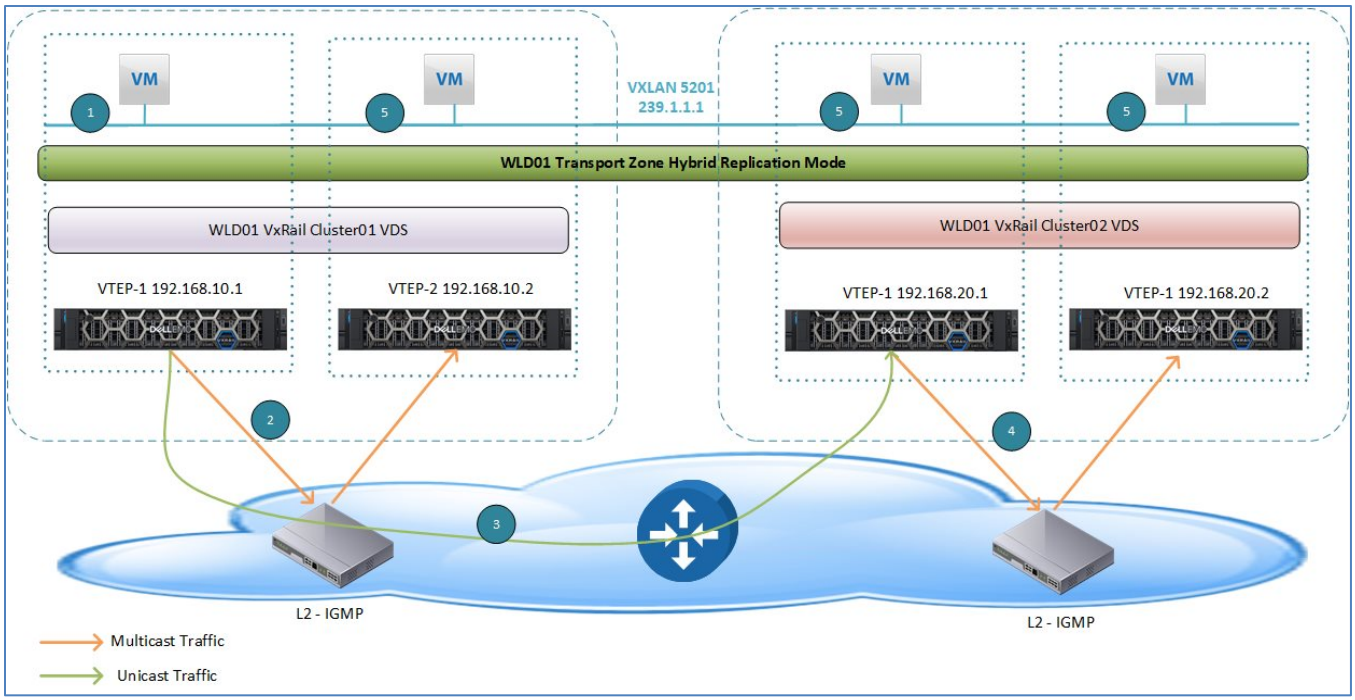

Figure 19 **Hybrid Replication Mode across Clusters in VI WLD**

6.1.6 Management WLD Application Virtual Network

The Application Virtual Network (AVN) is a term used to describe the network used to connect certain management applications running in the management workload to NSX-V backed logical switches. Starting with version 3.9.1, virtual routing is deployed for the management WLD using cloud builder to allow the deployment of the vRealize suite onto the AVN or NSX-V Universal logical switches. This is to prepare for a future release where a Multi-Region disaster recovery solution can be adopted to allow the failover and recovery of the vRealize suite in the case of a full site failure.

Note: Any VCF environments upgraded from 3.9 will not automatically get AVN deployed. The Log Insight VMs remain on the Management network and the vRealize Suite remains on a VLAN backed network. If migration to AVN is required, it will need to be done manually. See VCF 3.9.1 release notes for further details VCF 3.9.1 Release Notes.

The topology that is required for AVN is shown in Figure 20.

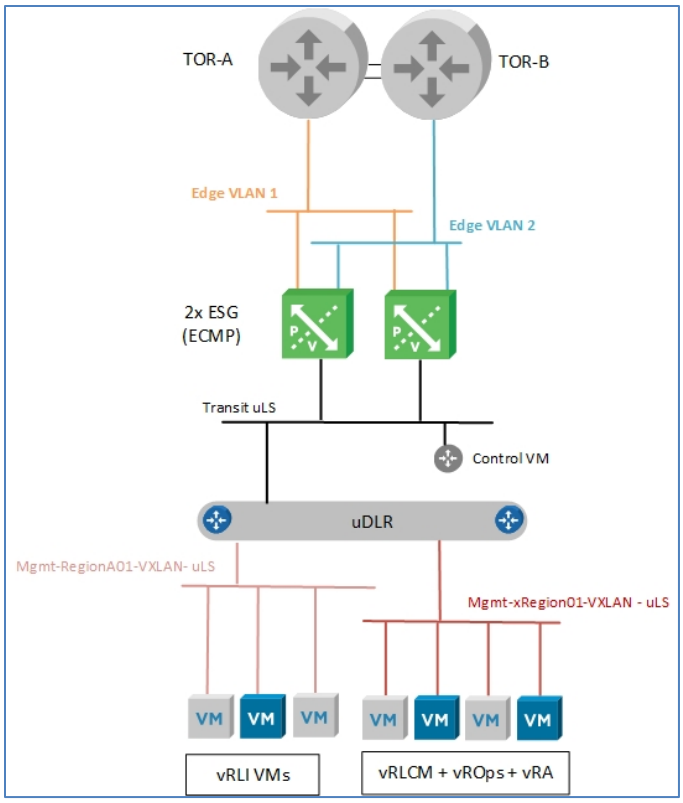

Figure 20 **AVN connectivity for Management WLD**

AVN contains the following components:

- Dual TOR switches running BGP protocol
- Two Edge VLANs are used as transit for the ESG uplinks to each upstream router.
- Two ESGs in ECMP mode
- One Universal Transit Logical Switch connecting ESGs to uDLR
- One uDLR
- Two Universal Logical Switches for the AVN

During the deployment, anti-affinity rules are used to ensure the NSX Controllers, uDLR Control VMs and the ESGs do not run on the same node simultaneously. This is critical to prevent impact to the network services if one host fails in the management WLD. The following diagram illustrates how these components are typically separated by the anti-affinity rules that are created and applied during the management WLD deployment.

Figure 21 **Mgmt WLD NSX Component Layout**

To expand on the routing design, Figure 22 shows a typical BGP configuration for the management workload domain deployment. This figure shows the eBGP peering between the Mgmt WLD ESGs and the upstream switches and iBGP is used between the ESGs and the uDLR.

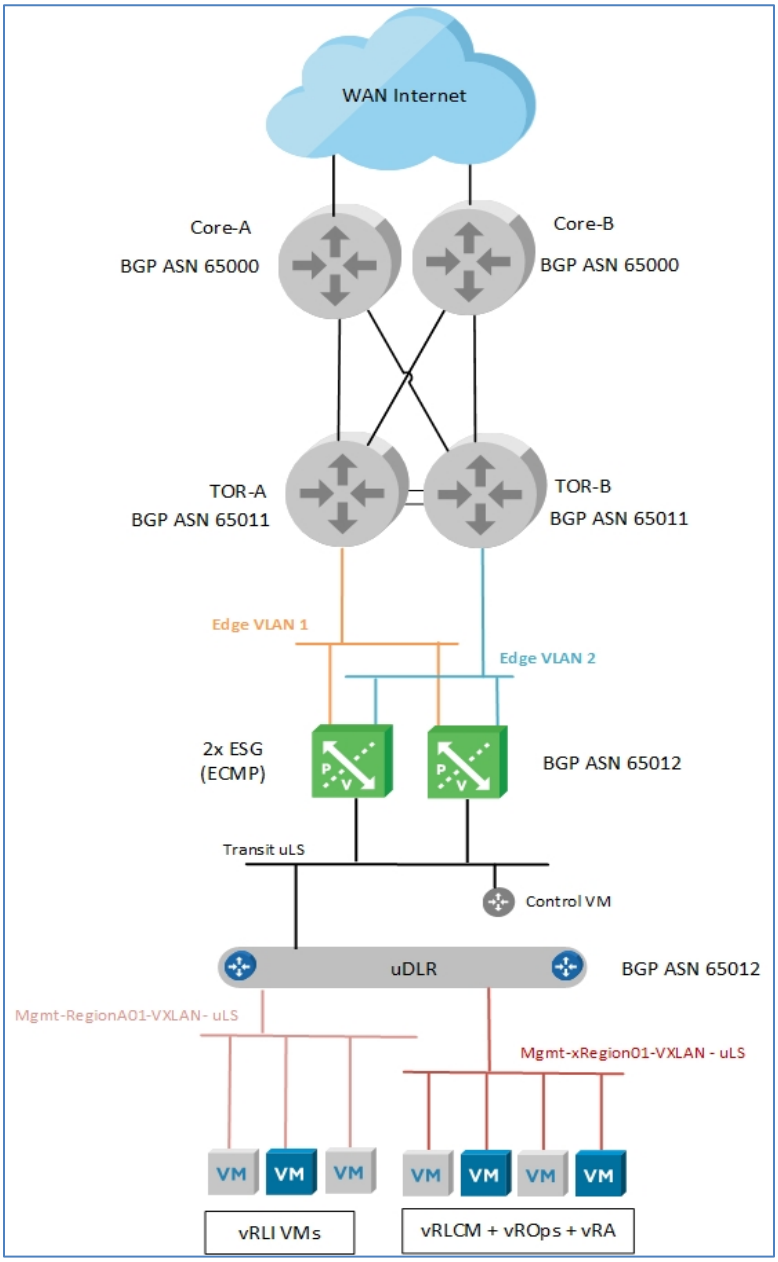

Figure 22 **AVN NSX Virtual Routing for Management WLD**

The following network BGP configuration is implemented as part of the deployment:

- eBGP neighbors that are created between both ESGs and both TORs
- BGP Password is applied in the neighbor configuration.
- BGP timers of 4/12 applied to the eBGP TOR neighbor
- Static routes created on the ESGs to protect against uDLR control VM failing.
- Redistribute connected and static routes from ESG to uDLR.
- iBGP configured between ESGs and uDLR
- BGP timers of 1/3 applied to iBGP uDLR neighbor
- Redistribute connected and static routes from uDLR to ESG.

As mentioned previously, the AVN networks must be routable to the VLAN backed management network for the management WLD and also any management networks for any additional VI WLDs that are deployed. This allows the management components on the AVN network to communicate to the VLAN backed management components on the VI WLD. For example, Log Insight VMs on the AVN network must communicate with hosts on a VI WLD management network. Figure 23 illustrates the connectivity and VLANs required for communication between management WLD components on the management VLAN backed network and the vRealize components on the AVN network. Any VI WLDs that are added would have similar connectivity requirements. In this example, the two AVN subnets 172.16.11.0/24 and 172.16.12.0/24 must be routable to the management network VLAN 100 and subnet 192.168.100.0/24.

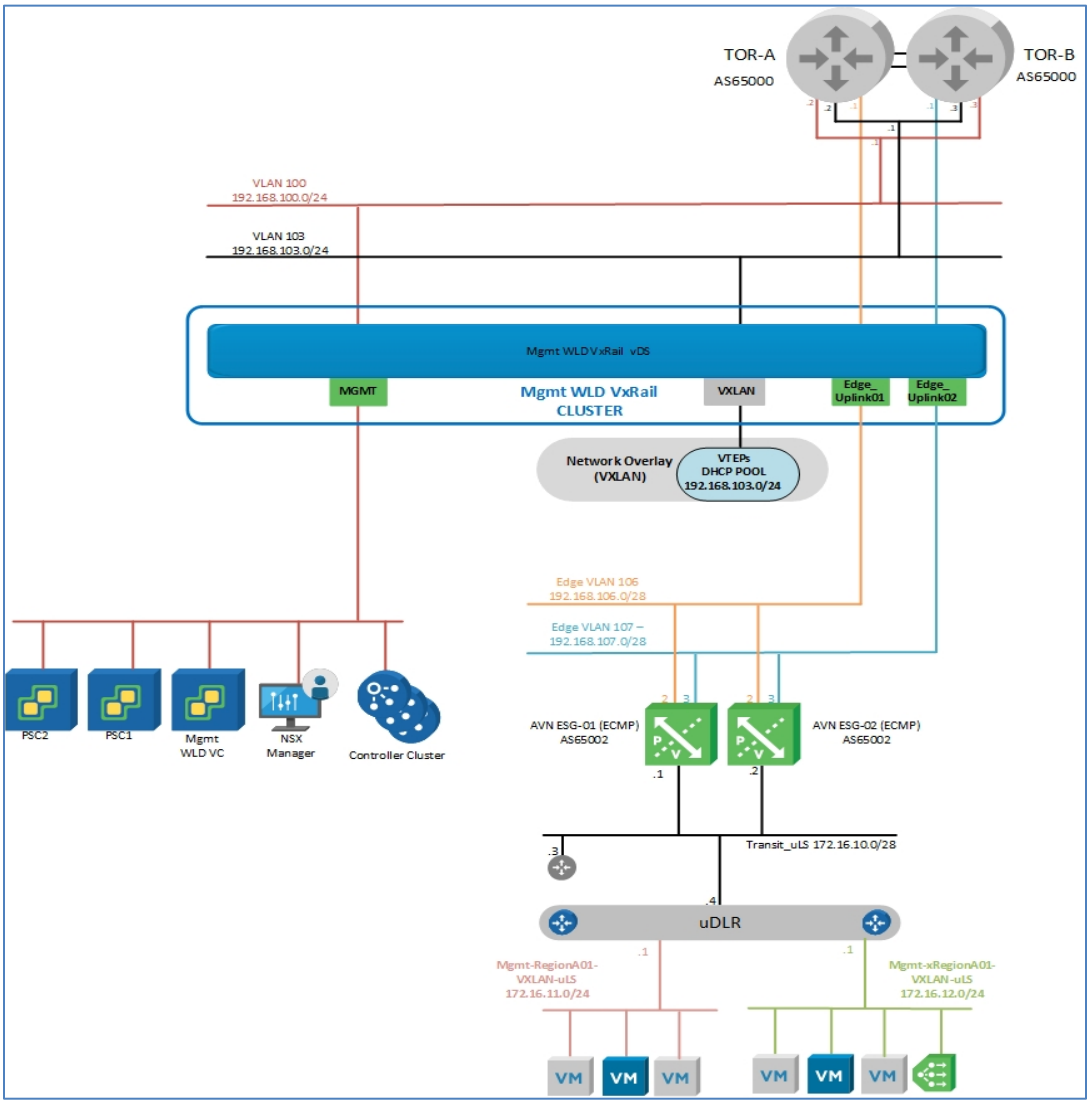

Figure 23 **AVN connectivity to Mgmt WLD management network**

6.2 NSX-T based VI WLD

The following section describes the design of the NSX-T VI WLD.

6.2.1 NSX-T physical network requirements

The following NSX-T external network requirements must be met before deploying any NSX-T based VI WLD from SDDC Manager.

- MTU 9000 for Geneve (Overlay) traffic
- Host Overlay VLAN is created on the physical switches.
- DHCP is configured for each VLAN to assign the Host TEPs IP.
- IP Helper on the switches if the DHCP server is in different L3 network.
- 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.

6.2.2 NSX-T deployment in VI WLD

The NSX-T components are installed when the first VxRail cluster is added to the NSX-T VI WLD. The SDDC Manager deploys NSX-T components onto the management and the VI WLD clusters. The following list highlights the major steps that are performed during the deployment process:

1. Deploy NSX-T Managers in Mgmt WLD 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 VLAN Transport zone.
8. Create NSX-T VLAN backed segments.
9. Create a NIOC profile.
10. Create an Uplink profile.
11. Create Transport Node Profile.
12. Prepare the hosts in the cluster for NSX-T.

Note: No additional NSX-T managers are needed when a second NSX-T based VI WLD is added. The vCenter is added as a Compute Manager and the ESXi hosts are prepared for use for NSX-T.

Figure 23 shows the NSX-V and NSX-T components deployed in the MGMT VI WLD. It shows the VI WLD with two NSX-T clusters added to the VI WLD.

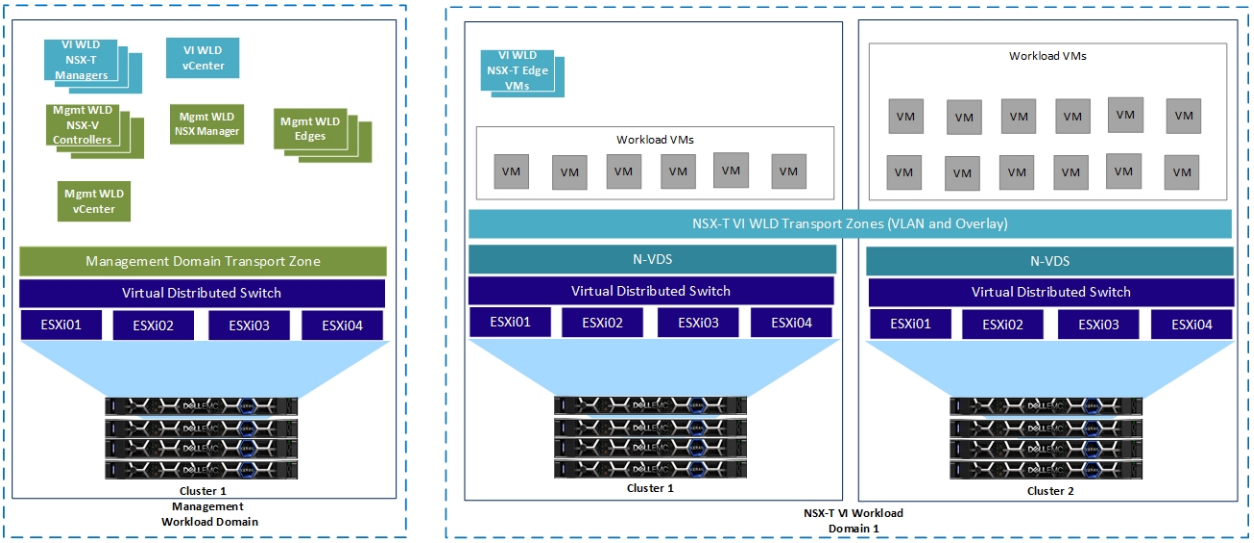

Figure 24 **NSX-T VI WLD Cluster Design**

6.2.3 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.

When the first cluster is added to the first VI WLD, SDDC Manager creates the Overlay and VLAN transport zones in NSX-T Manager. Two additional VLAN transport zones must be manually created on Day 2 for the Edge VM uplink traffic to the physical network.

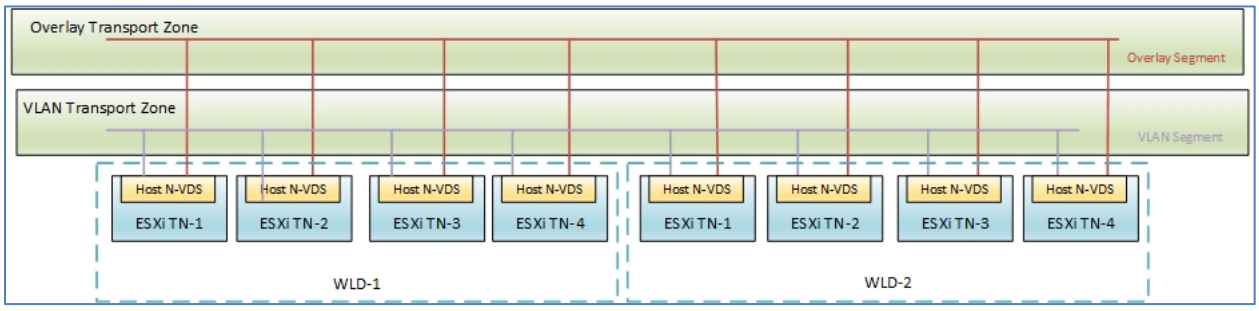

Figure 25 **NSX-T Transport Zones**

Note: When subsequent clusters are added to a WLD, or if a new WLD is created, all the nodes participate in the same VLAN and Overlay Transport Zones. . For each cluster the same VLAN or a different VLAN can be used for the TEP traffic for the Overlay.

6.2.4 NSX-T segments

Segments are used to connect VMs to Layer 2 networks, and they can be either VLAN or Overlay segments. Following is the complete list of segments that are needed to support the virtual infrastructure for an SDDC created on VCF on VxRail.

Segment Name	Uplink and Type	Transport Zone	VLAN (example)
Overlay (VCF Deployed)	None	VLAN-TZ	None
Edge-uplink01	None	VLAN-TZ	0-4094
Edge-uplink02	None	VLAN-TZ	0-4094
Edge-Overlay	None	VLAN-TZ	0-4094
uplink01	None	Uplink01-TZ	1647
uplink02	None	Uplink02-TZ	1648

Table 6. NSX-T Segments for VCF on VxRail

Note: In table 7, only the Overlay segment is created during the deployment of the NSX-T WLD. The other segments must be created manually on Day 2.

6.2.5 Uplink profile design

The uplink profile is a template that defines how an N-VDS that exists in each transport node (either host or Edge VM) connects to the physical network. It specifies:

- Uplinks to be used in the N-VDS
- Teaming policy that is applied to those uplinks
- VLAN used for the profile
- MTU applied to the traffic

Profile	Teaming Policy	Active Uplinks	VLAN (example)	MTU
Host-uplink (VCF Deployed)	Load Balance Source	uplink-1,uplink-2	1644	9000
Edge-overlay-profile	Failover Order	uplink-1	1649	9000
Edge-uplink01-profile	Failover Order	uplink-1	1647	9000
Edge-uplink01-profile	Failover Order	uplink-1	1648	9000

Table 7. NSX-T Uplink Profiles

Each time a new cluster is added to an NSX-T 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 clusters.
For a single cluster VI WLD, four different uplink profiles are required to complete the overall deployment of an NSX-T WLD, including the dynamic routing configuration. The host uplink profile is auto generated when the cluster is added to the VI WLD. The uplink profiles for the Edge connectivity need to manually created following the VVD guidance located here Create Uplink Profiles.

6.2.6 Transport node profiles

A transport node as described earlier is either a host or an edge VM that has an N-VDS component installed and is 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.

- Transport Zones for N-VDS participation – Overlay and VLAN TZ
- N-VDS name – VCF defined
- NIOC profile – VCF defined
- Uplink profile - See Table 7
- LLDP profile – Send LLDP packets
- IP Assignment type for the TEPs – DHCP
- Physical NIC Mapping – vmnics to uplinks, dependent on the available NICs on the VxRail node

During the VCF deployment of an NSX-T VI WLD, when a new cluster is added to the VI WLD,a transport profile is created with the settings in the preceding list. When the clusters are added to the NSX-T VI WLD, the transport node profile is applied to the nodes in the cluster, creating the N-VDS, adding the nodes to the transport zones, configuring the physical interfaces and creating and assigning an IP to a TEP so hosts can communicate over the overlay network. Figure 26 shows a compute node with Logical segments created that can use the N-VDS to communicate to VMs in the same transport zone.

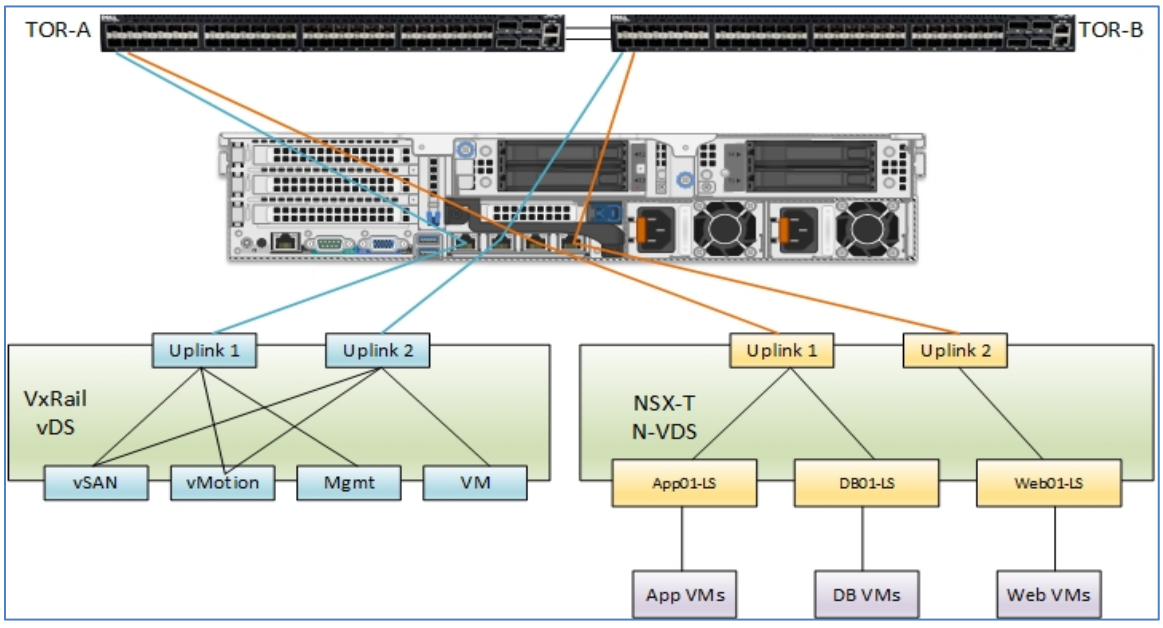

Figure 26 **Compute Node**

Note: The application logical segments can be either Overlay-backed or VLAN-backed segments.

6.2.7 NSX-T Edge Node design

The edge node design follows the VVD design and is a manual configuration for VCF on VxRail, two edge node VMs are deployed in the first VI WLD cluster. VCF on VxRail has a shared edge and compute cluster design meaning the edge node VMs overlay and uplink interfaces connect to the Host N-VDS for external connectivity. The management interface connects to the VxRail vDS port group as shown in Figure 27.

For additional details on the edge node connectivity design please, see the VVD documentation located here Transport Node and Uplink Policy Design.

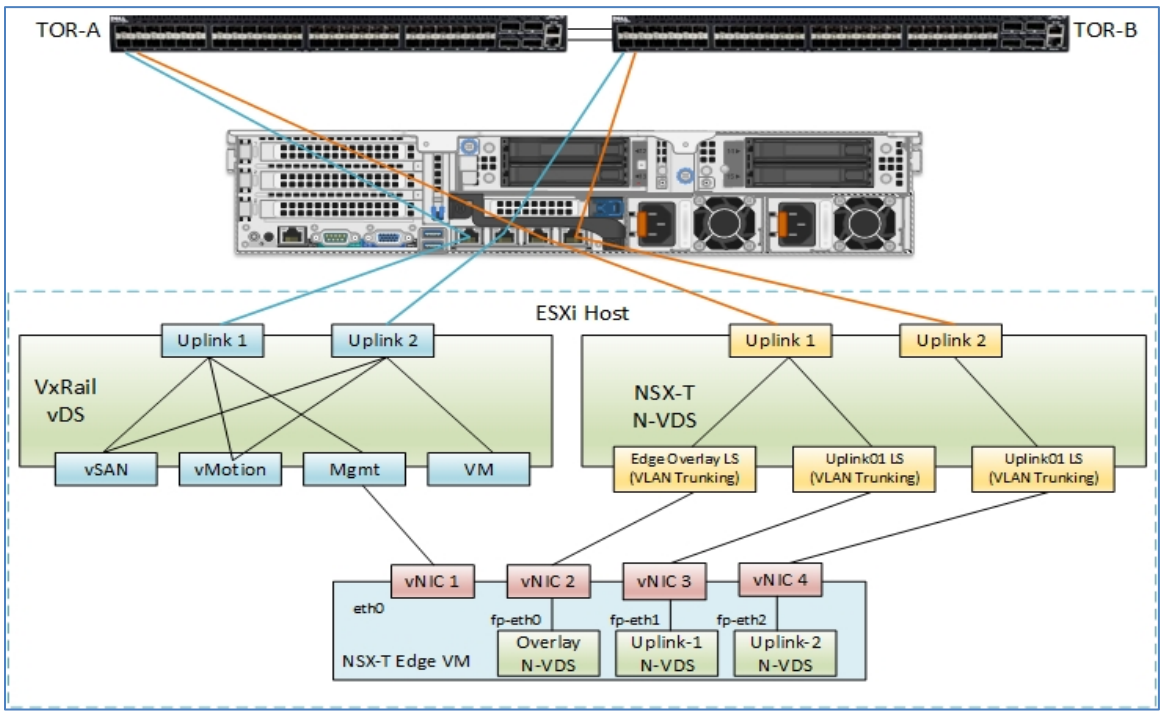

Figure 27 **Edge Node connectivity design**

Note: The Overlay and uplink segments used to connect the edge VM overlay and uplink interfaces are in trunking mode as the Edge transport node NVDS will use the VLAN tagging.

The NSX-T edge routing design is based on the VVD design located here Routing Design using NSX-T. A Tier-0 gateway is deployed in Active/Active mode with ECMP enabled to provide redundancy and better bandwidth utilization. Both uplinks are utilized. Two uplink VLANs are needed for North/South connectivity for the Edge virtual machines in the Edge Node cluster. 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.

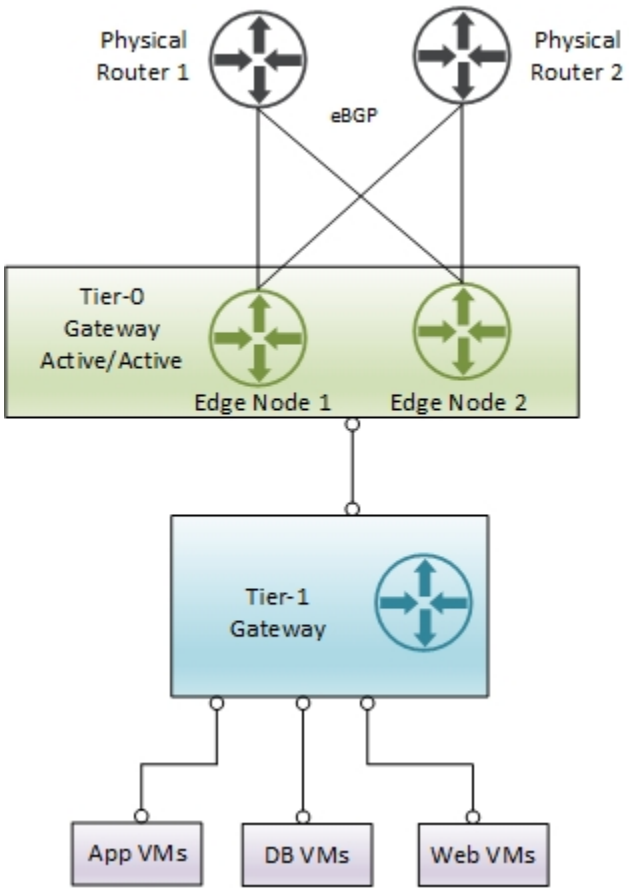

Figure 28 **Edge Node North/South connectivity**

The manual configuration steps to deploy the Edge node cluster following the VVD design are located here [Deploy NSX-T Edge Cluster on VxRail].

7 Physical network design considerations

The VCF on VxRail network design offers flexibility to allow for different topologies and different network hardware vendors. This enables users to use their 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 factors:

- NSX-V or 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 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.

7.1 Traditional 3-tier (access/core/aggregation)

The traditional 3-tier design is based on a Layer 2 fabric, as shown in Figure 29.

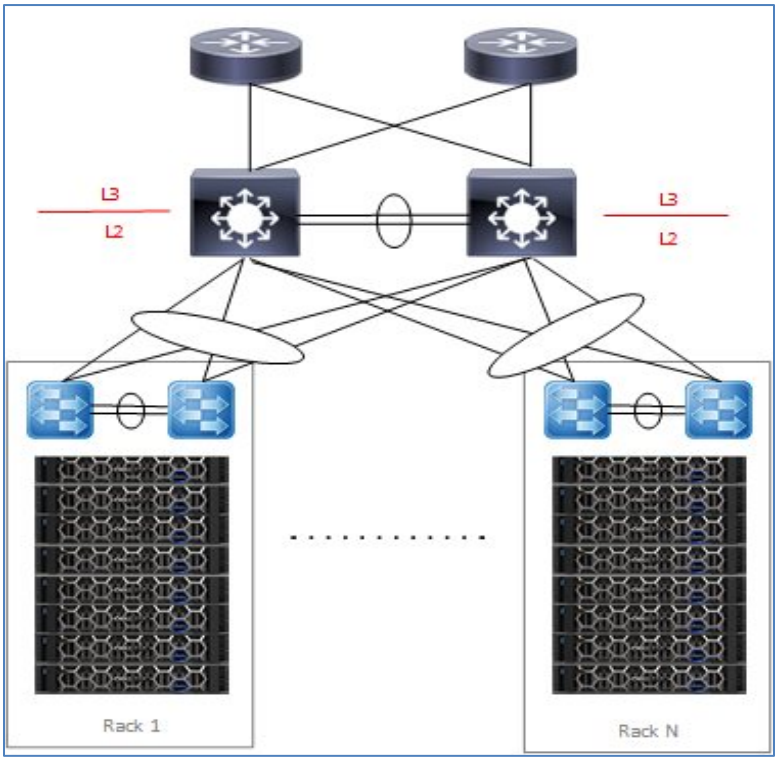

Figure 29 **Traditional 3 Tier Layer 2 Fabric Design**

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 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-V ESGs or NSX-T T0 Gateway will peer 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 clusters spanning racks will create large broadcast domains.
- Interoperability issues between different switch vendors can introduce spanning tree issues in large fabrics.
- The NSX-V ESGs or NSX-T T0 gateways for each WLD will need to peer at the aggregation layer. For large scale deployments with multiple WLDs, the configuration will become 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.

7.2 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 Figure 30.

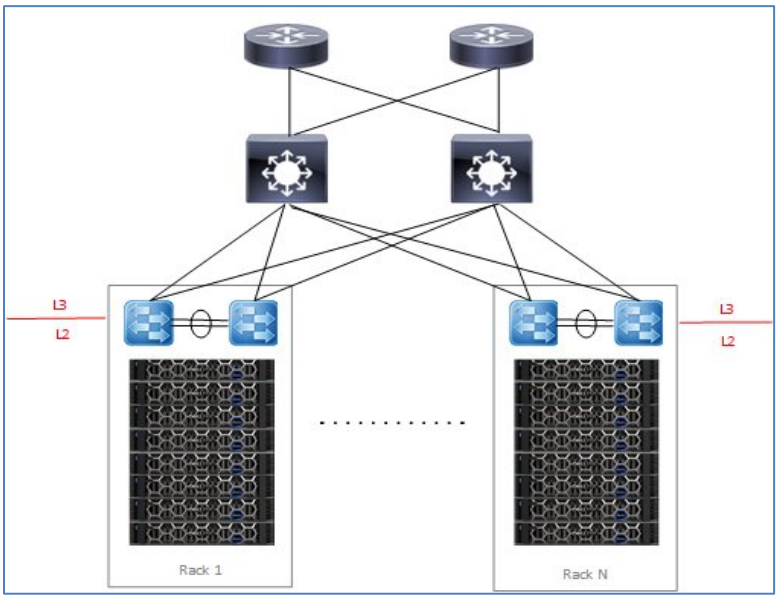

Figure 30 **Leaf and Spine Layer 3 Design**

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-V ESGs or NSX-T T0 Gateway for the WLD will peer with the leaf switches in each 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 are routed through the Spine.
- ESG peering is simplified by peering the WLDs with the leaf switches in the rack.

Disadvantages:

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

7.3 Multi-rack design considerations

It might be desirable to span WLD clusters across racks to avoid a single point of failure within one rack. The loudmouth protocol for VxRail node discovery requires VxRail nodes to reside on the same L2 private management discovery network. Additionally, VxRail does not yet support L3 for vSAN. L3 for vMotion is a post VxRail cluster deployment operation and the management VMs will also need L2 adjacency so the VMs can be migrated between racks. For a Layer 3 Leaf-Spine fabric, this is a problem as the VLANs are terminated at the leaf switches in each rack.

7.3.1 VxRail multi-rack cluster

VxRail multi-rack cluster is a network design that allows a single (or multiple) VxRail cluster(s) to span between racks. This particular solution uses a Dell networking switch 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 multi-rack 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 EMC Networking Guides.

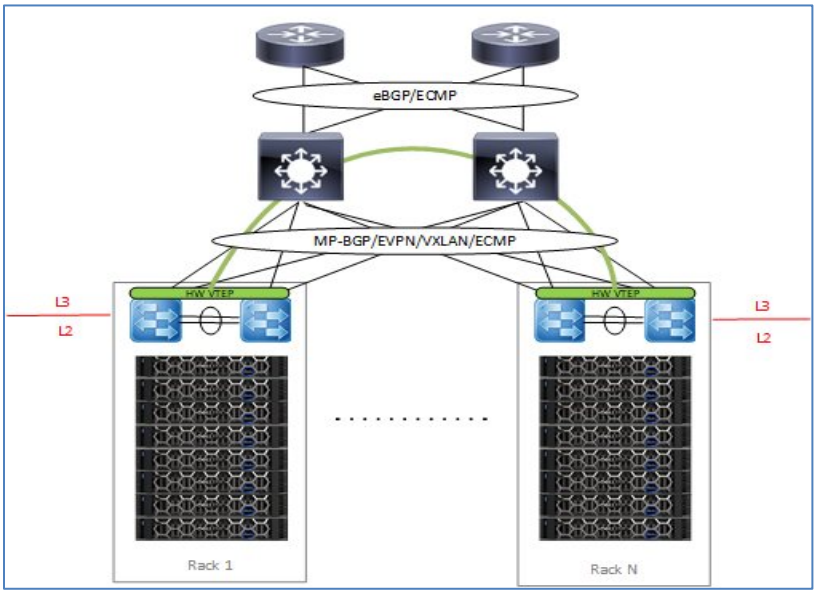

Figure 31 **Multi-Rack cluster with hardware VTEP**

7.4 VxRail Physical network interfaces

The VxRail can be deployed with either 2x10/2x25 GbE or with 4x10 GbE profile. It will need the necessary network hardware to support the initial deployment. There are two important considerations that must be kept in mind when planning or designing the physical network connectivity:

- NSX-V based WLDs (Mgmt or a VI WLD) deploy the VXLAN VTEP Port Group to the VxRail vDS.
- NSX-T based VI WLD require additional uplinks. The uplinks that were used to deploy the VxRail vDS cannot be used or the NSX-T N-VDS.

The following physical host connectivity diagrams illustrate the different host connectivity options for NSX-V and NSX-T based WLDs.

7.4.1 NSX-V based WLD physical host connectivity options

Figure 32 shows a VxRail deployed with a 2x10 profile on the NDC network card with NSX-V (Mgmt or VI WLD). The remaining two ports on the NDC can be used for other traffic types such as iSCSI, NFS, and Replication.

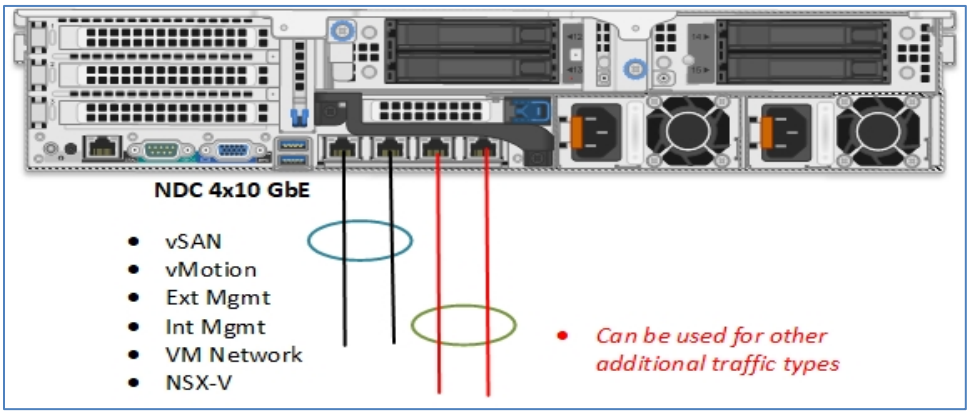

Figure 32 **VxRail 2x10 network profile with NSX-V (Mgmt or VI WLD)**

Figure 33 shows a VxRail deployed with a 4x10 profile with a 4-port NDC card. The NSX-V VXLAN traffic uses all four interfaces. An additional PCI-E card can be installed for additional network traffic if required.

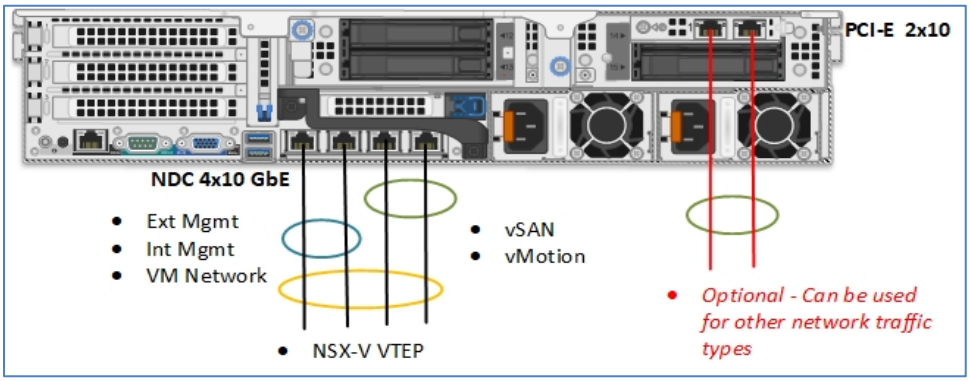

Figure 33 **VxRail 4x10 network profile with NSX-V (Mgmt or VI WLD)**

The VxRail can be deployed with a 2x25 profile on NDC with NSX-V (Mgmt or VI WLD) as shown in Figure 34. An additional PCI-E card can be installed to provide more connectivity for other traffic types.

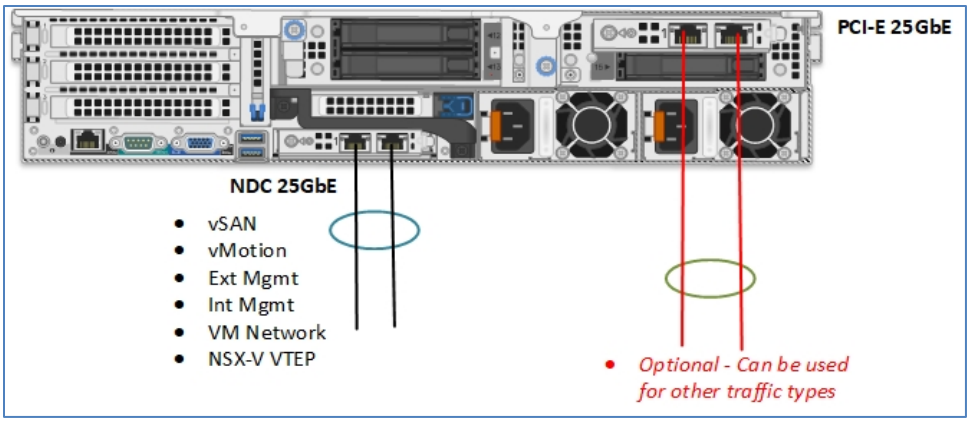

Figure 34 **VxRail 2x25 network profile on NDC with NSX-V (Mgmt or VI WLD)**

Note: VCF on VxRail does not support installing additional vDS. More physical interfaces must be added to the VxRail vDS to be used for additional traffic types.

7.4.2 NSX-T based VI WLD physical host connectivity options

This section illustrates the physical host network connectivity options for NSX-T based VI WLD using different VxRail profiles and connectivity options. Figure 32 illustrates a VxRail deployed with 2x10 profile on the 4-port NDC, the remaining two ports available after the deployment are used for NSX-T.

Note: For each cluster that is added to an NSX-T VI WLD, the user will have the option to select the two pNICs if there are more than two pNICs available. This can provide NIC redundancy if the pNICs are selected from two different NICs. Any subsequent nodes added to the cluster will use the same pNICs.

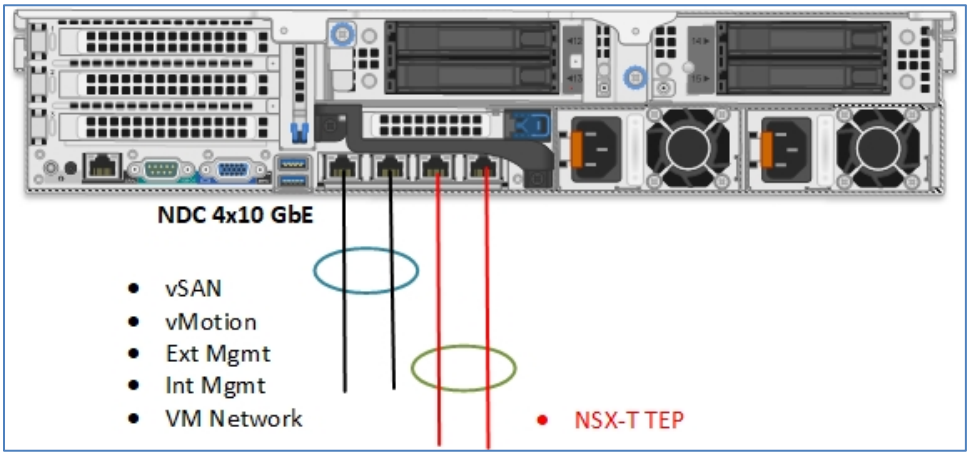

Figure 35 **VxRail 2x10 network profile with NSX-T VI WLD**

The next option is VxRail deployed with a 4x10 profile and NSX-T deployed onto the cluster. In this scenario, the NSX-T traffic uses the additional PCI-E card as shown in Figure 36, making this a 6xNIC configuration.

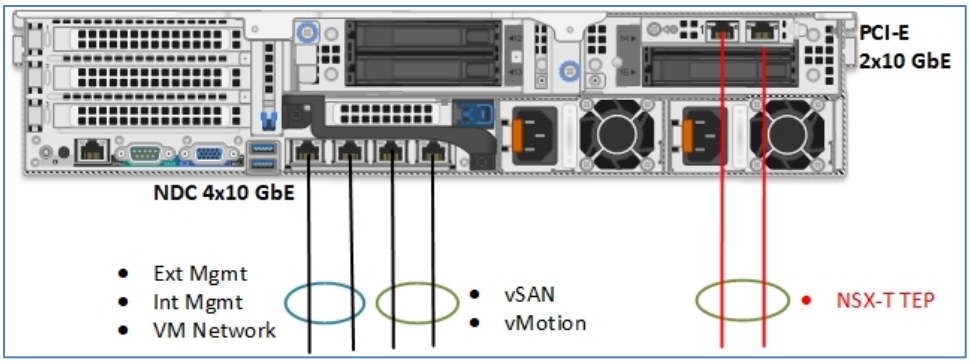

Figure 36 **VxRail 4x10 network profile with NSX-T VI WLD**

The final option that is covered here is a 2x25 profile for the VxRail using the 25GbE NDC and an additional 25GbE PCI-E. The VxRail system traffic uses the two ports of the NDC while the NSX-T traffic is placed on the two port PCI-E card.

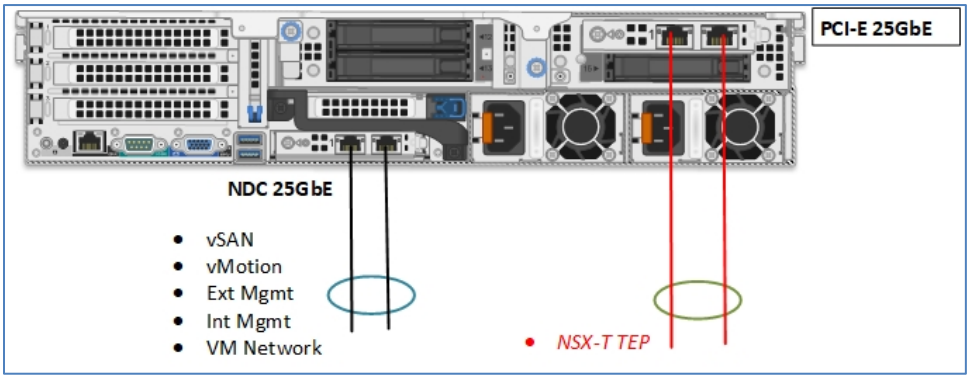

Figure 37 **VxRail 2x25 network profile with NSX-T VI WLD**

8 Multi-site design considerations

The VMware Cloud Foundation on VxRail solution natively supports a stretched-cluster configuration for the Mgmt WLD and VI WLDs between two availability zones. The stretched-cluster configuration combines both standard VxRail procedures and automated steps that are performed by using SOS commands from the SDDC Manager. The Witness deployment is manually deployed and configured, and the SDDC Manager automates the configuration of the vSAN stretched cluster.

Note: Dual Region disaster recovery is not yet supported in VCF version 3.9.1 on VxRail.

8.1 Multi-AZ (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. The following general requirements apply to a VMware Cloud Foundation on VxRail stretched-cluster deployments.

1. Witness deployed at a third site using the same vSphere version used in the VCF on VxRail release
2. All SC configurations must be balanced with the same number of hosts in AZ1 and AZ2.

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

The following network requirements apply for the Mgmt WLD and the VI WLD clusters that need to be stretched across the AZs:

- Stretched Layer 2 for the external management traffic.
- 5 millisecond RTT between data node sites
- Layer 3 between each data nodes site and Witness site
- 200 millisecond RTT between data node sites and the Witness site

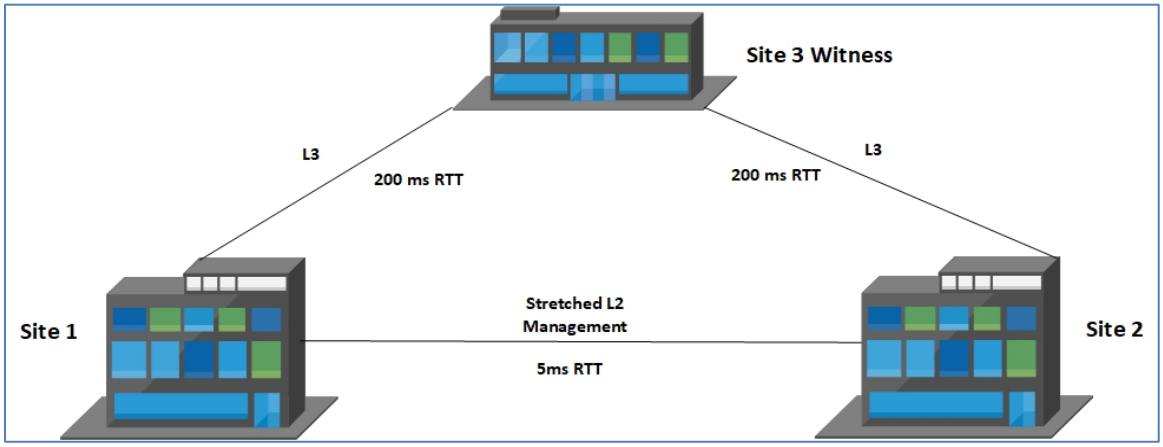

Figure 38 **Stretched Cluster Network Requirements**

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

8.1.1 NSX-V WLD

Table 8 shows the supported connectivity for the data nodes sites for the different traffic types for an NSX-V based WLD. This includes the management WLD and any NSX-V based VI WLD.

Traffic Type	Connectivity Options	Minimum MTU	Maximum MTU
vSAN	• L2 Stretched to AZ2 • L3 Routed to AZ2	1500	9000
vMotion	• L2 Stretched to AZ2 • L3 Routed to AZ2	1500	9000
External Management	• L2 Stretched to AZ2	1500	9000
VXLAN	• L2 Stretched to AZ2 • L3 Routed to AZ2	1500	9000
Witness vSAN	• L3 Routed to Witness Site	1500	9000

Table 8. NSX-V WLD Multi-AZ Connectivity Requirements

The vSAN and the vMotion traffic can be stretched Layer 2 or extended using Layer 3 routed networks. The external management traffic requires to be stretched Layer 2 only. This ensures the management VMs (VC/PSC/VxRM) do not need their IP addresses to be changed. It also ensures that no manual network reconfiguration is required when the VMs are restarted in AZ2 due to a site failure at AZ1. The VXLAN network uses one port group on the VxRail vDS that is connected to the nodes at AZ1 and AZ2. However, DHCP is used to assign the VTEP IPs for each host. DHCP enables you to stretch the VXLAN VLAN between sites and use the same subnet at each site for the VTEPs. Alternatively, you can use the same VLAN local at each site non-stretched, use a different subnet for the VTEPs, and route the VXLAN traffic between the two sites.

Note: The VXLAN VLAN ID is the same at each site whether using stretched Layer 2 or Layer 3 routed.

8.1.2 NSX-T WLD

Table 9 shows the supported connectivity for the data nodes sites for the different traffic types for an NSX-T WLD.

Traffic Type	Connectivity Options	Minimum MTU	Maximum MTU
vSAN	• L3 Routed	1500	9000
vMotion	• L2 Stretched • L3 Routed	1500	9000
External Management	• L2 Stretched	1500	9000
Geneve Overlay	• L2 Stretched • L3 Routed	1500	9000
Witness vSAN	• L3 Routed to Witness Site	1500	9000

Table 9. NSX-T WLD Multi-AZ Connectivity Requirements

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. The external management traffic must be stretched Layer 2 only. This ensures the VxRail Manager and the Edge Nodes can continue to have connectivity to the management network if they are restarted at AZ2 due to a site failover at AZ1. The Geneve overlay network can either use the same or different VLANs for each AZ so a single Geneve VLAN can be stretched Layer 2. 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.

Note: The vSAN traffic can only be extended using Layer 3 networks between sites. If only Layer 2 stretched networks are available between sites with no capability to extend with Layer 3 routed networks, an RPQ should be submitted.

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 VCF automation cannot be used for the stretched cluster configuration. It must be done manually using a standard VxRail SolVe procedure with some additional guidance.

8.1.3 Multi-AZ Component placement

During the 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 cluster of an NSX-T VI WLD.

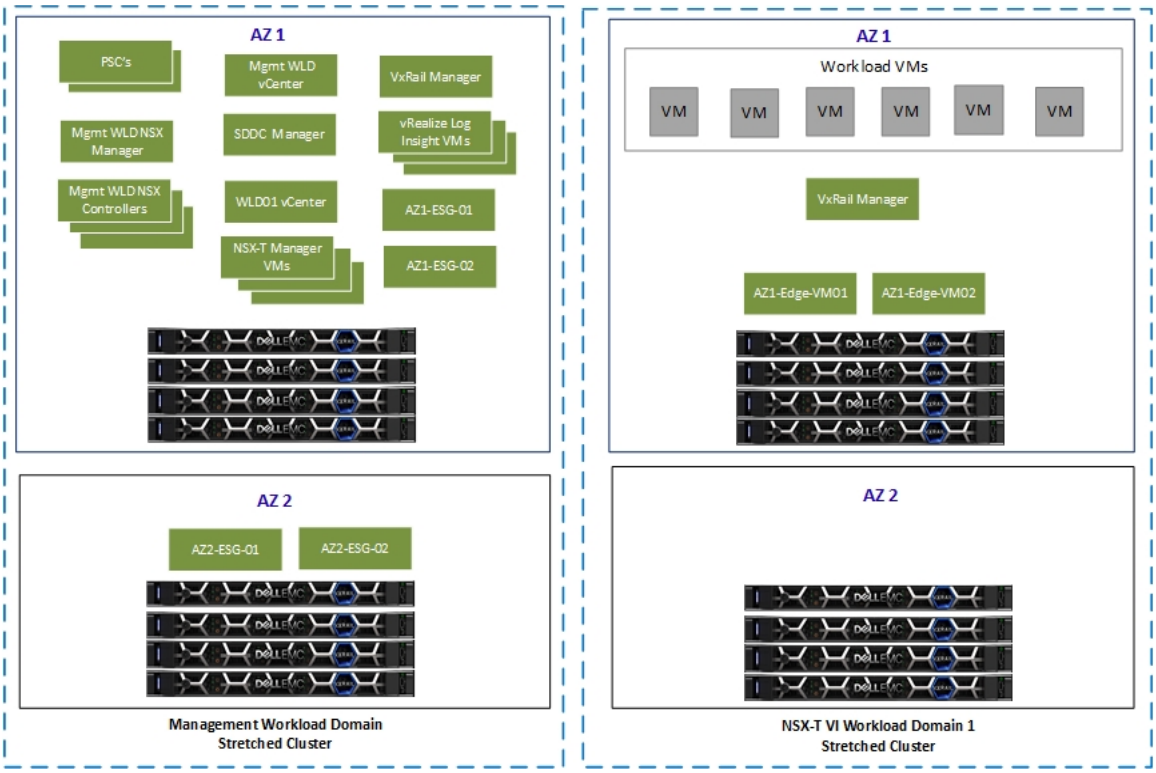

Figure 39 **Multi-AZ Component Layout**

8.1.4 Management WLD Multi-AZ – stretched cluster routing design

The management WLD multi-AZ aligns with the VVD reference architecture. This requires a manual Day 2 deployment using VVD documentation as a guide for the NSX ESGs at AZ2. The AVN network requires North–South routing at AZ2. If there is a failure at AZ1, the ESGs will need to be manually deployed and configured at AZ2. The physical network design varies depending on the existing network infrastructure.

Key points to consider for Multi-AZ management WLD routing design:

- Single management WLD vCenter/NSX Manager
- BPG used as the routing protocol for uDLR/ESG/TOR
- Site 1 ESGs deployed by Cloud Builder when management WLD is deployed
- Site 2 ESGs must be manually deployed.
- At Site 2 two uplink VLANs required for the ESGs
- Site 2 ESGs connect to the same Universal Transit Network as site 1 ESGs.
- Site 2 ESGs have the same ASN as Site 1 ESGs.
- Traffic flow from uDLR prefers Site 1 (primary) ESGs during normal operation; lower BGP weight is used between uDLR and ESGs at Site 2 for this purpose.
- BGP path prepend is needed on the Physical TORs to control Ingress traffic.
- Physical network design varies depending on existing network infrastructure.

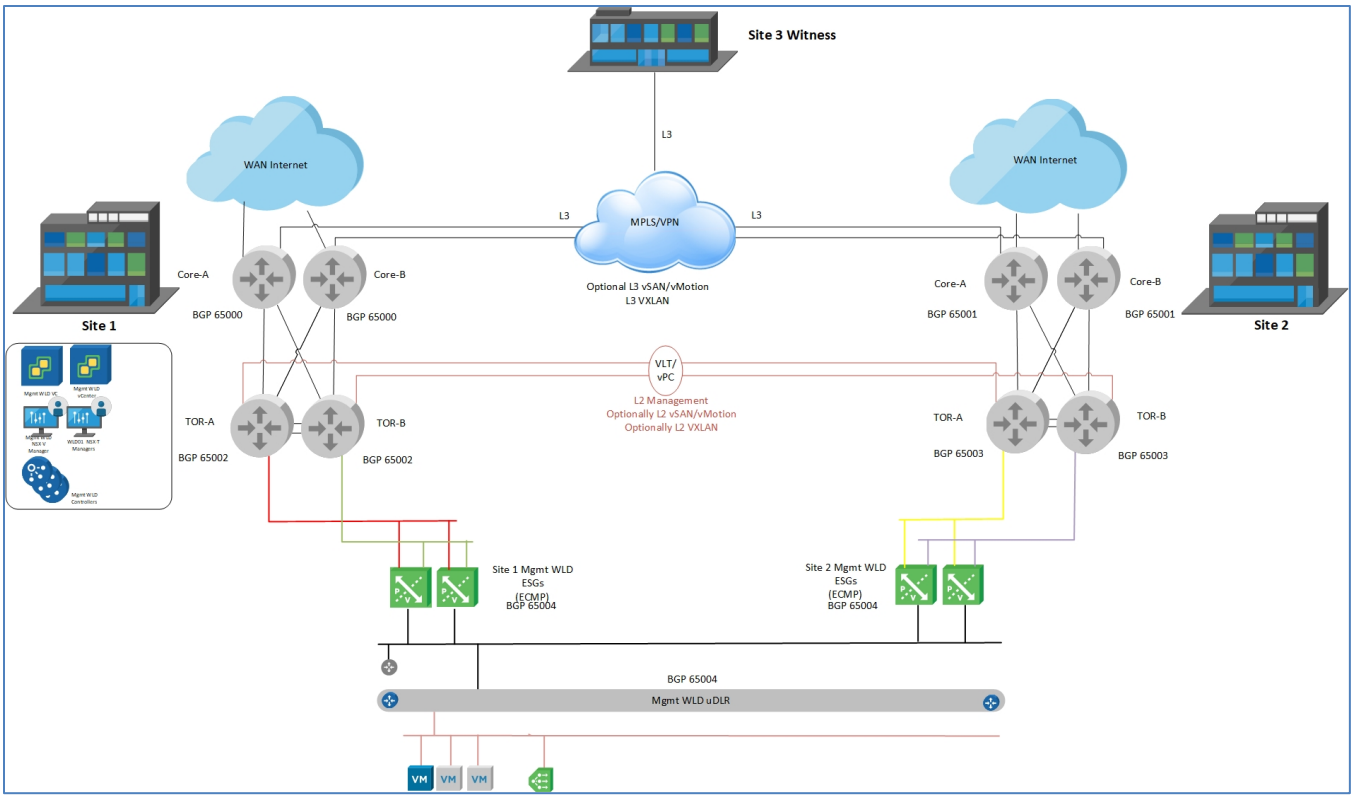

Figure 40 **Multi AZ – Mgmt WLD routing design example**

8.2 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 VMware Cloud Foundation 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 workload domain vCenters at each site. Following are the important factors to consider:

1. The vCenters for each VCF instance that participates in the same SSO Domain are connected using Enhanced Link Mode (ELM).
2. 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.
3. PSC replication configuration should be in a closed-loop design.
4. Recommended Site1 PSC1 > Site1 PSC2 > Site2 PSC1 > Site2 PSC2 > Site1 PSC1
5. Manual configuration is required to point Site2 PSC2 back to Site1 PSC.

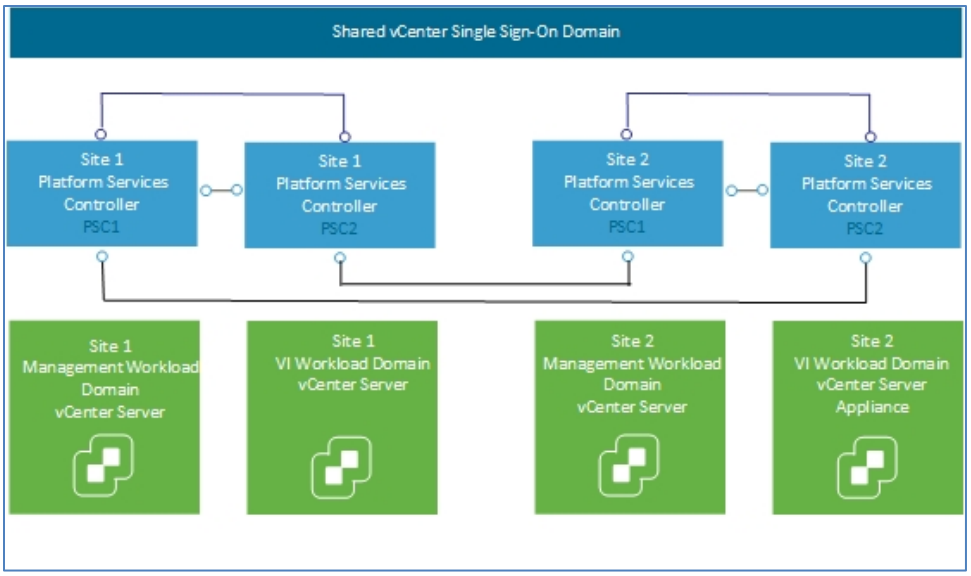

Figure 41 **Shared SSO Domain topology for two VCF Instances**

8.2.1 Upgrade Considerations

There are some factors that must be considered when it comes to upgrading the PSCs in 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:

1. 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.
2. The compatibility rules in VCF LCM do not extend to external VCF instances.

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

9 Operations Management Architecture

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. They are described in more detail in this section.

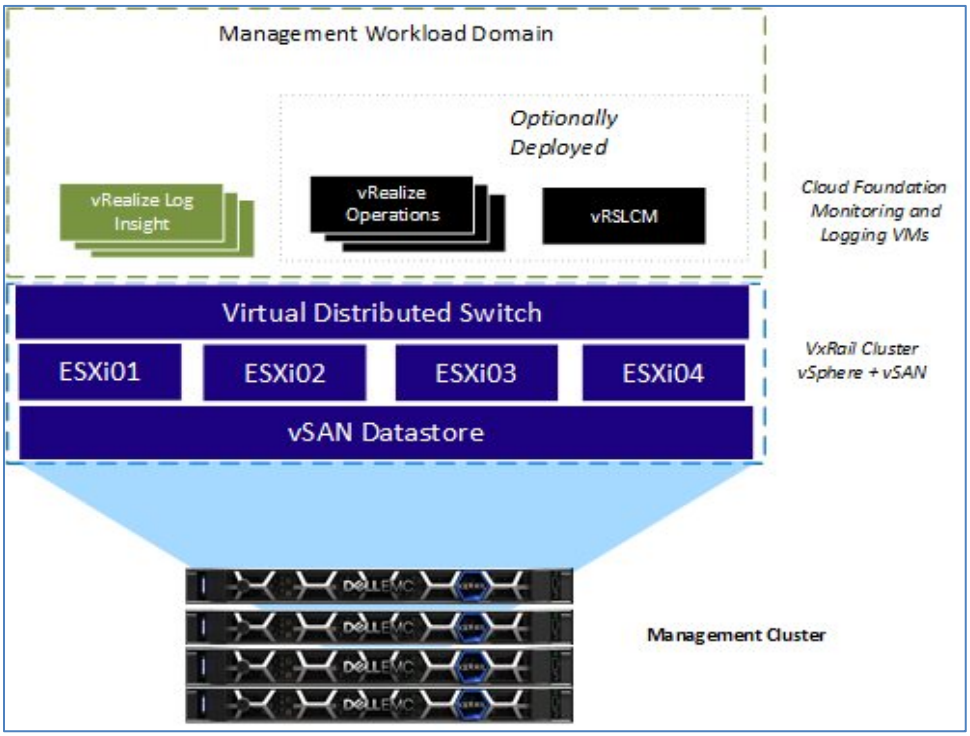

Figure 42 **Monitoring and Logging Operations**

9.1 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.

9.2 vRealize Operations

VMware vRealize Operations provides self-driving operations from applications to infrastructure to optimize, plan, and scale SDDC and multi-cloud deployments. This highly scalable, extensible, and intuitive operations platform automates and centralizes management for SDDC and cloud, delivering continuous performance optimization based on intent, efficient capacity management, proactive planning, and intelligent remediation. vRealize Operations Manager provides operations dashboards to gain insights and visibility into the health, risk, and efficiency of your infrastructure, performance management, and capacity optimization capabilities. vRealize Operations is an optional component that is deployed from the SDDC Manager UI. Before you can deploy vRealize Operations or vRealize Automation, you must deploy vRealize Suite Lifecycle Manager.

9.3 vRealize Log Insight

VMware vRealize Log Insight delivers automated log management through log aggregation, analytics, and search capabilities with an integrated cloud operations management approach. It provides the operational intelligence and enterprise-wide visibility that is required to enable service levels proactively and operational efficiency in dynamic hybrid cloud environments. vRealize Log Insight is deployed in the Mgmt WLD during the VCF bring-up process. It consists of three VMs deployed in a cluster, and the internal Load Balancer feature enabled.

10 Lifecycle Management

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 cluster. The VxRail hardware and software life cycle are orchestrated from the SDDC Manager. The underlying upgrade process for each cluster is managed by VxRail Manager to upgrade the hardware, firmware, and the vSphere ESXi and vSAN.

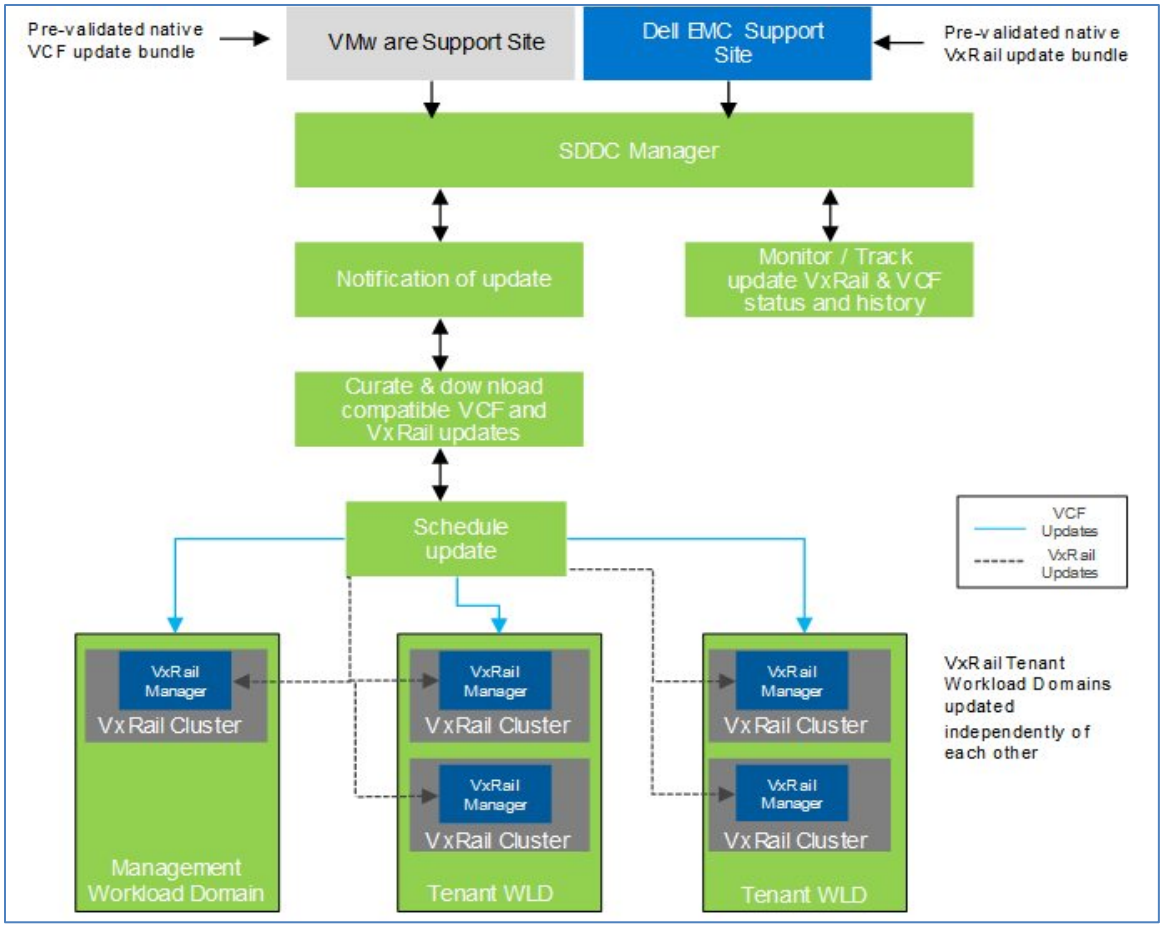

Figure 43 **VCF on VxRail LCM Components**

Credentials for a My VMware account **and** a Dell EMC Support account must be provided for the LCM process to download the appropriate upgrade bundles. VMware and Dell EMC validate updates and distribute them using native VCF and Dell EMC 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.

10.1 vRealize Suite Lifecycle Manager

The VMware vRealize suite Lifecycle Manager automates the LCM of the vRealize suite. It must be deployed before any 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.

11 Cloud Management Architecture

vRealize Automation provides self-service provisioning, IT services delivery, and life cycle management of cloud services across many multi-vendor virtual, physical, and cloud platforms using a flexible distributed architecture. vRealize Automation provides a secure portal where authorized administrators, developers, and business users can request new IT services and manage existing system resources from predefined user-specific service catalog. The two main functional elements of the architecture are the vRealize Automation appliance and the IaaS components.

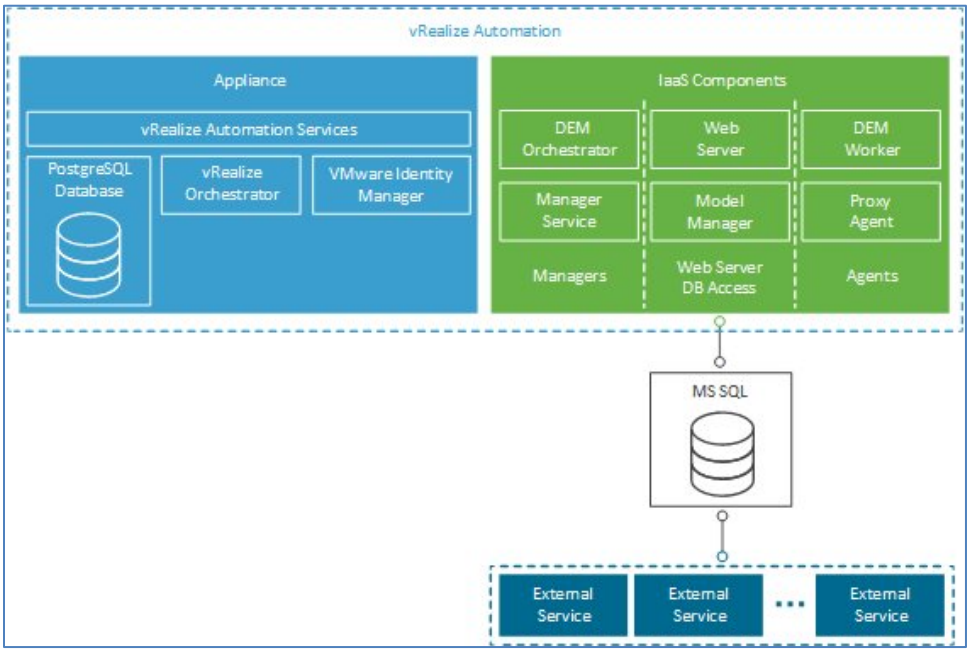

Figure 44 **vRealize Automation Components**

Before deploying the vRealize Automation suite from SDDC Manger, the vRealize Lifecycle Manger must be deployed from SDDC Manager. It is used to deploy and Lifecycle the vRealize Suite components. There are additional requirements for an external Microsoft SQL server and an IaaS Windows server OVA Template. Read the Requirements for vRealize Automation documentation for all the requirements that must be met before a vRealize Automation deployment can be initiated from SDDC Manager.

Note: The vRealize suite is deployed to a VLAN backed network. If these management components are going to be protected in a multi-site DR configuration, you must migrate the networking to NSX logical switches. This might also be desirable for a multi-site with stretched cluster.