VMware Cloud Foundation on VxRail Architecture Guide

VCF 5.2 on VxRail

December 2024

H04423

White Paper

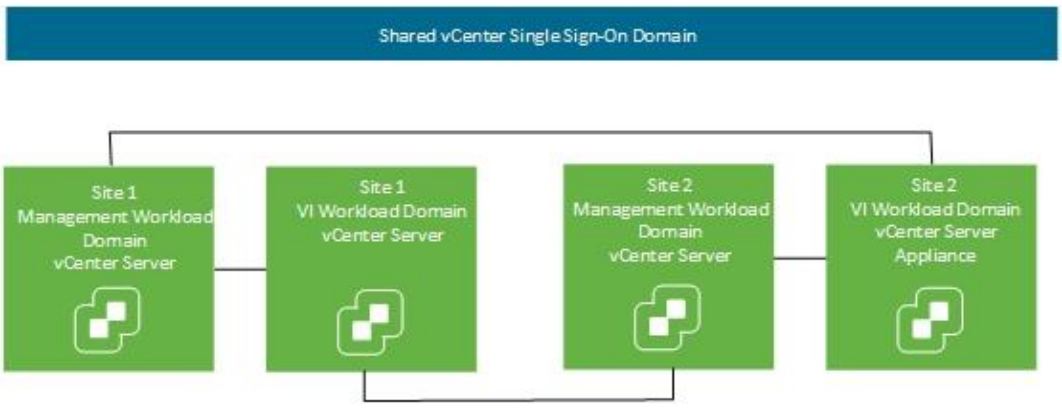
Abstract

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

DELL Technologies

Copyright

Contents



Executive summary

VMware Cloud Foundation on VxRail

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

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

Document purpose

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

Audience

This 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 business requirements. Readers should be familiar with the VMware vSphere, NSX, vSAN, and Aria product suites in addition to general network architecture concepts.

Revisions

Date	Part number/ revision	Description
December 2024		VCF 5.2.1.1
April 2024	H19988	Initial release of VCF 5.1

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: David Ring

Contributors: Martin Rolfe, Dave Ryan.

Note: For links to other documentation for this topic, see the VxRail Info Hub.

Architecture overview

Introduction

You can virtualize all your infrastructure and deploy a full VMware SDDC with the benefit of automated SDDC life-cycle management (LCM) by implementing a standardized VMware SDDC architecture on VxRail with Cloud Foundation. This solution includes NSX for Network Virtualization and Security, vSAN for SDS, VMware vSphere 8 for Kubernetes, Tanzu Kubernetes Grid, VMware Private AI Foundation with NVIDIA for AI workloads and SDDC Manager for SDDC LCM.

By virtualizing all your infrastructure, you can take advantage of what a fully virtualized infrastructure can provide, such as resource utilization, workload and infrastructure configuration agility, and advanced security. With SDDC software life-cycle automation provided by VCF (and in particular SDDC Manager, which is a part of VCF 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 manually performing updates and upgrades using multiple tools for all the SDDC SW and HW components of the stack. These processes are now streamlined using a common management toolset in SDDC Manager with VxRail Manager. You can begin to take advantage of 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 such as microsegmentation, which 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 VCF, 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 be assured that there is an automated and validated path forward to get from one known good state to the next across the end-to-end stack.

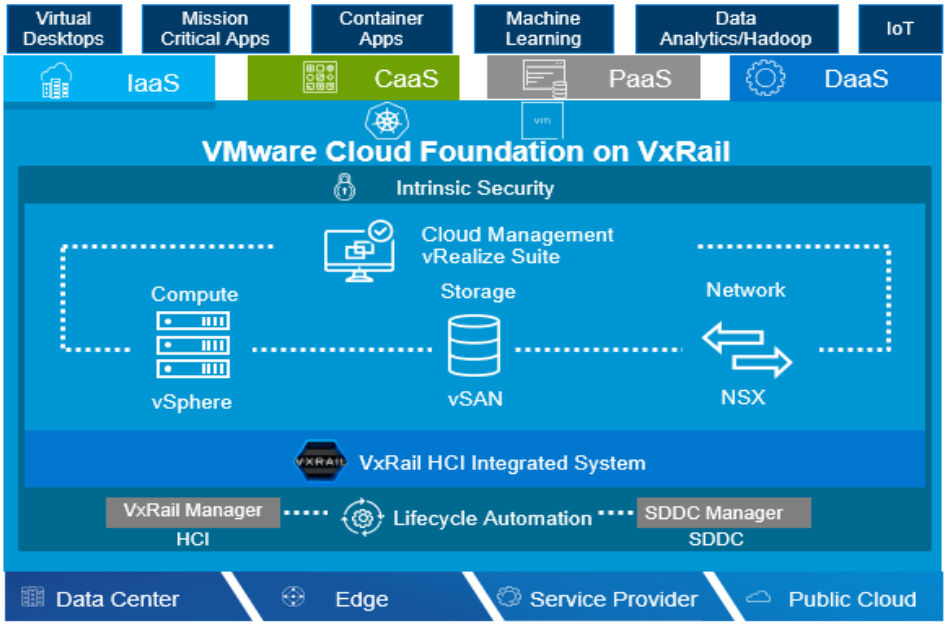


Figure 1. Architecture overview

VxRail Manager

VCF on VxRail uses VxRail Manager to deploy and configure VxRail clusters that are powered by vSAN or external storage. It is also used to perform the LCM of VMware ESXi, vSAN, and hardware 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 unique turnkey hybrid cloud experience not available on any other infrastructure.

VxRail Manager provides the glue for the HCI hardware and software and is all life-cycle-managed together. Focusing on the glue and automation across the deployment, updating, monitoring, and maintenance phases of the 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 them having to spend time maintaining the infrastructure.

SDDC Manager

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

Network virtualization

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

NSX Data Center reproduces the entire network model in software, enabling any network topology—from simple to complex multi-tier networks—to be created and provisioned in seconds. You can create multiple virtual networks with diverse requirements, using a combination of the services that NSX offers. These services include microsegmentation and 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.

VMware Av Load Balancer (formerly known as NSX Advanced Load Balancer) allows you to implement centrally managed distributed load balancing for your application workloads within VMware Cloud Foundation and configure enterprise grade load-balancing, global server load balancing, application security, and container ingress services.

Starting with VMware Cloud Foundation 5.2, you can use SDDC Manager to deploy Avi Load Balancer as a high availability cluster of three VMware Avi Controller instances, each running on a separate VM.

Operations management

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

Logging and analytics

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

Self-service cloud

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

Workload domain architecture

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

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

Two types of WLDs can be deployed:

- A Management WLD (Mgmt WLD), single per VCF instance

- A Virtual Infrastructure (VI) WLD, also known as a tenant WLD

The following section provides more detail about each type of WLD.

Management WLD

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

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

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

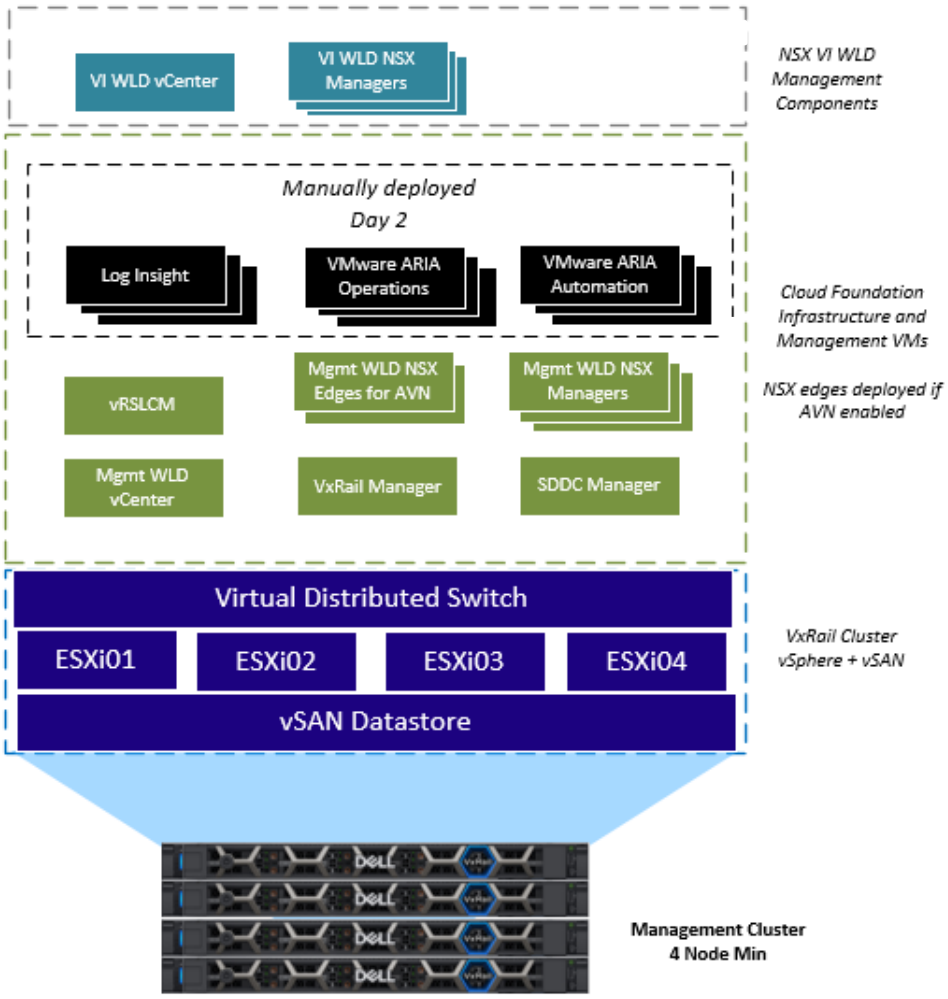


Figure 2. Management domain components

vCenter design

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

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

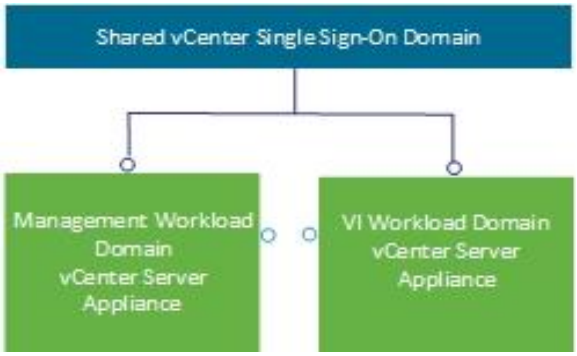


Figure 3. vCenter design

VI workload domain

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

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

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

You can dedicate a VxRail cluster for only the Edge node components if either dedicated compute or bandwidth is required for the Edge node cluster. Bare-Metal NSX Edge nodes are not supported.

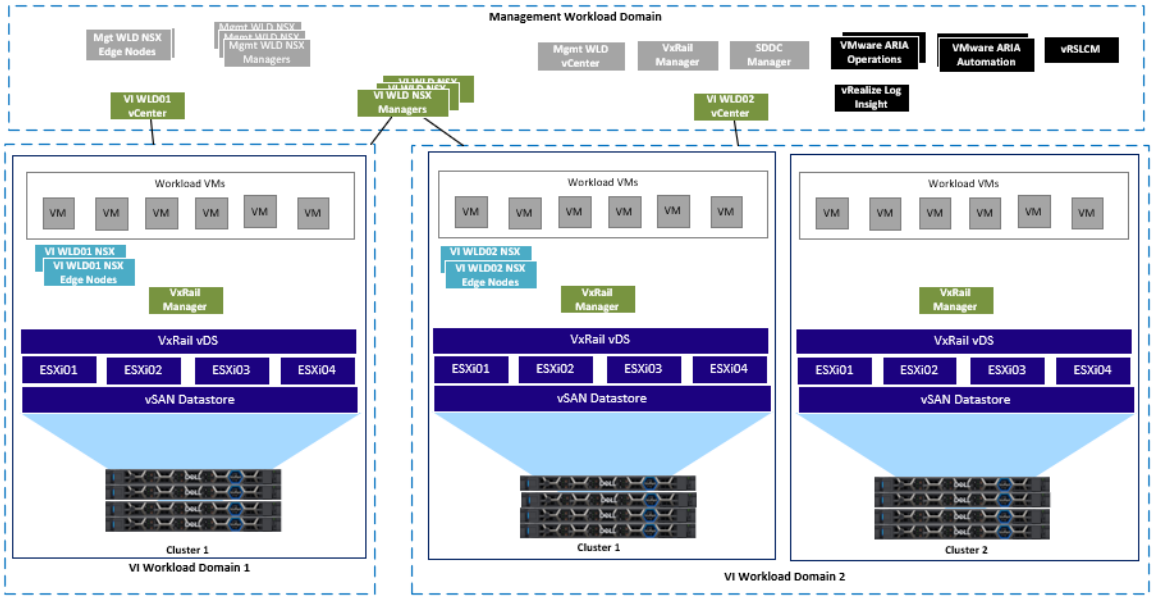


Figure 4. VI WLD component layout with NSX everywhere

vCenter design

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

Isolated SSO workload domains

Isolated SSO WLDs give administrators the option to configure new WLDs using a separate Single Sign On (SSO) instance. This scenario is useful for large enterprises that need workload isolation and for Managed Service Providers (MSPs) who can allocate WLDs to different tenants with their own SSO domains. Isolated SSO domains are each configured with their own NSX instance. The added benefit is that configuring WLDs as an isolated WLD also allows the option to configure a separate identity provider (Active Directory or LDAP).

- WLD scaling also increases from 15 to 25 WLDs when using isolated WLDs within a single VCF instance. Note that WLDs configured to use the shared management domain SSO are still limited to a maximum of 15 domains.

Consolidated architecture

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

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

- The conversion of consolidated to standard requires a new VI WLD domain to be created. The tenant workloads must be migrated to the new VI WLD. The recommended method for this migration is to use HCX.
- Use cases that require a VI WLD to be configured to meet specific application requirements cannot run on a consolidated architecture. The singular management

WLD cannot be tailored to support management functionality and these use cases. If your plans include applications that require a specialized VI WLD (such as Horizon VDI or PKS), plan to deploy a standard architecture.

- Life-cycle management can be applied to individual VI WLDs in a standard architecture. If the applications targeted for VCF on VxRail have strict dependencies on the underlying platform, consolidated architecture 4 is not an option.
- Autonomous licensing can be used in a standard architecture, where licensing can be applied to individual VI WLDs. In a consolidated architecture, autonomous licensing is not an option.
- Scalability in a consolidated architecture has less flexibility than a standard architecture. Expansion is limited to the underlying VxRail cluster or clusters supporting the single management WLD because all resources are shared.
- If a VxRail cluster was built using two network interfaces, consolidating VxRail traffic and NSX traffic, additional nodes are limited to two Ethernet ports being used for VCF for VxRail.

Remote VxRail clusters

With Remote Clusters, you can deploy a VI WLD that has its VMware vSphere cluster at a remote location. You can also enable VCF with Tanzu on a cluster deployed at a remote site. All the VCF operational management can be administered from the central or regional data center out to the remote sites, which is important because:

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

The following figure illustrates a VCF Deployment with two VI WLD Domains, one local and the other remote which is located at a single Edge site where the VxRail cluster is deployed:

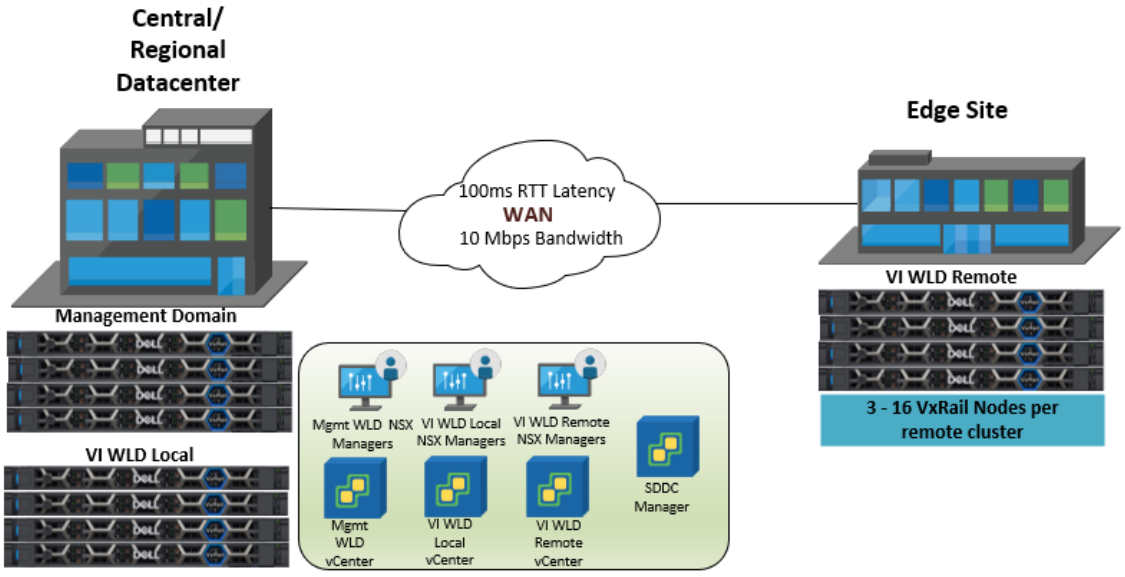


Figure 5. Remote VxRail cluster deployment

Remote VxRail cluster deployments

The following requirements must be met for remote VxRail cluster deployments. Failure to adhere to these requirements will lead to system integrity, instability, resiliency, and security issues of the Edge workload.

- 10 Mbps bandwidth.
- 100-millisecond RTT latency.
- Support for three to 16 nodes per remote site per VI WLD domain. VI WLD domain can include a local cluster or a remote cluster but not both.
- VCF 5.1 supports a single VCF instance with up to eight VI WLDs with single remote clusters.
- Primary and secondary active WAN links (not required, but highly recommended).
- DNS and NTP Server available locally or reachable from central site to Edge site.
- A DHCP server that is available for the NSX host overlay (Host TEP) VLAN of the WLD. When NSX creates Edge Tunnel End Points (TEPs) for the VI WLD, the TEPs are assigned IP addresses from the DHCP server. The DHCP server should be available locally at the Edge site.

Remote VxRail cluster network design

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

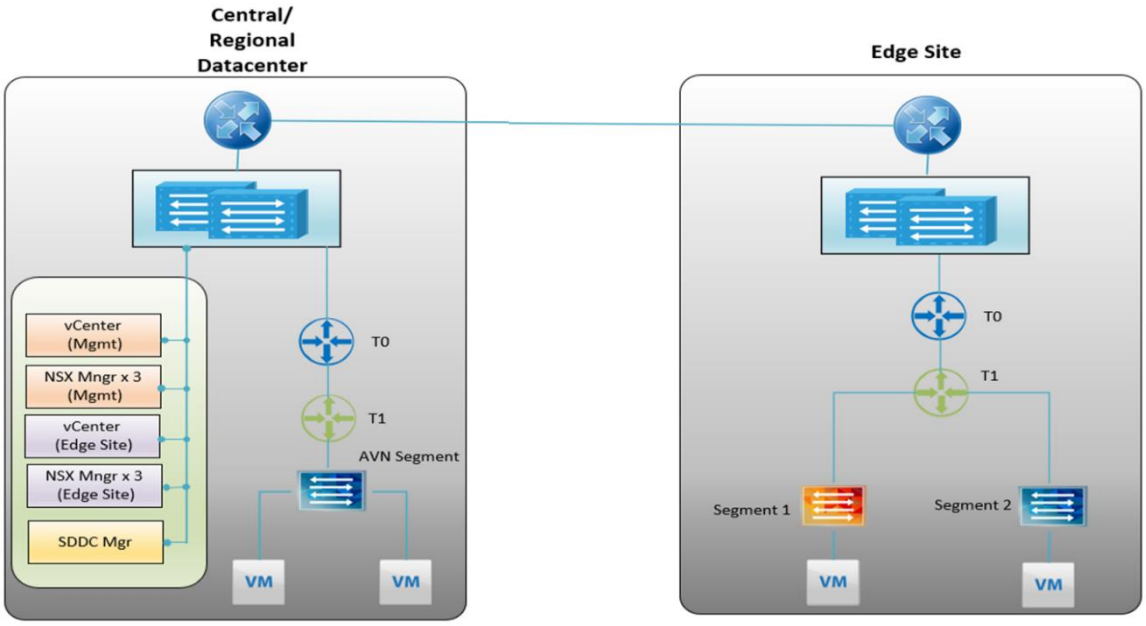


Figure 6. Remote VxRail cluster network design

Physical WLD layout

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

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

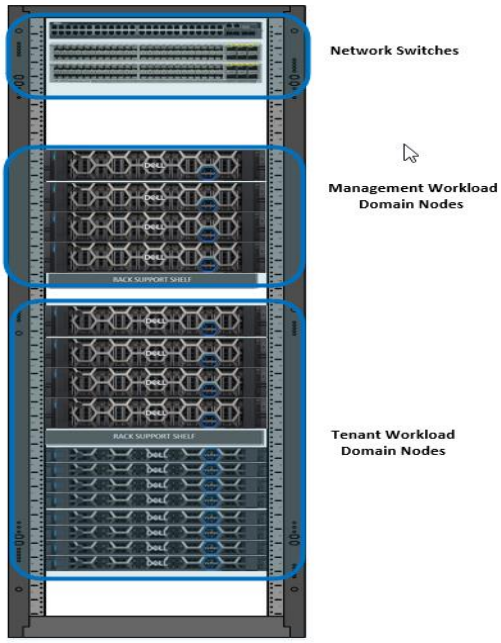


Figure 7. Single-rack WLD mapping

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 that needs redundancy might have to be stretched across multiple adjacent racks, as shown in the following figure:

Figure 8. WLDs spanning racks

VxRail hardware options

VCF on VxRail Supports 14G/15G/16G VxRail platforms. See the VCF on VxRail Support Matrix for detailed information.

Depending on the customer workload and application requirements, the correct VxRail hardware platform must be selected. Work with your Dell account representative for guidance on sizing for the SDDC components for the different VxRail hardware platforms.

VxRail virtual network architecture

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

VCF 5.1 on VxRail 8.0.200 enables traffic isolation between management VMs and ESXi Management VMkernel Interfaces. This feature enables end users to configure different subnets/VLANs between ESXi mgmt DvpPG and mgmt VM DvPG. Before this release, the default networking topology deployed by VCF on VxRail comprised of the ESXi host management interfaces (VMkernel interface) and management components (vCenter server, SDDC Manager, NSX components, VxRail Manager, so forth) being applied to the same distributed virtual port group (DvPG).

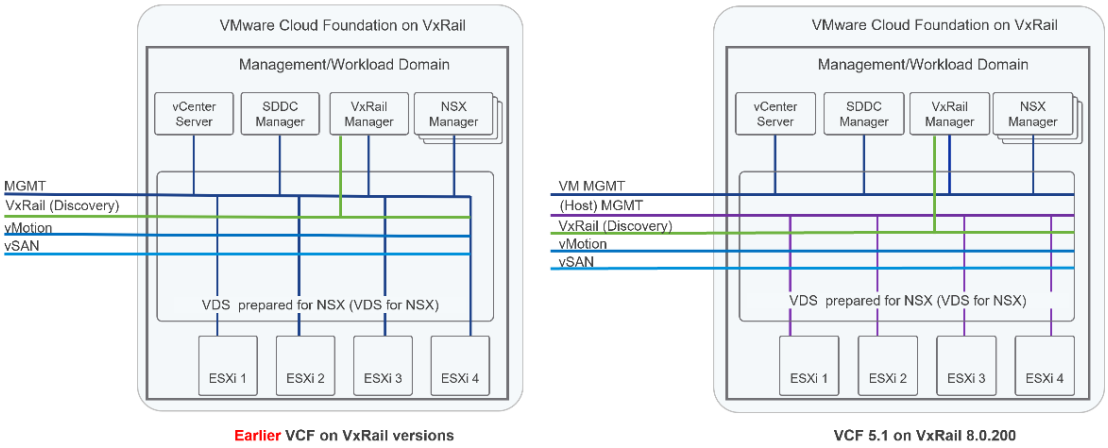


Figure 9. Traffic isolation

Separation of DVPG for management appliances and ESXi host management prior to and from 5.1 release. Note that in VxRail the default port groups are labeled as follows:

- **'Management Network'** DvPG' = ESXi Management VMkernel Interfaces.
- **'vCenter Server Network'** DvPG = vCenter server, VxRail Manager.

VxRail virtual distributed switch system (vDS)

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

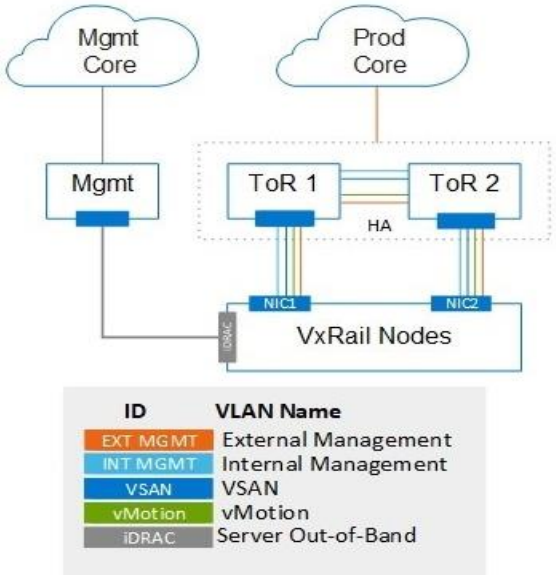


Figure 10. VxRail cluster VLANs

VxRail vDS NIC teaming

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

VxRail predefined profiles

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

Table 1. Predefined 2x10 or 2x25 GbE profile

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

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

Table 2. Predefined 4x10 profile

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

Finally, a 4x25 profile is available with the following network layout:

Table 3. Predefined 4x25 profile

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

VxRail vDS custom profiles

With custom profiles, you can essentially select what uplinks/VMNICs pairings to use for each type of system traffic. For more details about custom profiles, see **Error! Reference source not found.**.

Additional VCF NSX networks

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

Note: The Edge cluster for the management WLD is deployed as a Day-2 operation from the SDDC Manager.

Table 4. VCF VLANs for management WLD deployment

Network traffic	Sample VLAN
NSX Host TEP	103
NSX Edge TEP	104 (Can be the same as Host TEP)
Edge Uplink01	105
Edge Uplink02	106

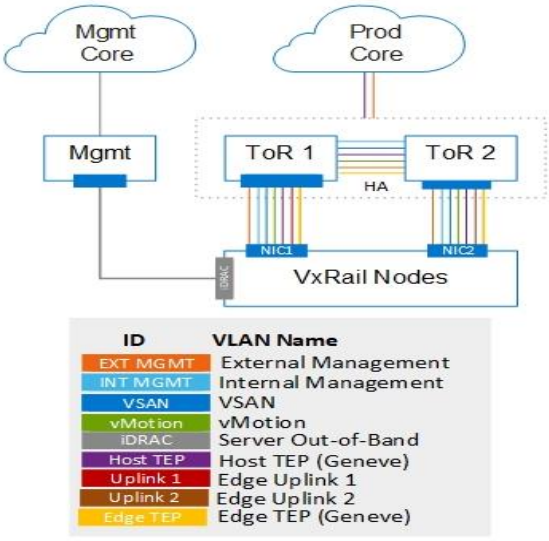


Figure 11. VxRail Management WLD VxRail cluster VLANs (uplink and Edge with AVN enabled)

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

Table 5. VCF VLANs for VI WLD deployment

Workload domain type	Network traffic	Sample VLAN
VI WLD	NSX Host TEP	203
VI WLD (Edge deployment only)	NSX Edge TEP	204 (Can be same as Host)
VI WLD (Edge deployment only)	Edge Uplink01	205
VI WLD (Edge deployment only)	Edge Uplink02	206

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

VxRail vDS with NSX networks

VxRail vDS and predefined network profiles

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

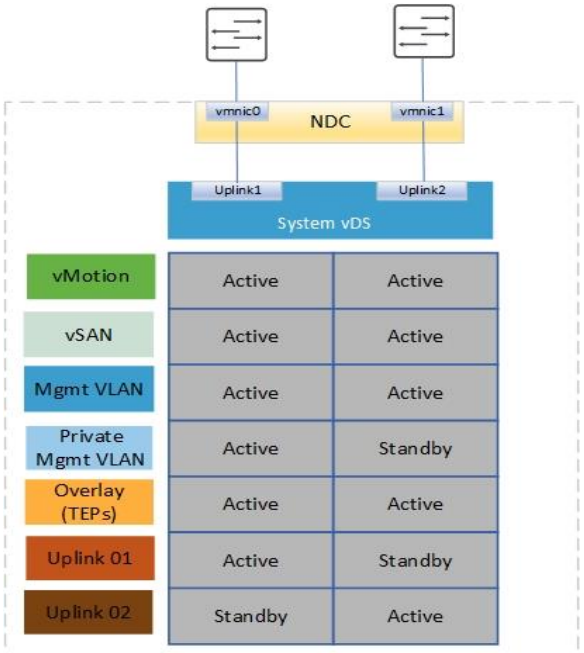


Figure 12. Single vDS using 2x10/2x25 predefined network profile

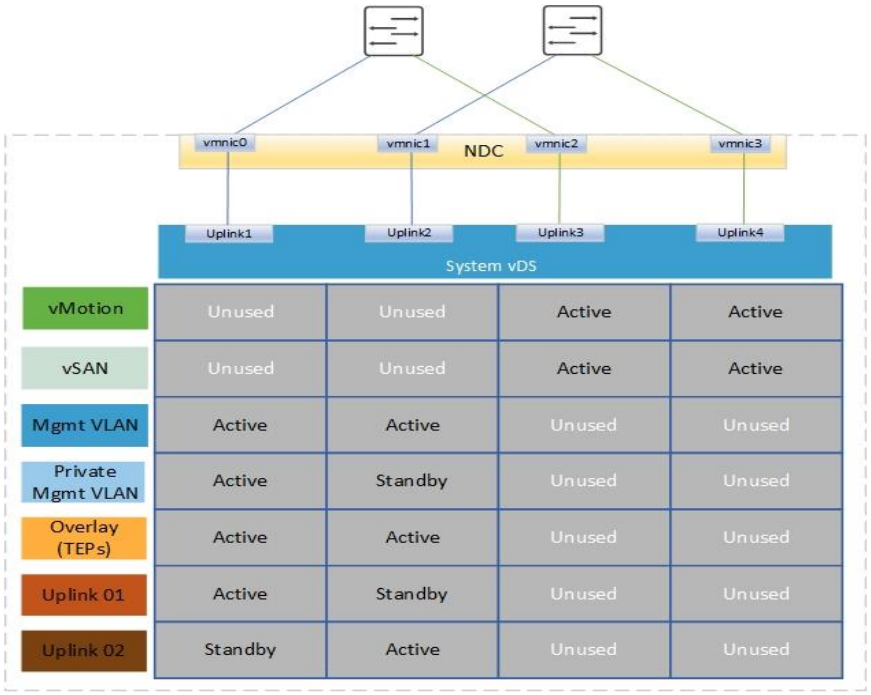


Figure 13. Single vDS with 4x10 predefined network profile

The following figure illustrates the 4x25 profile. This profile uses both an NDC/OCP and PCIe to achieve NIC-level redundancy for the system traffic. This profile is not recommended because it results in a nonstandard wiring configuration, as shown in the figure.

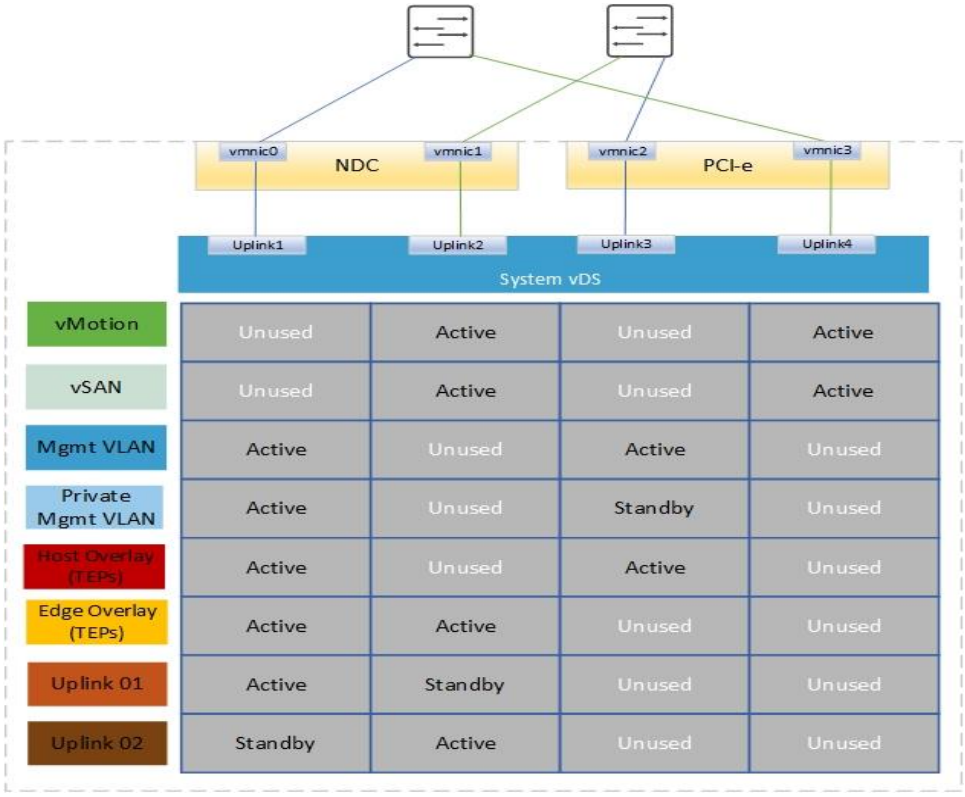


Figure 14. Single vDS with 4x25 predefined network profile

VxRail vDS and custom profiles

Notice that in the previous figure, the cabling of the NIC ports to the switches is a little unorthodox. We normally have VMNIC2 going to Fabric A and vminc3 going to Fabric B.

Note: user may select the VMNICs used for external Mgmt, vSAN, or vMotion to be used for the host TEP traffic and Edge traffic.

Table 6. The recommended method to achieve NIC-level redundancy for a VCF on VxRail cluster with 4x25 GbE is to configure the custom profile for the VxRail vDS using the configuration of the VMNIC/uplink mappings and the uplink to port group mapping shown in the following two tables.
VxRail vDS uplink to pNIC mapping

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

Table 7. VxRail vDS port group uplink mapping

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

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

The configuration of the vDS and uplink to pNIC mapping that is shown in the following figure provides NIC-level redundancy for a 4x25 GbE deployment using the VxRail custom profiles.

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

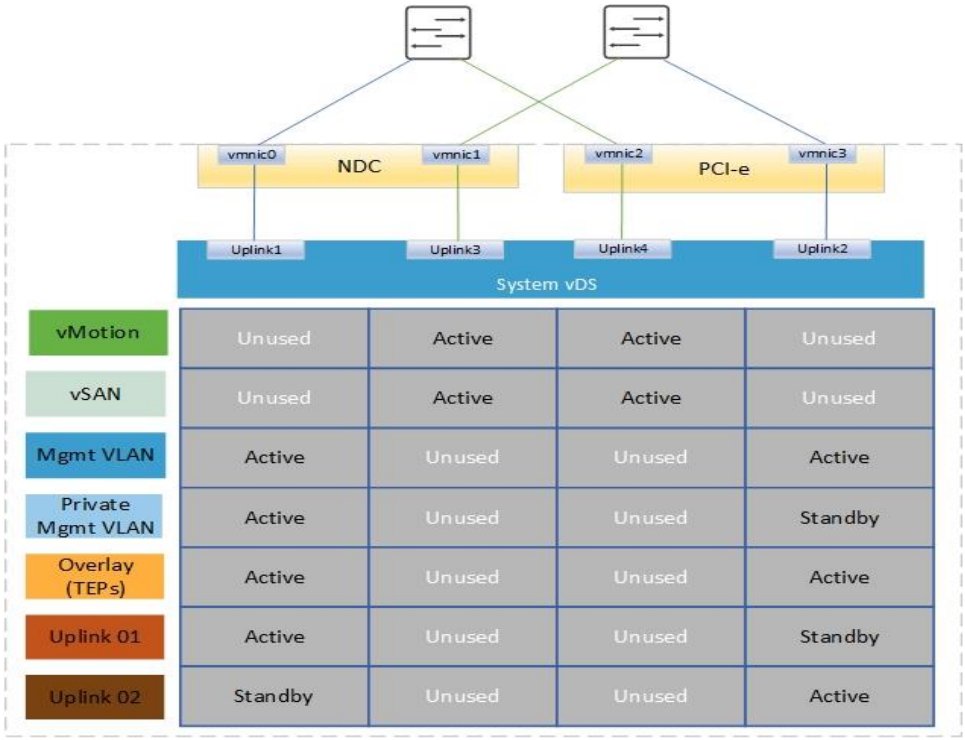


Figure 15. Single vDS with 4x25 using custom profile

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

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

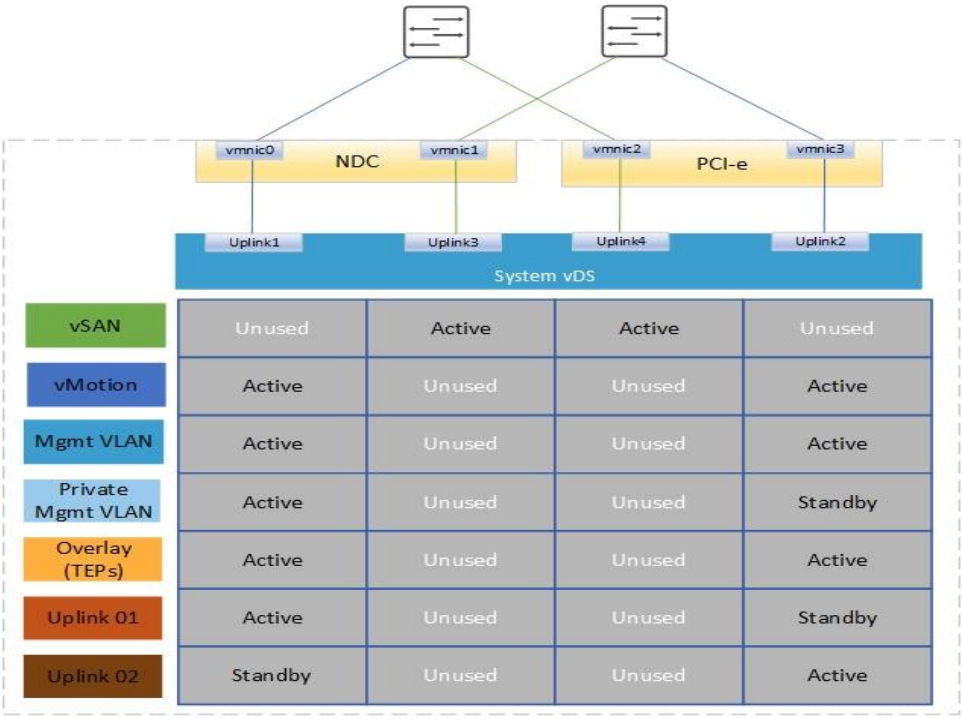


Figure 16. Single vDS using custom profile with NIC-level redundancy for dedicated vSAN

NSX vDS

VCF 5.1 provides the ability to fully separate VxRail system traffic and NSX traffic for both the management WLD and VI WLD using a secondary system vDS along with a third vDS which can be dedicated for NSX traffic. For the management WLD, additional inputs would be required in the deployment parameter sheet. Cloud Builder can create a third vDS during VCF bring-up allowing the ability to have two system vDS and an NSX dedicated overlay vDS. When deploying CB using a two system vDS VxRail Manager should create the system vDSs during the first run. With VCF 5.1 each vDS created must have a Transport Zone Type assigned to it defined in the input parameter sheet. Zone type can be either VLAN, Overlay or both. This supports new functionality allowing NSX DFW rules to be applied to the System VLAN Portgroups. In VCF 5.1 the VI WLD, SDDC Manager UI now supports creating VxRail VI workload domains using an advanced NIC profile allowing a total of three vDSs per cluster. Between the three only a single vDS can support NSX overlay traffic. Workflow Optimization support now has feature parity between WFO UI and WFO script, as shown in the following table.

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

Table 8. Workflow optimization support

Feature	WFO UI*	WFO Script
Single VDS (predefined network profiles)	✓	✓
Separate overlay VDS	✓	✓
Multiple system VDSs	✓	✓
Advanced VxRail VDS (NIC mapping, uplink PG mapping...)	✓	✓
FC Storage	✓	✓
Custom VDS/PG names	✓	✓
Isolated WLD Domains	✓	✓
DVPG (mgmt vmk separation from VM mgmt)	✓	✓

Secondary System and NSX network topologies

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

Two vDS (system and NSX) – 4pNIC topology

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

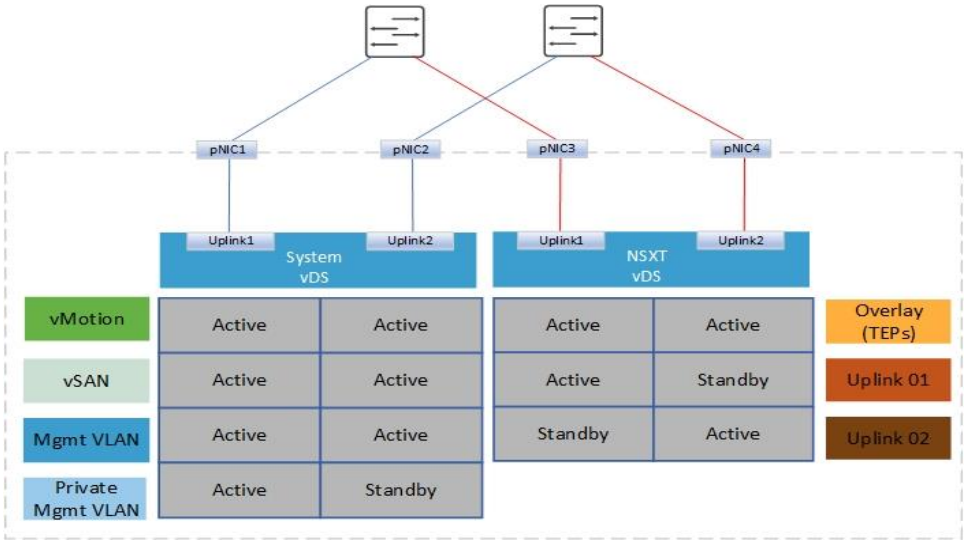


Figure 17. Two vDS with four pNICs

If we now consider the two-vDS design and NIC-level redundancy, the VxRail vDS must be deployed using a custom profile using uplink1/uplink2 for all traffic. Uplink 1 must be

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

Note: Both ports from NDC must connect to switch A, and both ports from the PCIe must connect to switch B. These connections are required for VCF VMNIC lexicographic ordering.

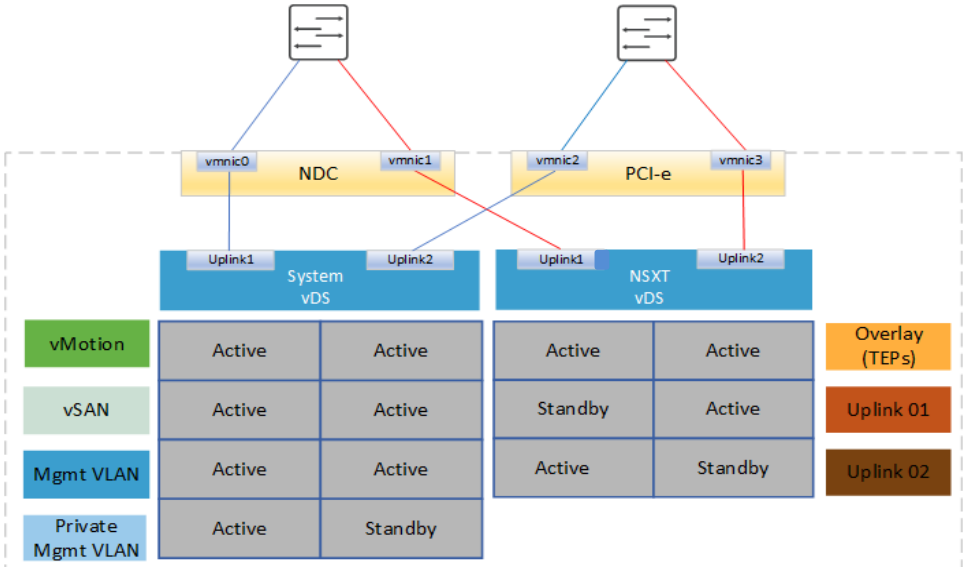


Figure 18. Two vDS with custom profile and NIC-level redundancy

Two vDS (system and NSX) – 6pNIC topologies

There are two options in a 6-pNIC design with two vDS. For the first option, we have four pNICs on the VxRail vDS and use two additional pNICs dedicated for NSX traffic on the NSX vDS. This option might be required to keep Mgmt and vSAN/vMotion on different physical interfaces and also if NSX needs its own dedicated interfaces.

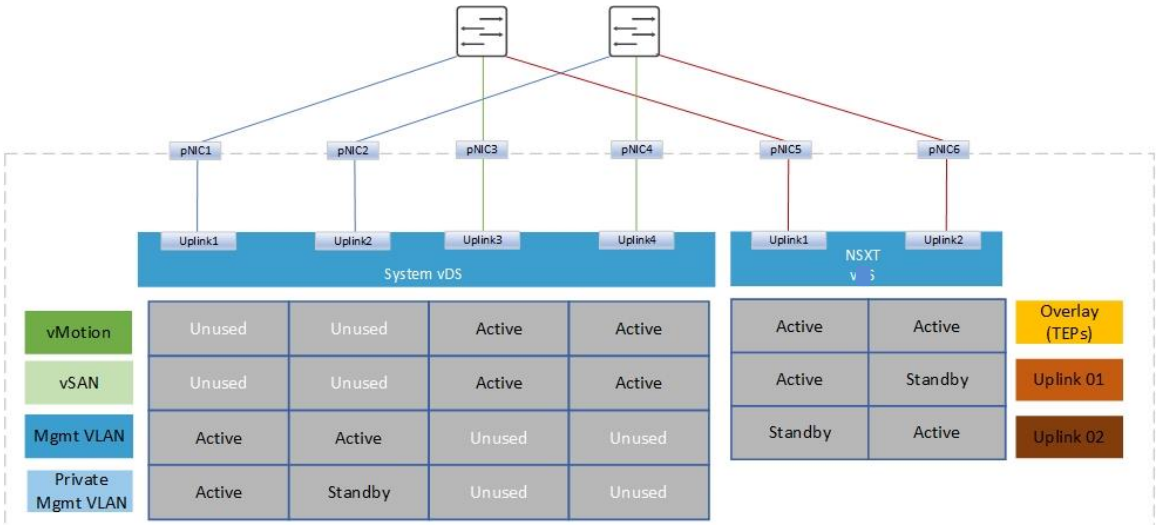


Figure 19. Two vDS with six pNICs option 1

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

Note: Admin can select which uplinks to use on the vDS for the Edge VM traffic when a dedicated NSX vDS has been deployed with four uplinks.

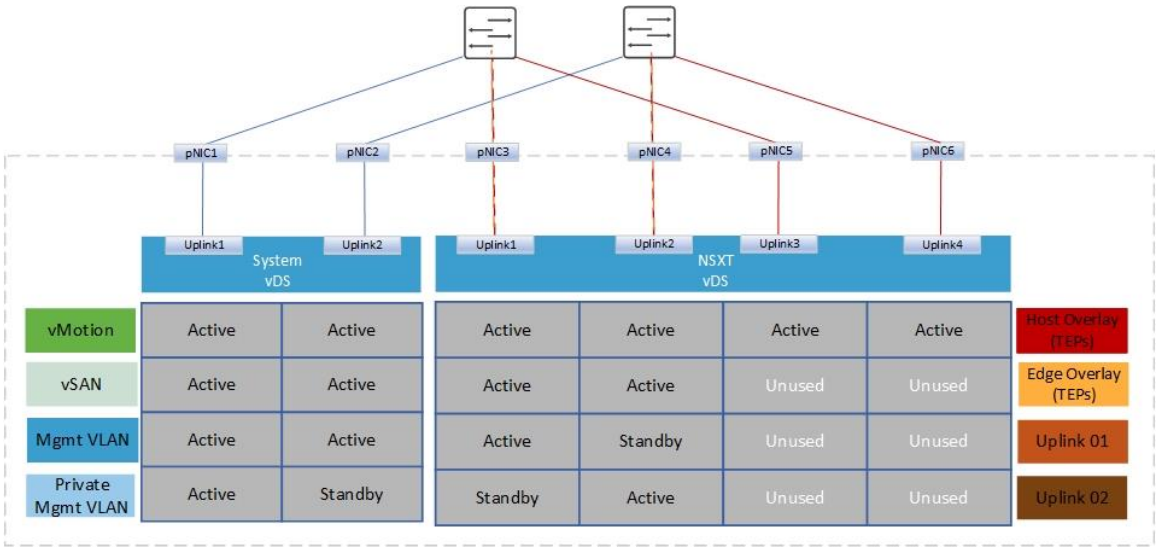


Figure 20. Two vDS with six pNICs option 2

Two vDS (system and NSX) – 8pNIC topologies

The 8-pNIC option that is illustrated in the following figure provides a high level of network isolation and also the maximum bandwidth for NSX east-west between host transport nodes. At the cost of a large port count on the switches, each host requires four ports per switch. The VxRail vDS (system) uses four uplinks and the NSX vDS also uses four uplinks.

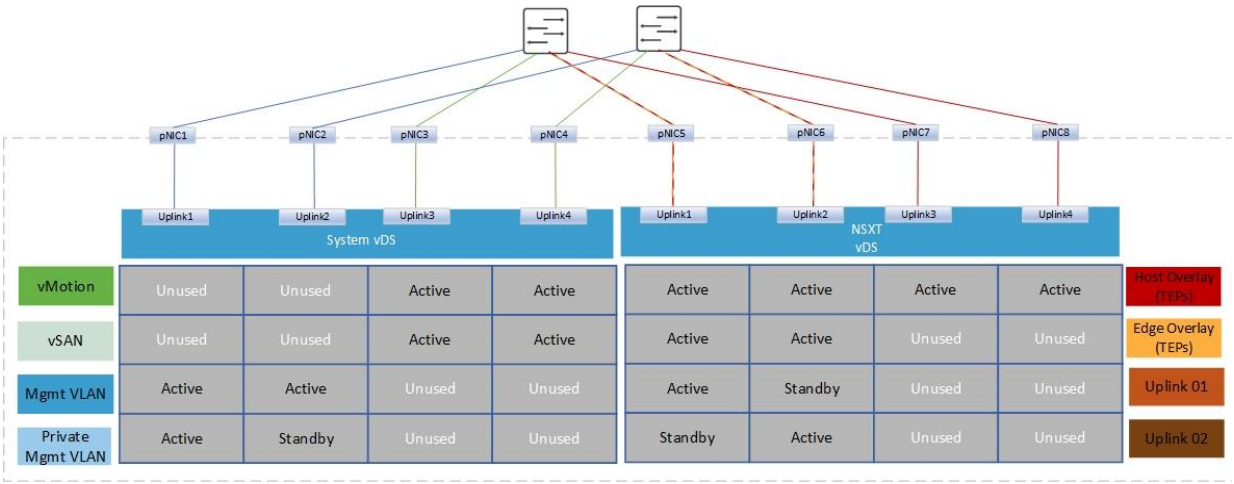


Figure 21. Two vDS (system and NSX) designs with eight pNICs

Two system vDS

Users can configure two VxRail system vDS to separate system traffic onto two different vDS. For example, vMotion and external management traffic can be on one vDS and vSAN on another vDS. Either one of the two VxRail vDS can be used for NSX traffic. Alternatively, you can use a dedicated NSX vDS, which results in three vDS in the network design. Sometimes physical separation is needed for management/vMotion, vSAN, and NSX traffic. The three-vDS design provides this capability. This section describes sample topologies.

Two system vDS – four pNIC

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

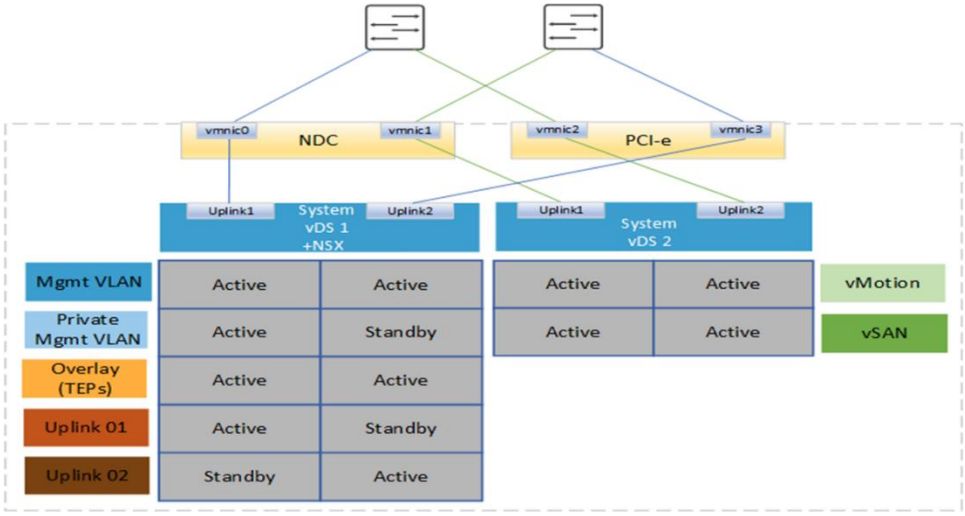


Figure 22. Two-system vDS design with four pNIC

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

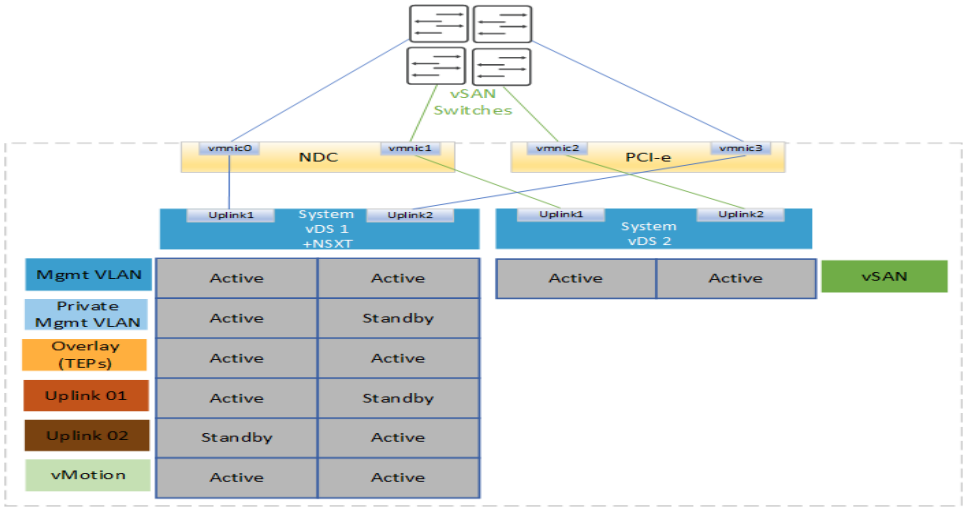


Figure 23. Two-system vDS design with four pNIC and isolated vSAN

Two system vDS – six pNIC

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

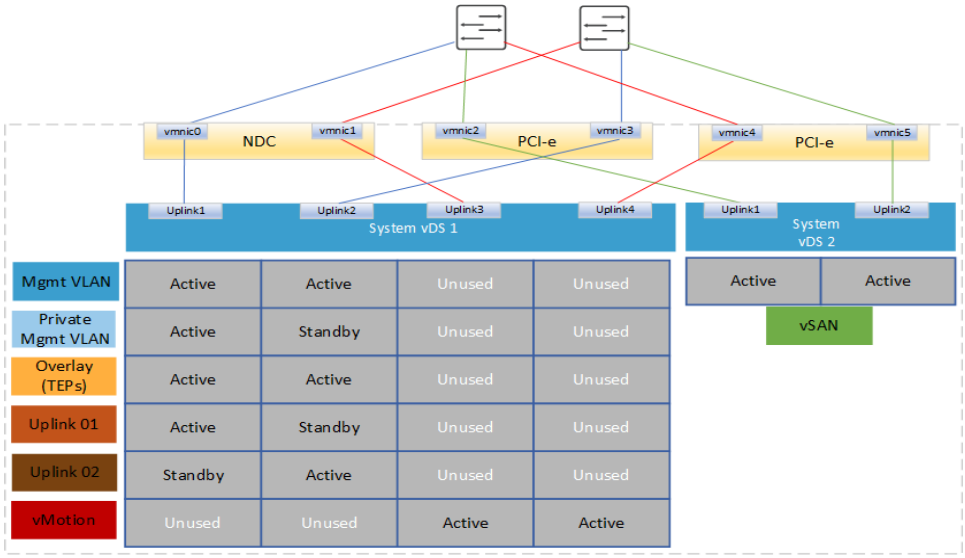


Figure 24. Two-system vDS design with six pNIC

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

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

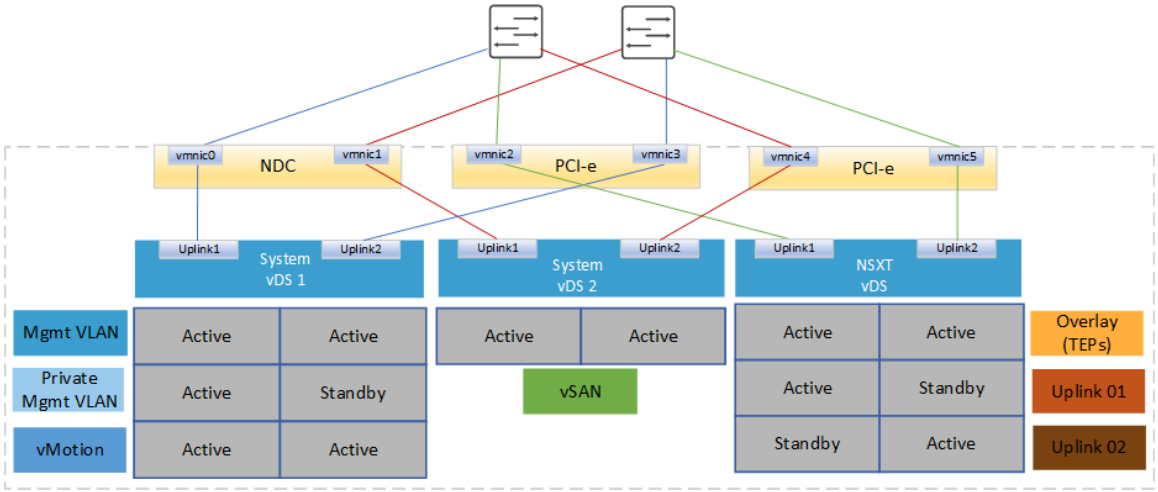


Figure 25. Two system vDS and NSX vDS

Network Virtualization

NSX provides the foundation of the network virtualization layer for VCF on VxRail. The solution provides a software-defined networking approach that delivers Layer 2 to Layer 7 networking services (for example, switching, routing, firewalling, and load balancing) in software. These services can then be programmatically assembled in any arbitrary combination, producing unique, isolated virtual networks in a matter of seconds. For multicloud connectivity and security, NSX provides the best possible features and provides native support for Kubernetes, PKS, and Cloud Native applications.

NSX architecture

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

- NSX Managers
- NSX Transport Nodes
- NSX Edge Nodes
- NSX Gateway T0/T1
- NSX Distributed Routers (DR)
- NSX Service Routers (SR)
- NSX Segments (Logical Switches)

Management plane

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

Note: Bare-metal NSX servers and Edges are not supported.

Control plane

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

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

Data plane

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

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

NSX network services

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

Segments (logical switch)

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

Gateway (logical router)

A logical router, also known as a gateway, consists of two components: distributed router (DR) and service 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 configured services such as NAT or DHCP.

Transport zones

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

Transport node

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

NSX 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 such as north-south routing or Stateful NAT require the SR component of logical routers to run on the Edge Node. The Edge Node is also a transport node just like the compute nodes in NSX. Similar to a compute node, it can connect to more than one transport zone. The Edge Node typically connects to one for overlay and other for north-south peering with external devices.

NSX 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 Edge cluster does not have a one-to-one relationship with a VxRail cluster. NSX Edge clusters can be distributed across multiple VxRail clusters.

Distributed firewall

The NSX 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 distributed firewall provides stateful protection of the workload at the vNIC level. DFW enforcement occurs in the hypervisor kernel, helping to deliver microsegmentation. A uniform security policy model for on-premises and cloud deployment supports multi-hypervisor (that is, ESXi and KVM) and multi-workload, with a level of granularity down to VM and container attributes.

NSX WLD Design

Application Virtual Network (AVN)

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

NSX transport zone design

A transport zone defines the span of the virtual network because logical switches extend only 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, including the Edge VM communications.

For the Mgmt WLD, only one VxRail cluster exists in standard architecture. All nodes in the VxRail cluster are added to an Overlay network. This network is used for AVN if the feature is enabled. VLAN-backed transport zone can be used for any VLAN-backed segments that are created in NSX.

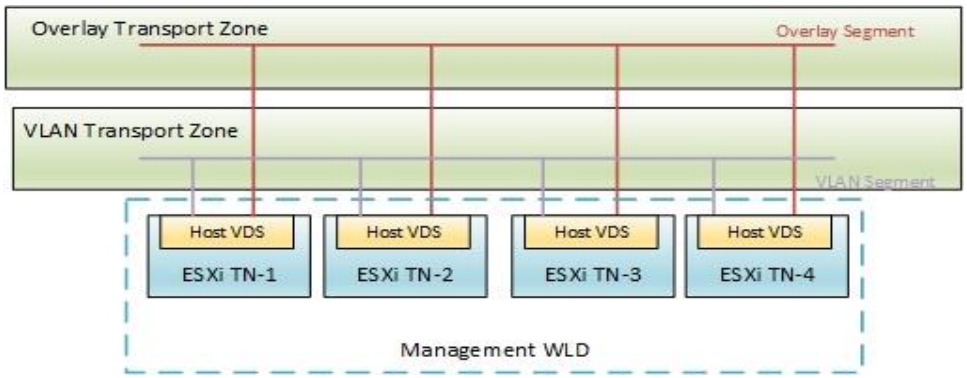


Figure 26. Mgmt WLD transport zones

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

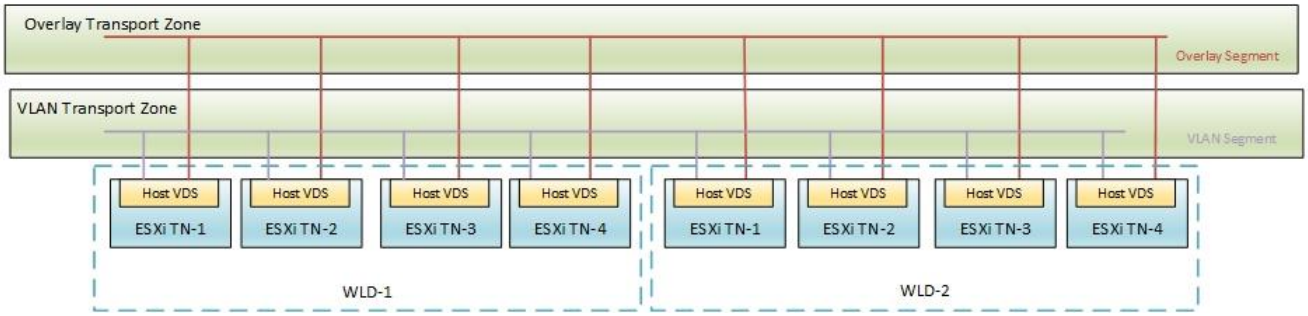


Figure 27. VI WLD 1: Many NSX transport zones

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

NSX segments

Segments are used to connect VMs to Layer 2 networks, and they can be either VLAN or Overlay segments. For the Mgmt WLD, when the NSX Edge Nodes and AVNs are deployed from SDDC Manager, two segments are created for the AVNs—Region A and X-Region—to be used for the VMware Aria Suite of components. You can create overlay-backed NSX segments or VLAN-backed NSX segments. Both options create two NSX segments (Region-A and X-Region) on the NSX Edge cluster deployed in the default management vSphere cluster. Those NSX segments are used when you deploy the VMware Aria Suite products. Region-A segments are local instance NSX segments; X-Region segments are cross-instance NSX segments. Two segments are created to allow Edge Node traffic to flow through the ESXi hosts to the physical network. The following table shows the segments that are created to support AVN in the virtual infrastructure of the SDDC.

Table 9. NSX Segments for Mgmt WLD

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

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

Uplink profile design

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

- Uplinks to be used by the transport node
- Teaming policy that is applied to those uplinks
- VLAN used for the profile
- MTU applied to the traffic

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

Table 10. VxRail or NSX vDS with two uplinks - Mgmt and VI WLD uplink profiles

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

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

Table 11. Second vDS with four uplinks—Mgmt and VI WLD uplink profiles

WLD type	Profile	Default teaming policy	Active uplinks	Transport VLAN (example)	Recommended MTU
Mgmt WLD	Host Overlay profile (deployed by Cloud Builder)	Load-Balance Source	uplink-1, uplink-2, uplink-3, uplink-4	103	9000
Mgmt WLD	Edge uplink profile (Day-2 deployed by Edge cluster automation from SDDC Manager)	Load-Balance Source	Any two of the four uplinks can be selected during deployment.	108	9000
VI WLD01	Host Overlay profile (deployed by SDDC Manager)	Load-Balance Source	uplink-1,uplink-2, uplink-3, uplink-4	203	9000
VI WLD01	Edge uplink profile (Day-2 deployed by Edge cluster automation from SDDC Manager)	Load-Balance Source	Any two of the four uplinks can be selected during deployment.	208	9000

Note: The Edge uplink profiles also include a Named Teaming policy that uses failover order for the uplink traffic. Thus, north-south traffic can be pinned to each physical network router.

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

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

Transport node profiles

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

- Name
- Transport zones for vDS participation – Overlay and VLAN TZ
- Uplink profile
- IP assignment type for the TEPs – DHCP or IP pool
- Physical NIC mapping – VMNICs to uplinks

The underlying vDS determines which pNICs are mapped in the transport node profile. If a system vDS is used for NSX, the pNIC can be mapped to the same pNICs used for

external management, vMotion, or vSAN. If a dedicated vDS is used for NSX, the pNICs are selectable, and either two or four uplinks can be mapped.

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

Table 12. NSX transport node profiles with two uplinks

WLD type	Transport zones	Uplink profile	IP assignment	Physical NIC mapping
Mgmt WLD	Host Overlay, VLAN	Mgmt WLD Host Uplink Profile	DHCP, IP Pool	pNIC1, pNIC2
VI WLD01	Host Overlay	VI WLD Host Uplink Profile 01	DHCP, IP Pool	pNIC1, pNIC2

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

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

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

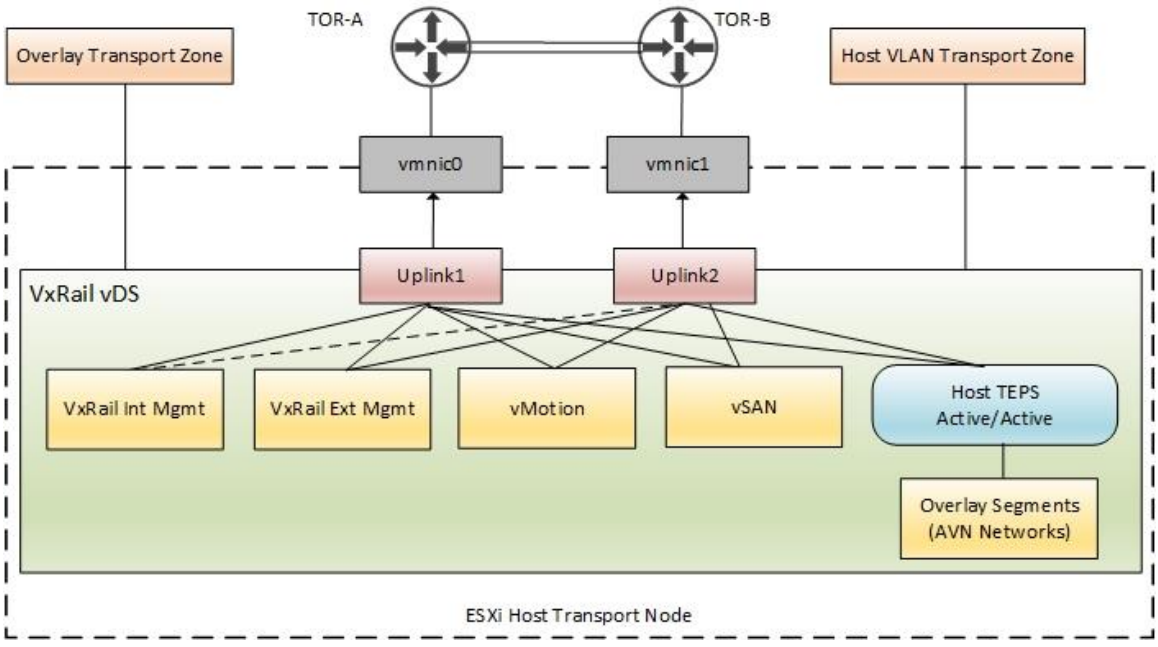


Figure 28. Mgmt WLD transport node – single VxRail vDS (only two uplinks)

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

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

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

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

Table 13. NSX transport node profiles with four uplinks

WLD type	Transport zones	Uplink profile	IP assignment	Physical NIC mapping
Mgmt WLD	Host Overlay, VLAN	Mgmt WLD Host Uplink Profile	DHCP, IP Pool	User Selectable: pNIC1, pNIC2, pNIC3, pNIC4
VI WLD01	Host Overlay	VI WLD Host Uplink Profile 01	DHCP, IP Pool	User Selectable: pNIC1, pNIC2, pNIC3, pNIC4

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

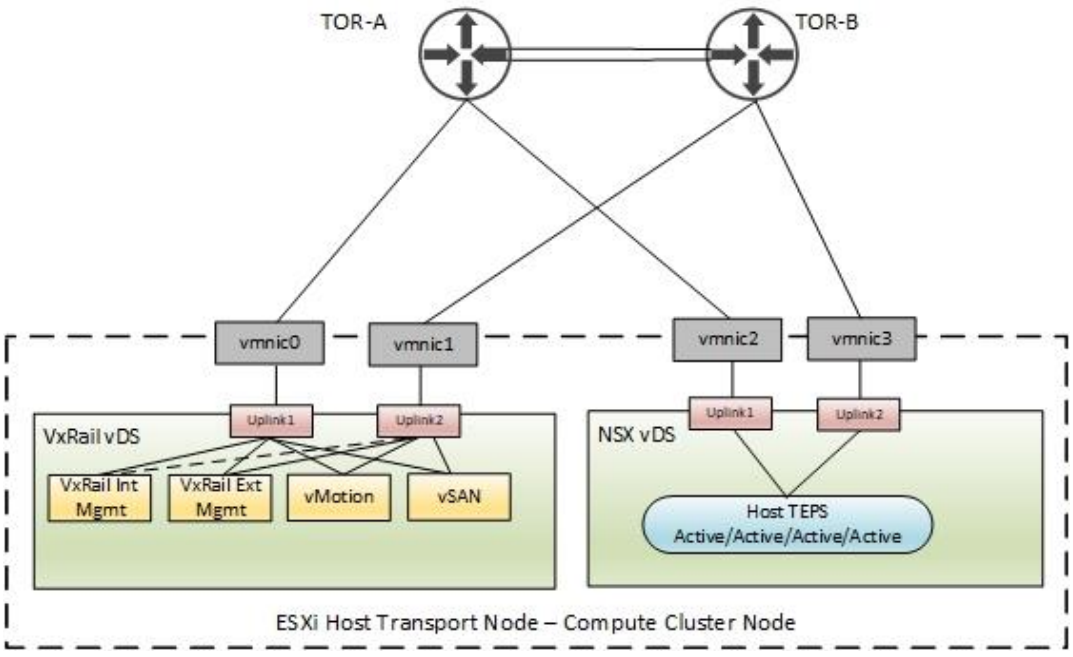


Figure 29. VI WLD transport node – Second NSX vDS (can be two or four uplinks).

NSX Edge Node design

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

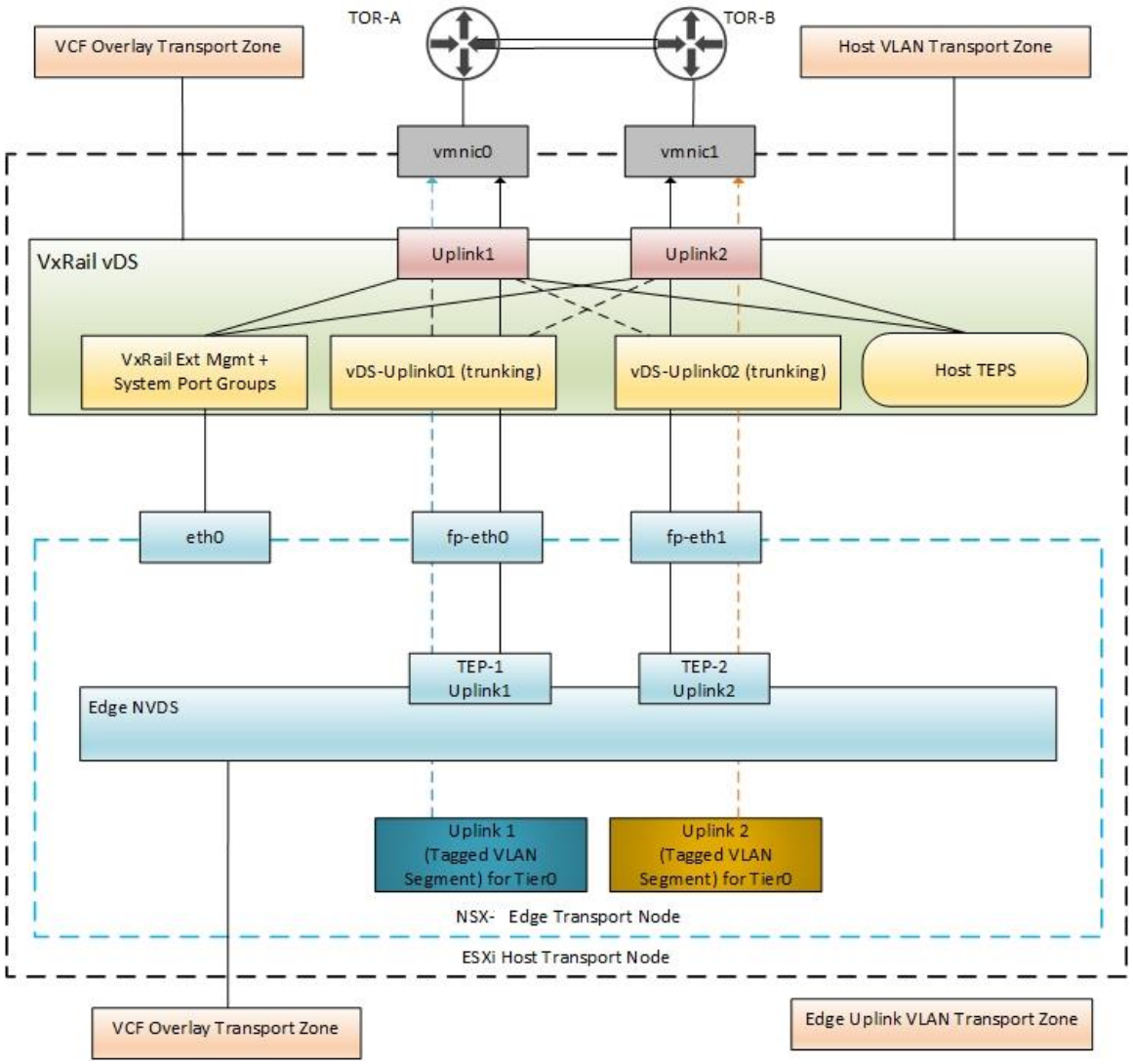


Figure 30. Mgmt WLD - Edge Node connectivity with single vDS

Note: The uplink port groups used to connect the Edge VM overlay interfaces are configured as trunk because the N-VDS does the VLAN tagging. The uplink profile defines the VLAN for the Edge overlay.

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

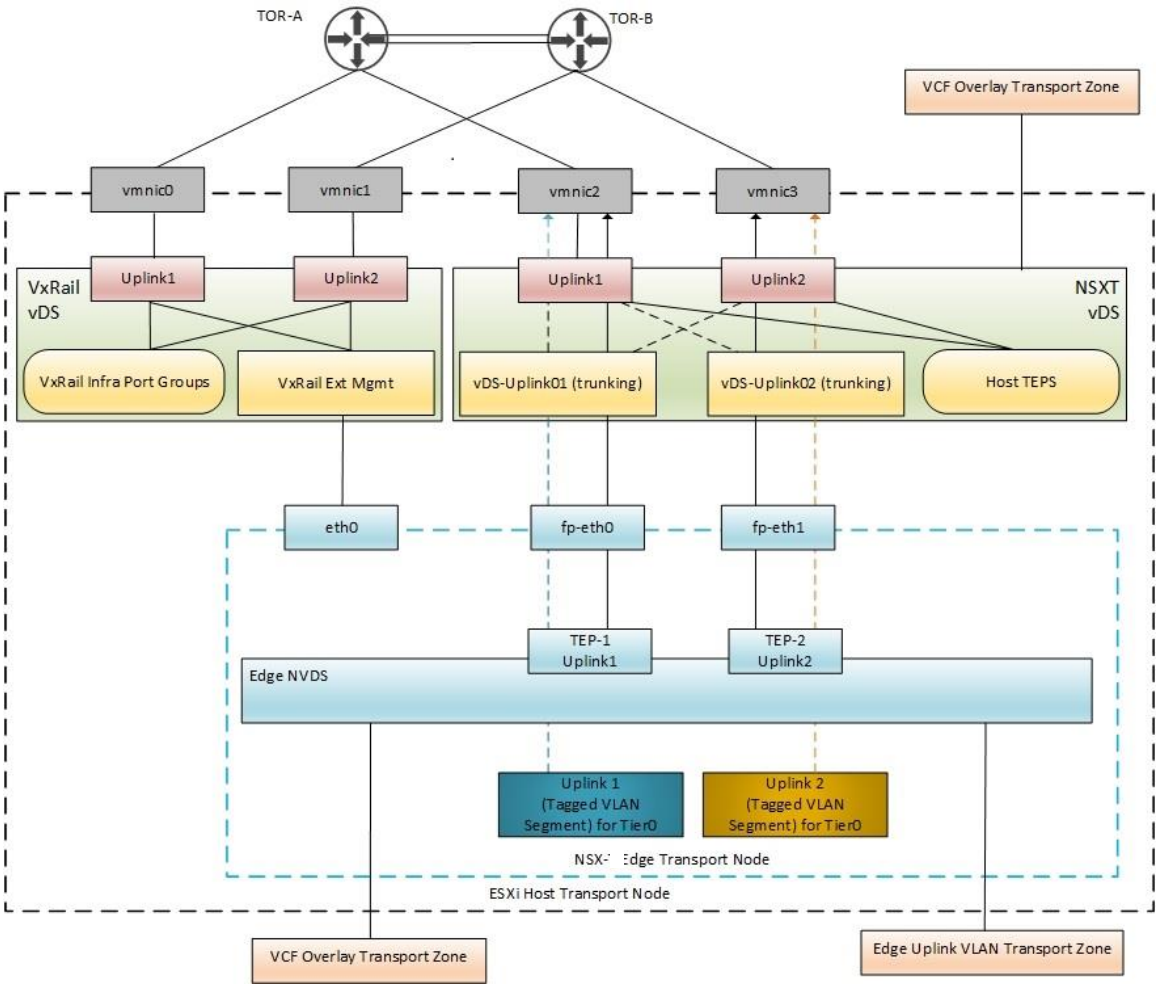


Figure 31. VI WLD – Edge Node connectivity with dedicated NSX vDS

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

NSX Edge north-south routing design

The NSX Edge routing design is based on the VCF design (see VMware's Network Design for the NSX Edge Nodes for the Management Domain). A Tier-0 gateway is deployed in active/active mode with ECMP enabled to provide redundancy and better bandwidth utilization. Both uplinks are used. Two uplink VLANs are needed for north-south connectivity for the Edge VMs in the Edge Node cluster. The dedicated uplink profile that is created for the Edge transport nodes defines named teaming policies. These

policies are to be used in the Edge uplink transport zone and in the segments that are created for the Tier 0 gateway and used as a transit network to connect the Tier 0 interfaces. The named teaming policy allows traffic from the Edge Node to be pinned to an uplink network/VLAN connecting to the physical router. BGP provides dynamic routing between the physical environment and the virtual environment. eBGP is used between the Tier-0 Gateway and the physical TORs. An iBGP session is established between the T0 Edge VMs SR components.

Support for static routes if BGP is not a viable option for the customer.

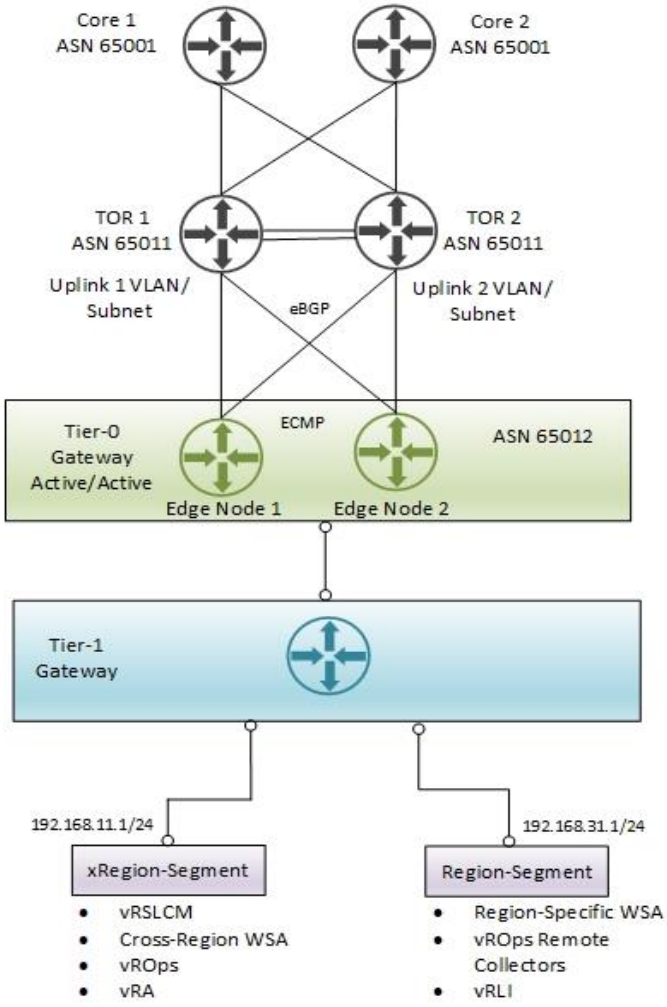


Figure 32. Mgmt WLD Edge Node north-south routing design

NSX Mgmt WLD physical network requirements

The following NSX external network requirements must be met before deployment of the Mgmt WLD:

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

- If DHCP is used for the Host TEPs, IP Helper is required on the switches if the DHCP server is in a different L3 network.

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

- Layer 3 license for peering with T0 Edges
- BGP configured for each router peering with a T0 Edge
- Two uplink VLANs for T0 Edge external connectivity to physical network
- Edge Overlay VLAN created on the physical switches
- Uplink and Overlay VLANs added to the trunk ports connecting to VxRail nodes

NSX VI WLD physical network requirements

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

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

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

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

NSX deployment in Mgmt WLD

Cloud Builder is used to deploy the NSX components in the Mgmt WLD VxRail. The deployment process consists of the following general steps:

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

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

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

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

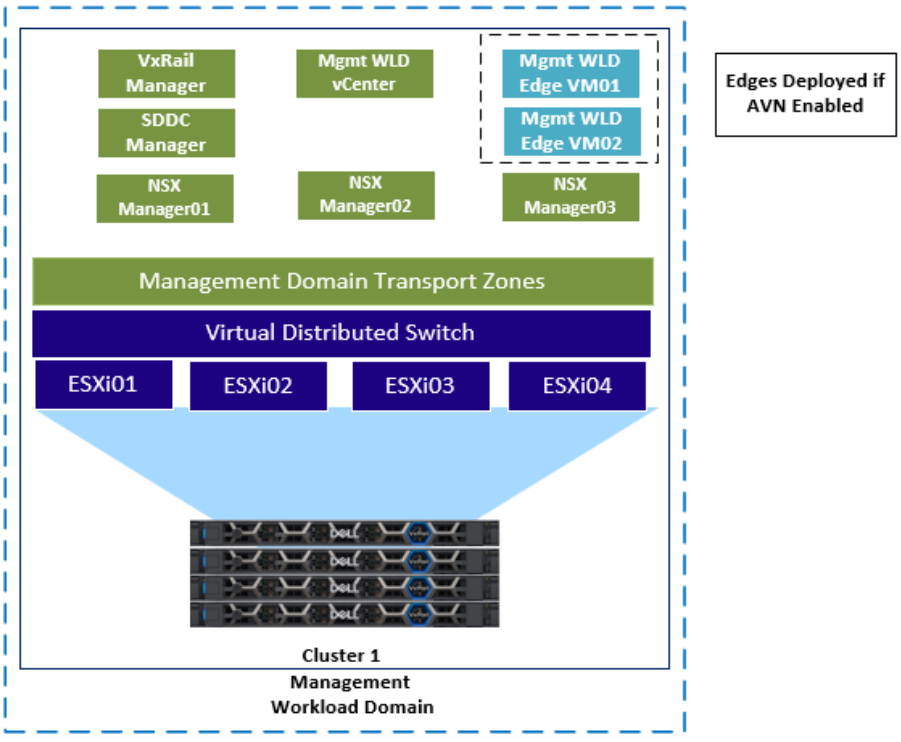


Figure 33. Mgmt WLD after deployment

NSX deployment in VI WLD

The NSX components are installed when the first VxRail cluster is added to the first NSX VI WLD. The SDDC Manager deploys the NSX Managers in the Mgmt WLD. It assigns an IP address to each NSX Manager virtual appliance, and it assigns a front-end virtual IP

address for those appliances. After assigning those addresses, SDDC Manager configures the VI WLD VxRail cluster to be used for NSX services.

The major steps of the deployment process are as follows:

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

Note: No additional NSX Managers are needed when a second NSX based VI WLD is added and part of the same SSO Domain. However, you can deploy a new NSX domain for each WLD if that is a requirement, 1:1 NSX for each VI WLD.

NSX Edges can be deployed using automation for the VI WLDs using SDDC Manager. This feature allows the Edges to be automatically deployed consistently per VCF guidance.

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

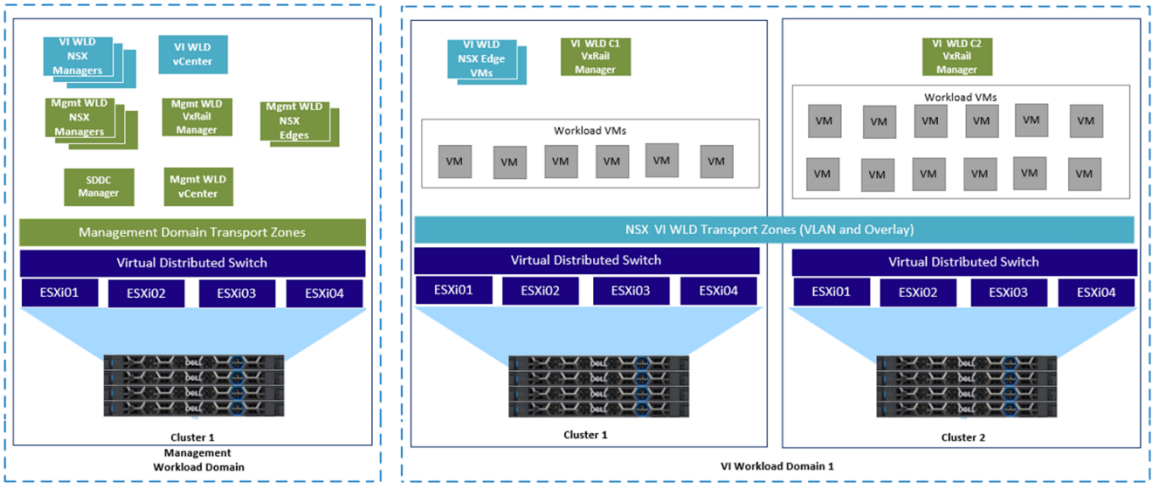


Figure 34. NSX VI WLD VxRail cluster design

Enabling VCF with Tanzu Features on workload domains

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

Prerequisites

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

- Licensing: Within a WLD, all hosts within the selected VxRail clusters must have the proper VMware vSphere for Kubernetes licensing to support Workload Management.
- Workload Domain: A VI WLD deployed as workload-management-ready must be available.
- NSX Edge cluster: At least one NSX Edge cluster must be deployed from SDDC Manager and available.

IP addresses:

  - Define a subnet for pod networking (nonroutable), a minimum of a /22 subnet.
  - Define a subnet for Service IP addresses (nonroutable), a minimum of a /24 subnet.
  - Define a subnet for Ingress (routable), minimum of a /27subnet.
  - Define a subnet for Egress (routable), minimum of a /27 subnet.

VCF with Tanzu detailed design

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

Physical network design considerations

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

- The investment that you have today in your current physical network infrastructure
- The advantages and disadvantages for both Layer 2 and Layer 3 designs

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

Traditional 3-tier (access/core/aggregation)

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

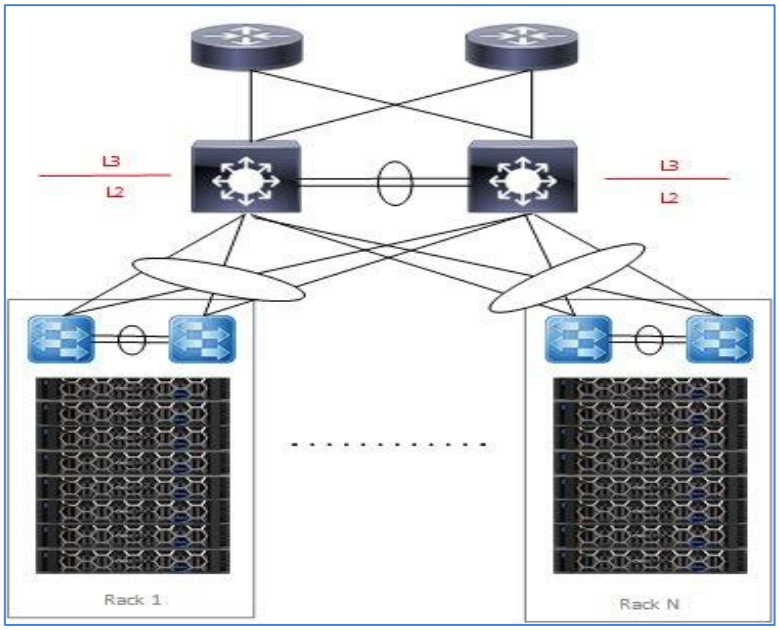


Figure 35. Traditional 3-tier Layer 2 fabric design

It has the following characteristics:

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

Advantages:

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

Disadvantages:

- Large VxRail clusters spanning racks will create large broadcast domains.
- Interoperability issues between different switch vendors can introduce spanning tree issues in large fabrics.
- The NSX T0 gateways for each WLD need to peer at the aggregation layer. For large-scale deployments with multiple WLDs, the configuration becomes complex.
- The size of such a deployment is limited because the fabric elements have to share a limited number (4,094) of VLANs. With NSX, the number of VLANs could be reduced so this limitation might not be an issue.

Leaf and spine Layer 3 fabric

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

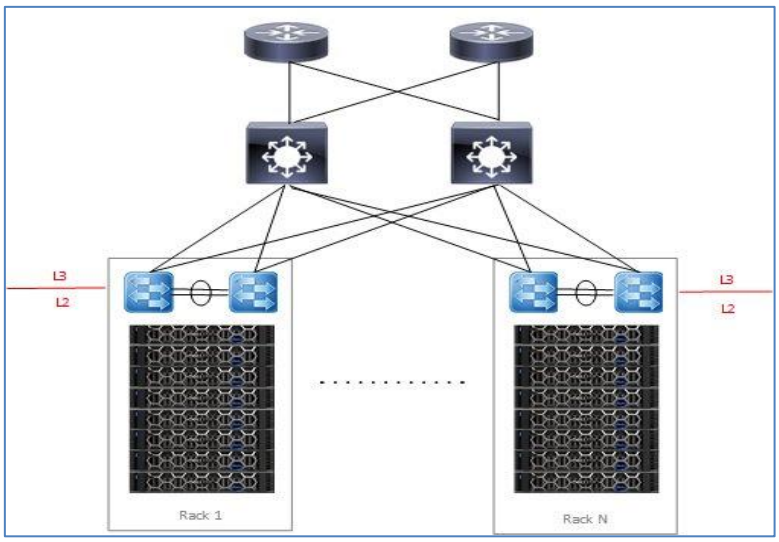


Figure 36. Leaf and spine Layer 3 design

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 T0 Gateway for the WLD peers with the leaf switches in one rack.

Advantages:

- Vendor agnostic - Multiple network hardware vendors can be used in the design.
- Reduced VLAN span across racks, thus smaller broadcast domains.
- East-west for an NSX domain can be confined within a rack with intra-rack routing at the leaf.
- East-west across NSX domains or cross-rack is routed through the spine.
- NSX Tier 0 peering is simplified by peering the WLDs with the leaf switches in the rack.

Disadvantages:

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

Multirack design considerations

You might want to span WLD VxRail clusters across racks to avoid a single point of failure within one rack. The management VMs running on the Mgmt WLD VxRail cluster and any management VMs running on the VI WLD require VxRail nodes to reside on the same L2 management network. This requirement ensures that the VMs can be migrated between

racks and maintain the same IP address. For a Layer 3 Leaf-Spine fabric, this requirement is a problem because the VLANs are terminated at the leaf switches in each rack.

SDDC Manager now provides the option to select Static or DHCP-based IP assignments to Host TEPs. This option can also be used for stretched clusters and L3 aware workload domain clusters. VMware Cloud Foundation 5.1 on VxRail 8.0.200 supports the configuration of a Sub-Transport Node profile (Sub-TNP) within NSX as a new topology for vSAN stretched clusters. This is useful when a cluster spans multiple racks and when the transport nodes of this cluster must use different transport VLANs or acquire their IP addresses for their tunnels using different IP pools.

VxRail cluster across racks

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

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

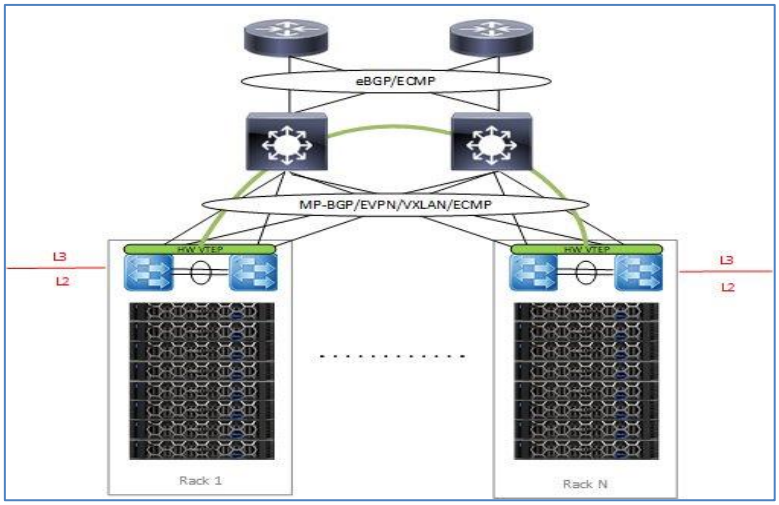


Figure 37. Multi-rack VxRail cluster with hardware VTEP

VxRail physical network interfaces

VxRail can be deployed with either 2x10/2x25 GbE, 4x10 GbE, 4x25 GbE or 2x100Gbe predefined profiles. Beginning with VxRail version 7.0.130, custom profiles can be used. VxRail needs the necessary network hardware to support the initial deployment. The following table illustrates various physical network connectivity options with a single system vDS and a dedicated NSX vDS with and without NIC-level redundancy. For this table, a standard wiring configuration can be used for connectivity—odd-numbered uplinks cabled to Fabric A and even-numbered uplinks cabled to Fabric B.

Notes:

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

Table 14. Physical network connectivity options

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

The following figures illustrate some of the different host connectivity options from the preceding table for the different VxRail deployment types for either the Mgmt WLD or a VI WLD. For the Mgmt WLD and the VI WLD, the Edge overlay and the Edge Uplink networks will be deployed when NSX Edges are deployed using Edge automation in SDDC Manager. There are too many different options to cover in this section. The following sections describe the most common options from Table 14.

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

Single VxRail vDS connectivity options

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

10 GbE connectivity options

This figure illustrates option A in Table 14. The VxRail deployed with 2x10 predefined network profile on the 4-port NDC. The remaining two ports are unused and can be used for other purposes if required.

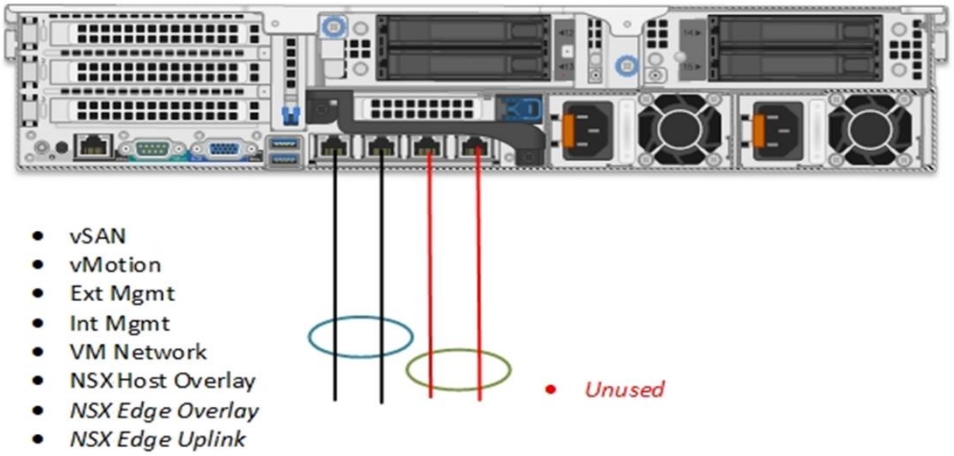


Figure 38. Single VxRail vDS - 2x10 predefined network profile

The next figure refers to option C in Table 14. The VxRail is deployed with a 4x10 predefined network profile. This places vSAN and vMotion onto their own dedicated physical NICs on the NDC and NSX traffic will use vmnic0 and vmnic1 shared with management traffic. More PCI cards can be installed and used for other traffic if that is required.

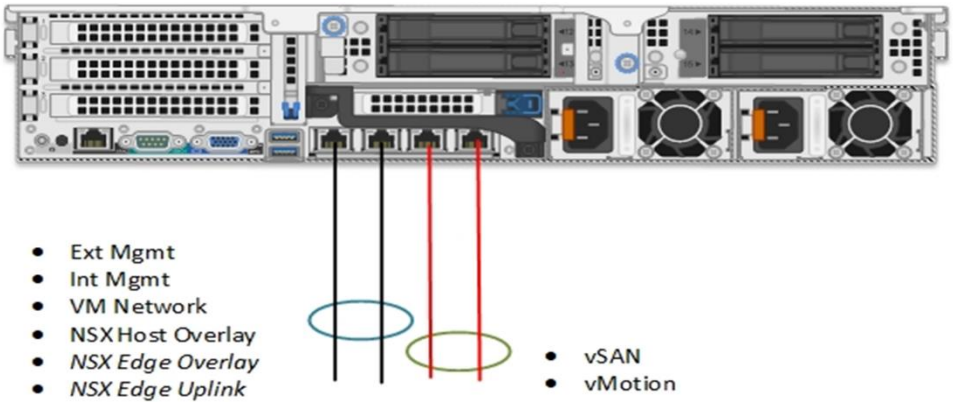


Figure 39. Single VxRail vDS - 4x10 predefined network profile

The final 10 GbE option provides NIC-level redundancy. To achieve this redundancy, use an NDC and PCIe with a custom profile to deploy the VxRail vDS. The following figure illustrates this option, which is option D in Table 14.

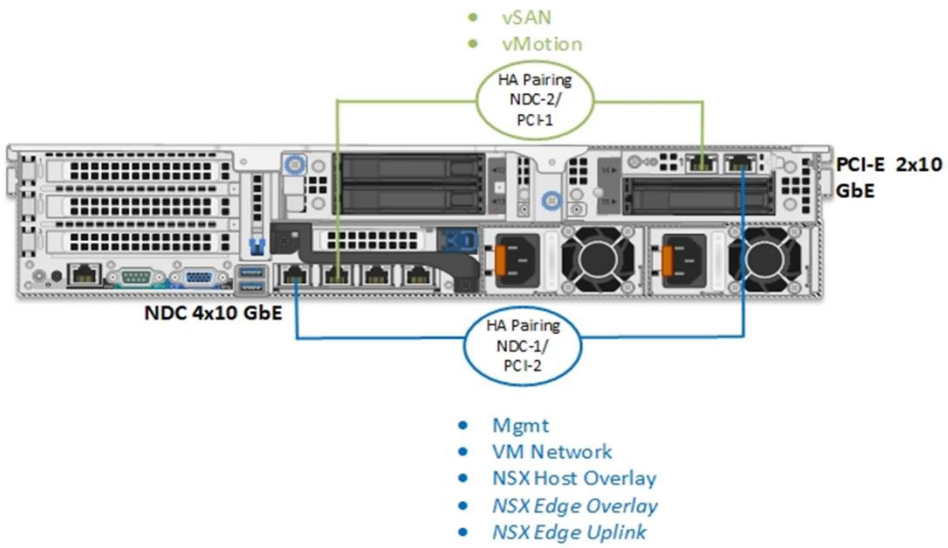


Figure 40. Single VxRail vDS – 4x10 custom profile and NIC-level redundancy

25 GbE connectivity options

The first option aligns with option A in Table 14. A single VxRail vDS using 2x25 network profile for the VxRail using the 25GbE NDC. The VxRail system traffic uses the two ports of the NDC along with NSX traffic.

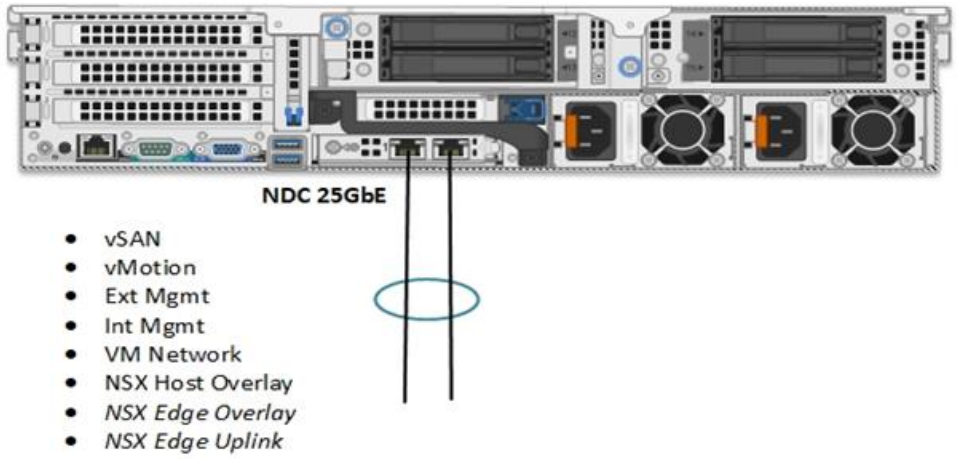


Figure 41. Single VxRail vDS - 2x25 network profile

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

The second option aligns with option D in Table 14. A single VxRail vDS using a custom network profile which provides NIC-level redundancy for the VxRail system traffic and the NSX TEP and Edge uplink traffic. A standard physical cabling configuration can be achieved with the logical network configuration described in section VxRail vDS custom profiles.

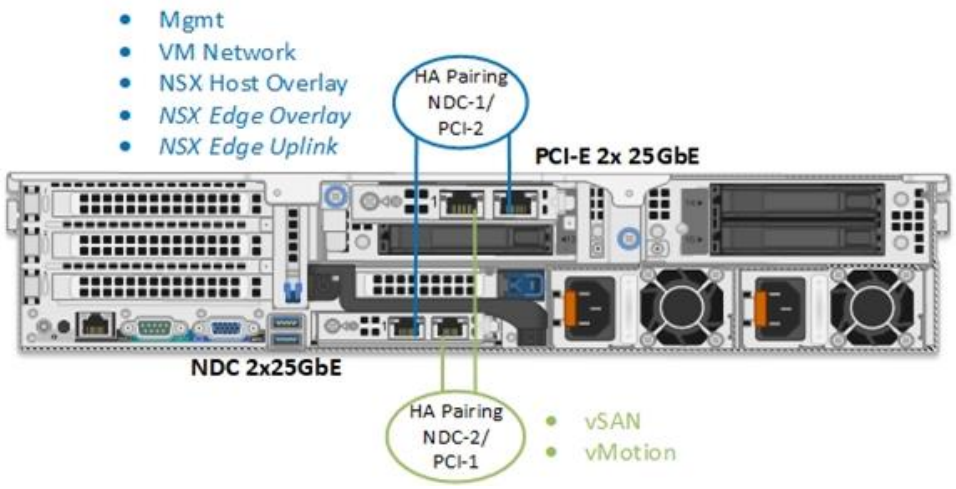


Figure 42. Single VxRail vDS - 4x25 custom network profile

NSX vDS connectivity options

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

10 GbE connectivity options

The following figure illustrates a VxRail deployed with 2x10 predefined network profile on the 4-port NDC which aligns to option E in Table 14. The remaining two ports are used for the second vDS for the NSX traffic.

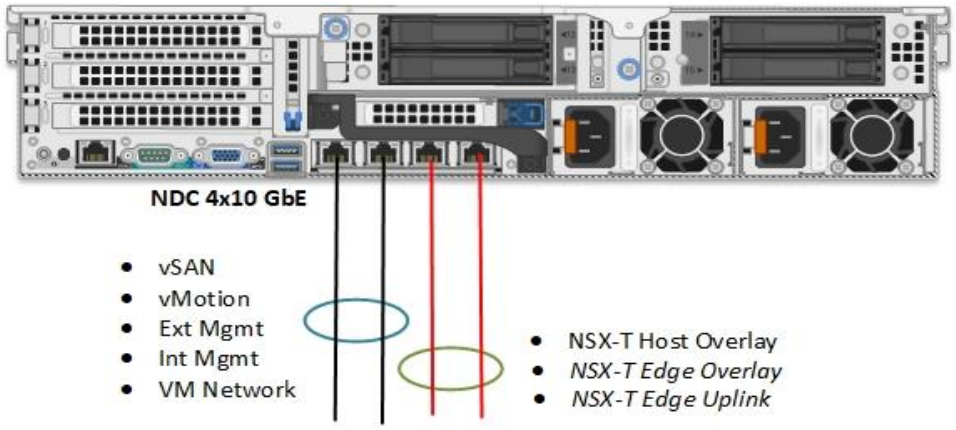


Figure 43. VxRail vDS and NSX vDS using two ports each on 4x10 NDC

The second option is VxRail deployed with a 4x10 predefined network profile consuming all four ports of the NDC. This option places vSAN and vMotion onto their own dedicated physical NICs on the NDC. The NSX traffic uses a dedicated vDS and uplinks connecting to pNICs on the PCI-E 10 GbE, as shown in the following figure, which represents option H in **Error! Reference source not found.**.

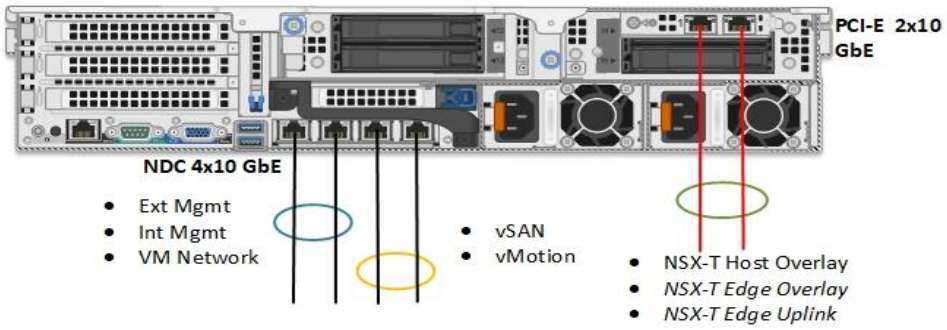


Figure 44. VxRail vDS using four ports of NDC and NSX vDS using PCI-E

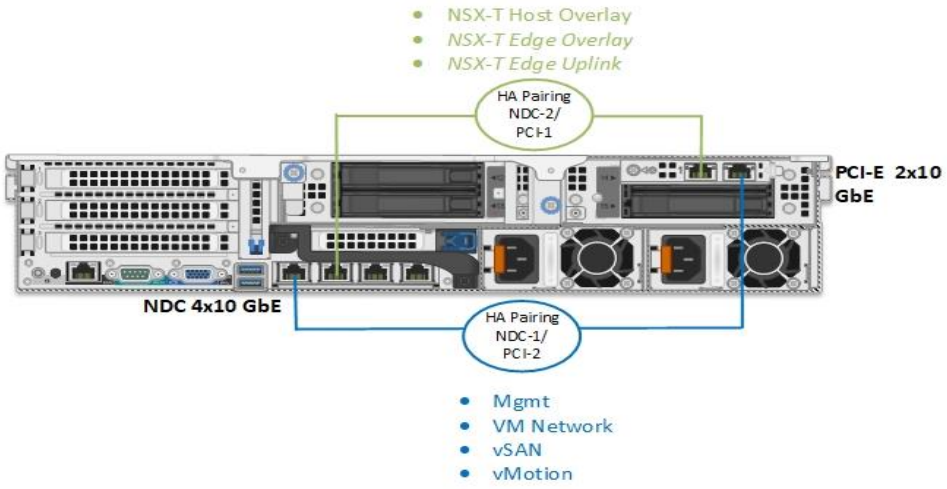


Figure 45. VxRail vDS and NSX vDS with NIC-level redundancy

The final option for a 10 GbE network environment that we want to illustrate provides NIC-level redundancy across the NDC and PCIe. Redundancy is for both system traffic on the VxRail vDS and NSX traffic on the dedicated NSX vDS. VxRail is deployed with the custom profile option using one port from NDC and one from PCIe. Similarly, the NSX vDS uses a port from each NIC. The following figure represents option G in **Error! Reference source not found.**.

25 GbE connectivity options

The first 25 GbE option uses the 2-port 25 GbE NDC for the VxRail vDS. A dedicated vDS is created for the NSX traffic using the two ports of the PCI-E card, which represents option F in **Error! Reference source not found.**.

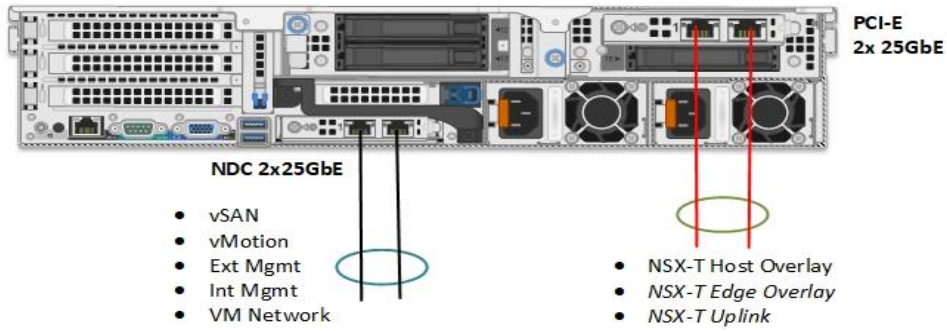


Figure 46. VxRail vDS using 25 GbE NDC and NSX vDS using 25 GbE PCI-E

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

The second option requires a total of six 25GbE ports. The VxRail is deployed with the 4x25 custom profile option, as previously discussed, using the two-port NDC. The two port PCIe and the second vDS for NSX traffic require an additional 2x25 GbE card. The following figure represents this configuration, which is option I in **Error! Reference source not found.**.

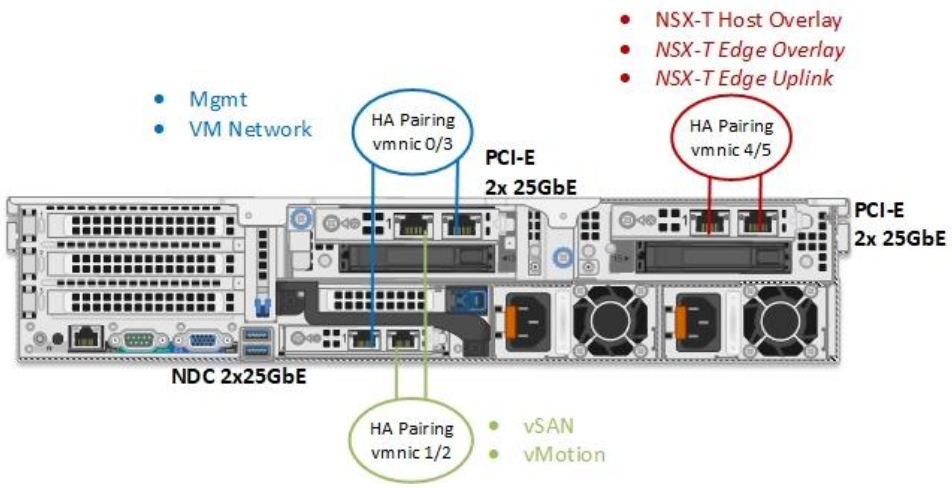


Figure 47. VxRail vDS using 2x25 GbE NDC and PCI-E, NSX vDS using 25 GbE PCI-E

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

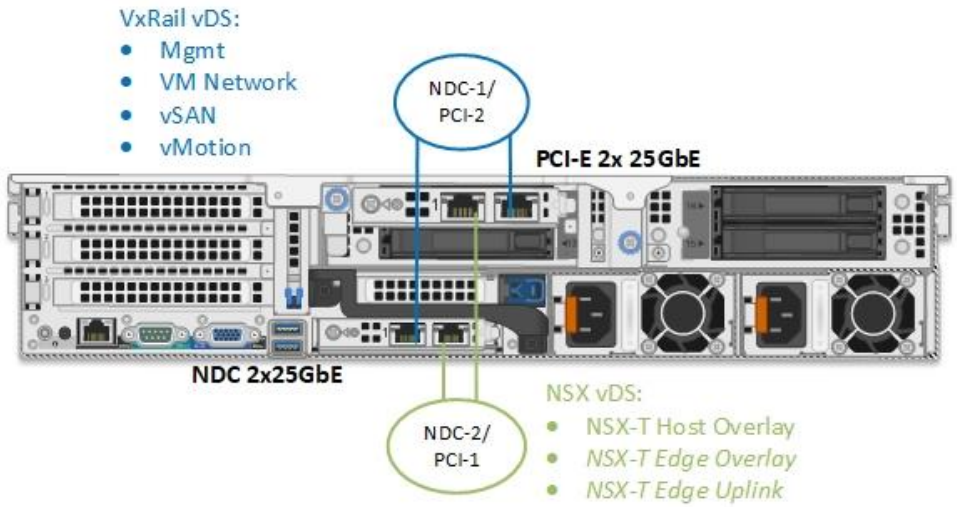


Figure 48. 25 GbE NIC-level redundancy for system and NSX traffic

There are additional options that support up to eight pNICs for both 10 GbE and 25 GbE networks that are not illustrated here, but they are in **Error! Reference source not found.** for reference.

Storage options

VCF on VxRail has flexible storage options that include vSAN and Fibre Channel storage as principal storage options for a VI WLD. The Management WLD primary VxRail cluster still requires vSAN storage as its principal storage. iSCSI, NFS storage, vVols, and vSAN HCI Mesh can be used as supplemental storage for both the Management and VI WLDs.

As of VCF 5.1, VMware vSAN ESA supports both Management and VI Workload Domain clusters (vSAN ESA requires vLCM). vSAN ESA is designed as a single-tier architecture built on NVMe platforms.

NOTE: For ESA stretch Clusters the minimum code version is VCF 5.2.1.1

Table 15. Storage options

Storage option	Mgmt WLD VxRail cluster	VI WLD VxRail cluster	non-vLCM	vLCM
vSAN OSA	Principal Storage	Principal Storage	Yes	Yes
vSAN ESA	Principal Storage	Principal Storage	No	Yes
vSAN OSA - Cross Cluster Shared Storage (Day-2)	Supplemental Storage	Supplemental / Principal for Dynamic nodes only	Yes	Yes
vSAN ESA - Cross Cluster Shared Storage (Day-2)	Supplemental Storage	Supplemental / Principal for Dynamic nodes only	No	Yes
FC Storage (VMFS)	Supplemental Storage	Supplemental / Principal for Dynamic nodes only	Yes	Yes
iSCSI Storage	Supplemental Storage	Supplemental Storage	Yes	Yes
NFS Storage	Supplemental Storage	Supplemental Storage	Yes	Yes
NVMe over TCP	Supplemental Storage	Supplemental Storage	Yes	Yes
NVMe over FC	Supplemental Storage	Supplemental Storage	Yes	Yes
PowerFlex	Supplemental Storage	Supplemental Storage	Yes	Yes
vVols	Supplemental Storage	Supplemental Storage	Yes	Yes

Notes:

- In a consolidated architecture, a second or additional VxRail cluster in the Management WLD would have the same support as a VI WLD VxRail cluster.
- Use of Supplemental storage for Consolidated Architecture is supported and follows the same guidelines as Management WLD VxRail Cluster.

vSAN

vSAN is deployed, scaled, and life-cycle-managed with SDDC Manager and VxRail HCI System software automation. WLDs with VxRail vSAN clusters can be configured quickly

and can be ready to use without having to make complex changes within VxRail hardware. vSAN Storage Policy Based Management (SPBM) allows for storage characteristics such as primary failures to tolerate and disk striping. Storage settings can be changed in software at a VM or object-level nondisruptively with only a few clicks. Compare vSAN with traditional array-based storage, which often requires hardware changes and updates across a whole LUN or volume through processes that are often time-consuming and risky.

vSAN HCI Mesh

vSAN HCI Mesh supports the sharing of spare vSAN storage capacity between VxRail clusters in a VI WLD. It can only be used as a secondary storage option; either vSAN or FC storage must be used for principal storage. vSAN HCI Mesh enables an alternative method to alleviate decreasing storage capacity in a VI WLD. This alternative can be useful for environments that are not compute-constrained in a VI WLD to avoid adding nodes, which increase both compute and storage resources.

VCF 5.1 introduced support for the use of HCI Mesh as principal storage in the case of VxRail dynamic nodes VI WLD clusters, created running the WFO API script method on SDDC Manager.

Prerequisites

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

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

Feature support

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

FC storage

VCF 5.1 supports the ability to deploy a VI WLD with external storage as principal storage with no VxRail managed vSAN. This feature allows support for a VxRail dynamic node cluster with three nodes or more to be added to a VI WLD. VCF 5.1 SDDC Manager allows 2-node clusters when using vLCM and connecting to external storage as primary. These nodes contain no internal storage disks normally required for vSAN.

Requirements

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

Supported storage arrays:

- PowerStore
- PowerMax
- Unity XT

Supported FC HBAs:

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

Storage configuration required:

- Zoning of the VxRail nodes to the storage system
- Creation and masking of the LUN to the VxRail nodes
- 900 GB of free space on the volume
- Formatting the LUN with a Virtual Machine File System (VMFS)

If multiple datastores are discovered during VxRail Day 1 workflow, the largest one will be selected as the primary datastore for the VxRail systems.

If multiple datastores are discovered of the same size, a random one will be selected during VxRail Day 1 workflow.

Multi-site design considerations

The VCF on VxRail solution natively supports two different multi-site options depending on the distance and latency between the sites and the type of protection needed for the workloads. VxRail vSAN stretched clusters offer multiple availability zones for the Mgmt WLD and VI WLDs. This option is typically only for sites within the same metro area, due to the latency requirements for stretched vSAN. VCF 5.0 supports NSX Federation. NSX Federation enables support for dual-region VCF instances, which can be located at much greater distances because a stretched cluster is not required.

Multi-AZ (VxRail vSAN stretched cluster)

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

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

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

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

The following network requirements apply for the Mgmt WLD and the VI WLD VxRail clusters that must be stretched across the AZs in accordance with the VMware VVS design:

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

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

You cannot stretch a VxRail cluster if remote vSAN datastores are mounted on any VxRail cluster.

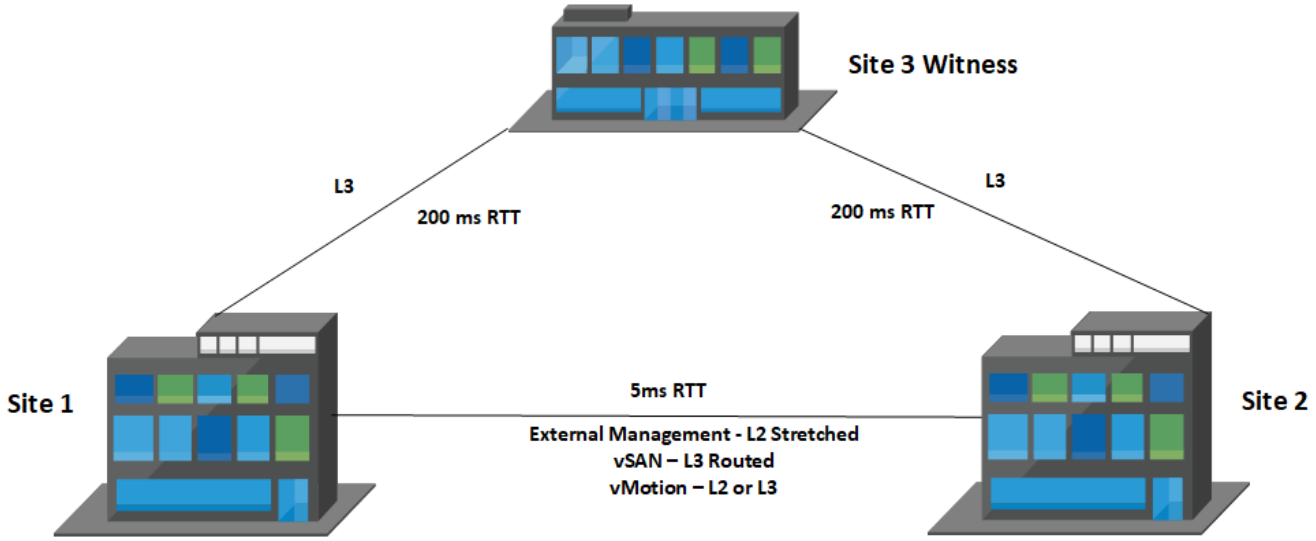


Figure 49. VxRail vSAN stretched cluster network requirements

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

Multi-AZ connectivity requirements

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

Table 16. Site connectivity and MTU

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

Increasing the vSAN traffic MTU to improve performance requires the MTU for the Witness traffic to the Witness site to also use an MTU of 9000. This requirement 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.

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

Table 17. Mgmt WLD sample VLAN and IP subnets

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

The VVS requirements for the VI WLD are the same as for the Mgmt WLD. If Edge Nodes are deployed, the Edge TEP and uplink networks must be stretched Layer 2 between sites. However, if stretched Layer 2 does not meet the requirements, implementing a different design might be possible. For alternative designs, consult VMware during the design phase of the project.

Table 18. VI WLD sample VLAN and IP subnets

Traffic type	AZ1	AZ2	Sample VLAN	Sample IP range
External Management	✓	✓	1631 (stretched)	172.16.31.0/24
VxRail Discovery	✓	✓	3939	N/A
vSAN	✓	×	1632	172.16.32.0/24
vMotion	✓	×	1633	172.16.33.0/24
Host TEP	✓	×	1634	172.16.34.0/24
Edge TEP	✓	✓	2731 (stretched)	172.27.31.0/24
Edge Uplink 01	✓	✓	2732 (stretched)	172.27.32.0/24

Traffic type	AZ1	AZ2	Sample VLAN	Sample IP range
Edge Uplink 02	✓	✓	2733 (stretched)	172.27.33.0/24
vSAN	×	✓	1641	172.16.41.0/24
vMotion	×	✓	1642	172.16.42.0/24
Host TEP	×	✓	1643	172.16.43.0/24

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

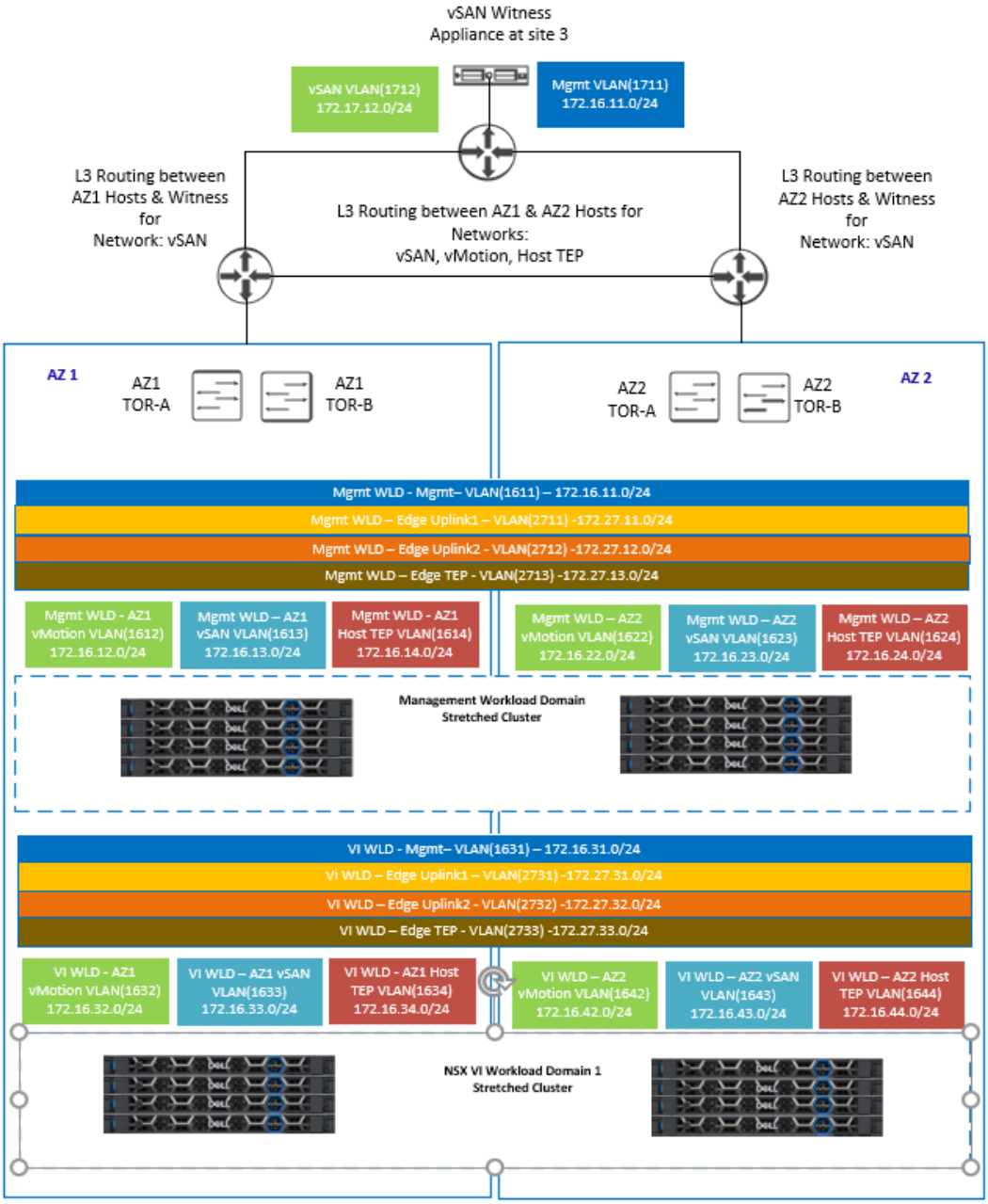


Figure 50. VLAN and network requirements for multi-AZ (VxRail vSAN stretched cluster)

Multi-AZ component placement

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

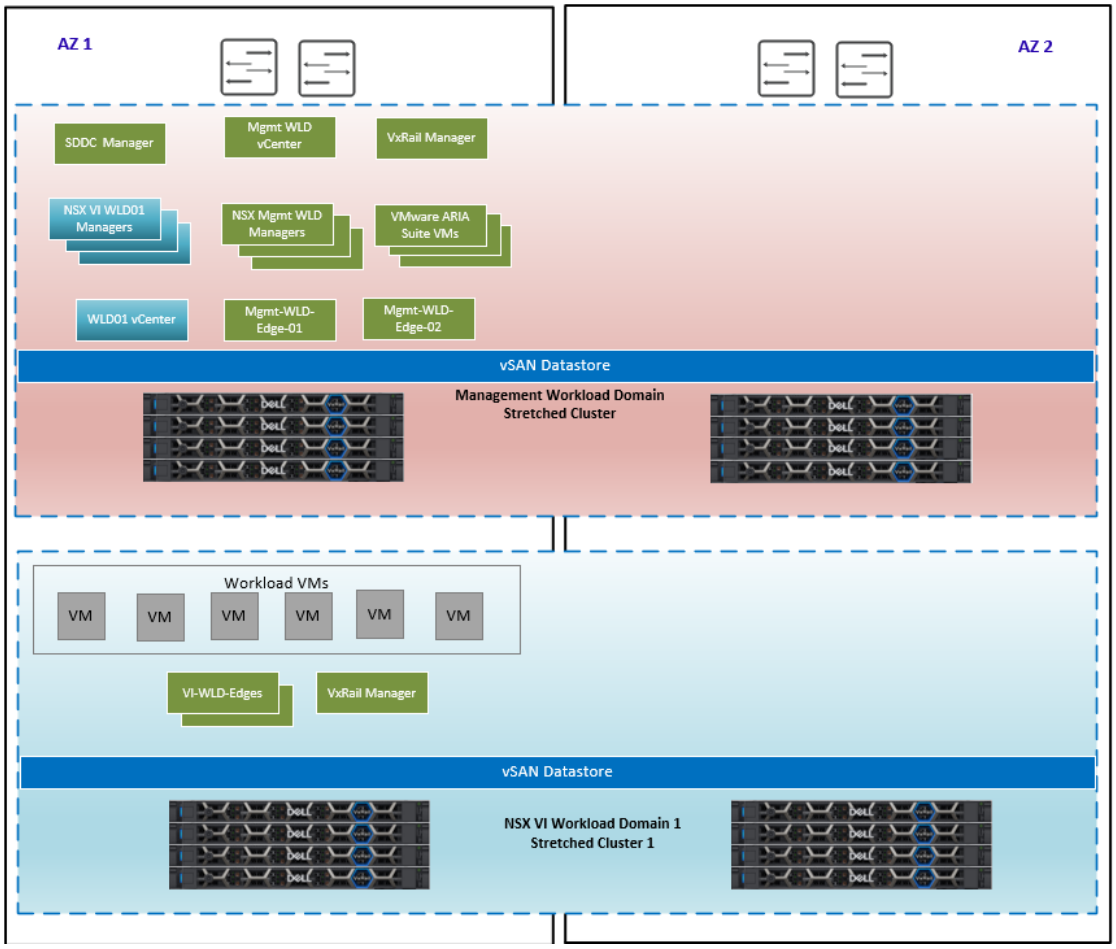


Figure 51. Multi-AZ component layout

Multi-AZ supported topologies

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

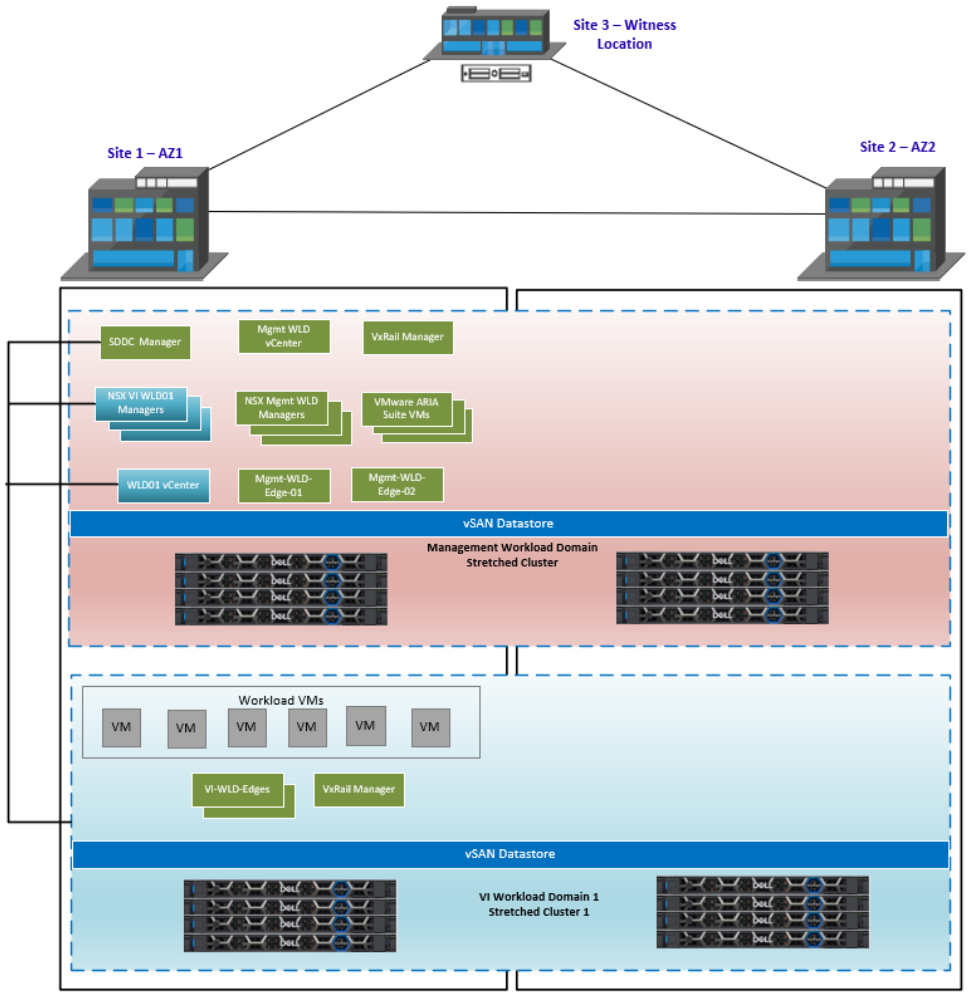


Figure 52. Mgmt and single VI WLD stretched

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

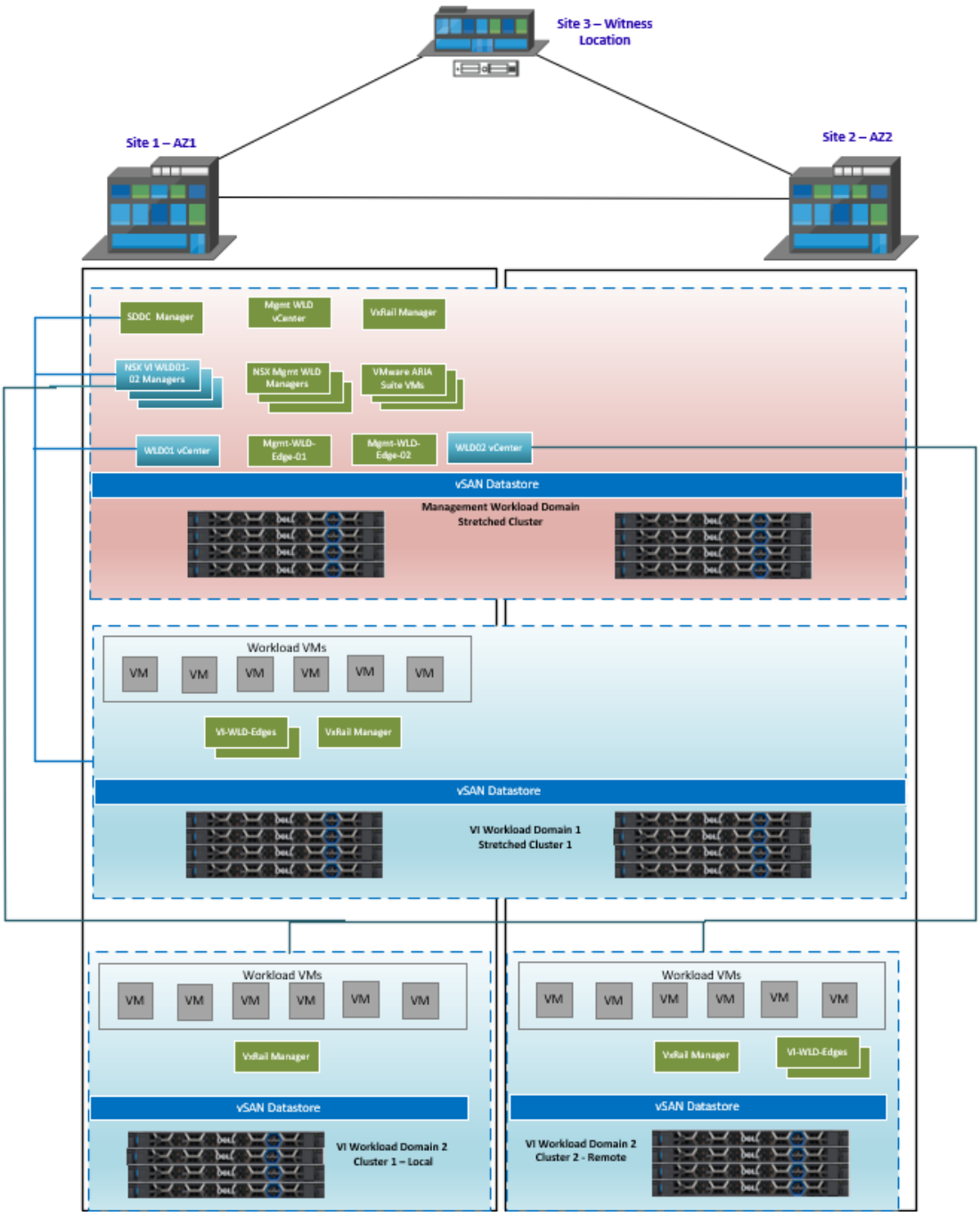


Figure 53. Mgmt and two VI WLD stretched

The next figure illustrates the concept of mixing local and stretched clusters in dual AZ. In this scenario, we have a stretched management WLD and two VI WLDs with a single NSX instance. The first VI WLD has one stretched VxRail cluster, and the second VI WLD has two clusters—one cluster at site 1 and the second at site 2. A dedicated NSX Edge cluster is deployed on the cluster at site 2.

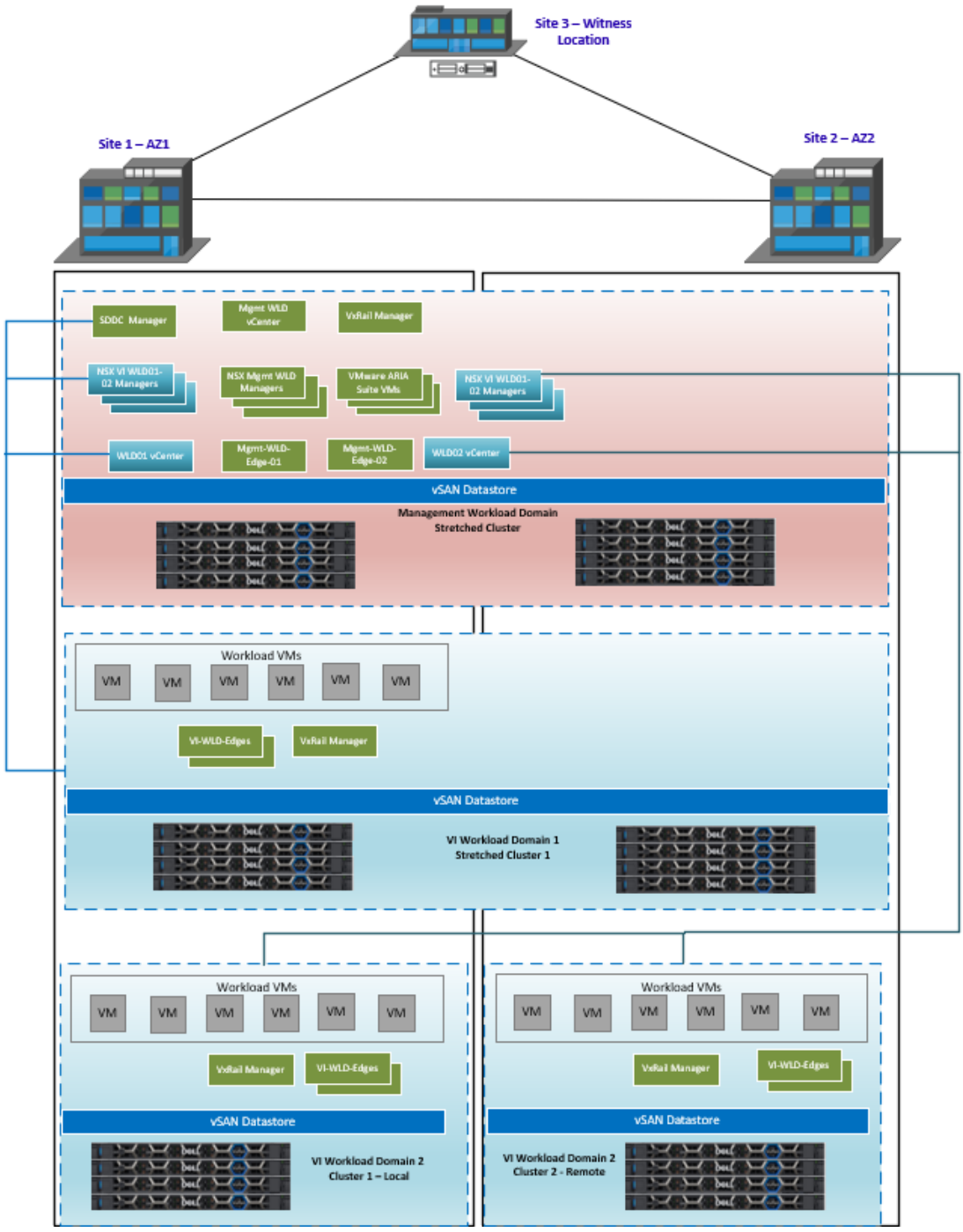


Figure 54. Mgmt and VI WLD01 stretched, nonstretched in VI WLD02

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

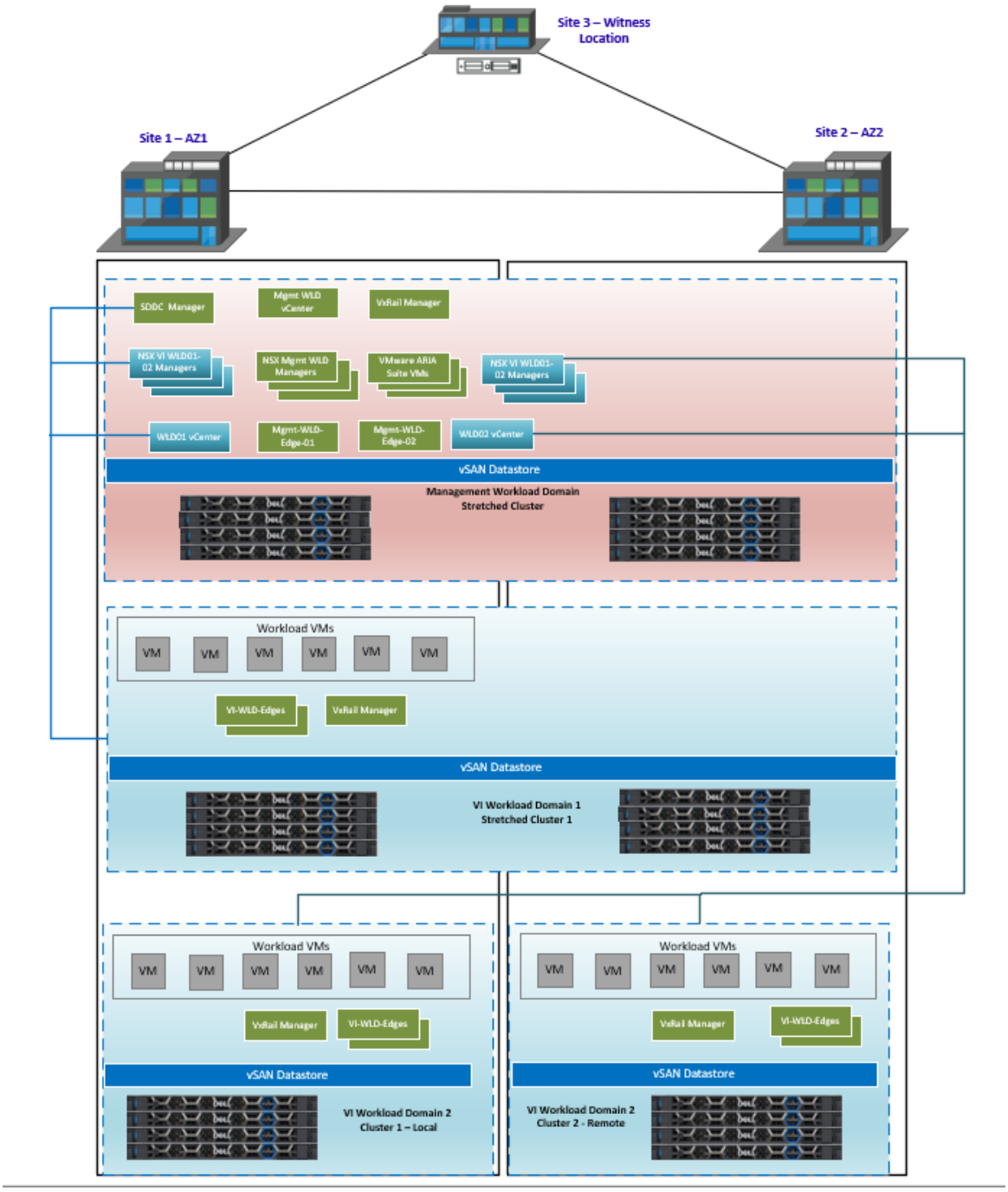


Figure 55. Mgmt and VI WLD01 stretched, nonstretched VI WLD02 with 1:1 NSX

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

As previously mentioned, with AVN overlay networks deployed, the Edge Nodes are deployed and configured to enable the management components in the Aria Suite to use this network. With multi-AZ, the north-south routing that occurs through AZ1 would need to fail over to AZ2 if there is a full site failure. This failover ability is achieved through the addition of the AZ2 TOR switches as BGP neighbors to the Tier 0 gateway so that traffic from the Tier 1 can flow through the TORs at either site. Using both BGP local preference and Path prepend configured on the Tier 0 gateway to steer the traffic out of AZ1 in

normal operating conditions requires manual Day 2 configuration. This configuration is outlined in NSX Data Center Configuration for Availability Zone 2.

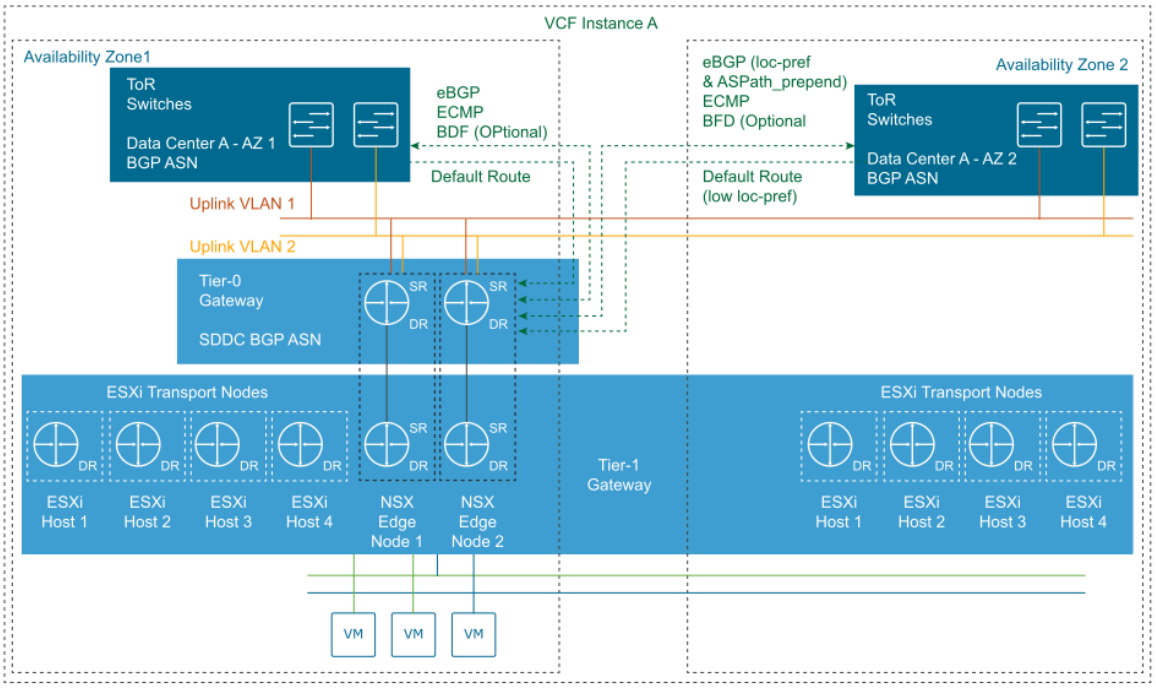


Figure 56. Multi AZ – Mgmt WLD VCF routing design

Multi-site (dual region)

Support for NSX Federation, which is the foundation to support a multisite dual-region deployment. It allows two separate VCF instances in data centers at two different locations in large distance regions to be connected. This connection provides for centralized management, consistent networking, and security policy configuration with enforcement and synchronized operational state. With NSX Federation, VCF can use stretched networks and unified security policies across multi-region VCF deployments, providing workload mobility and simplified disaster recovery. The deployment and configuration are done manually following prescriptive guidance in the VMware VVS documentation.

NSX Global Manager

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

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

NSX Federation requirements

Consider the following additional requirements for an NSX Federation deployment:

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

Dual-region component placement

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

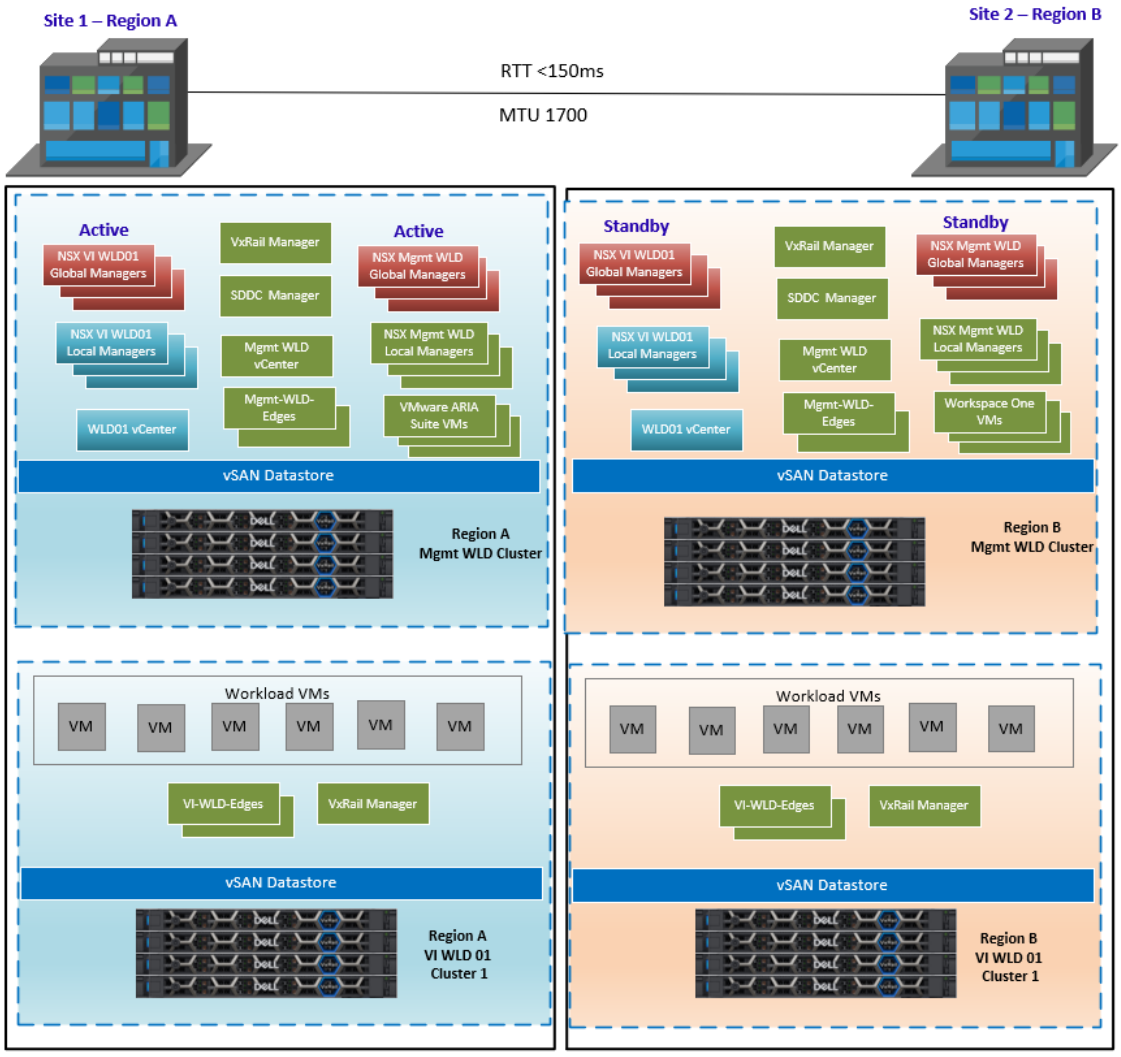


Figure 57. Multi-site – Dual-region NSX Global Manager placement

Inter-region connectivity

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

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

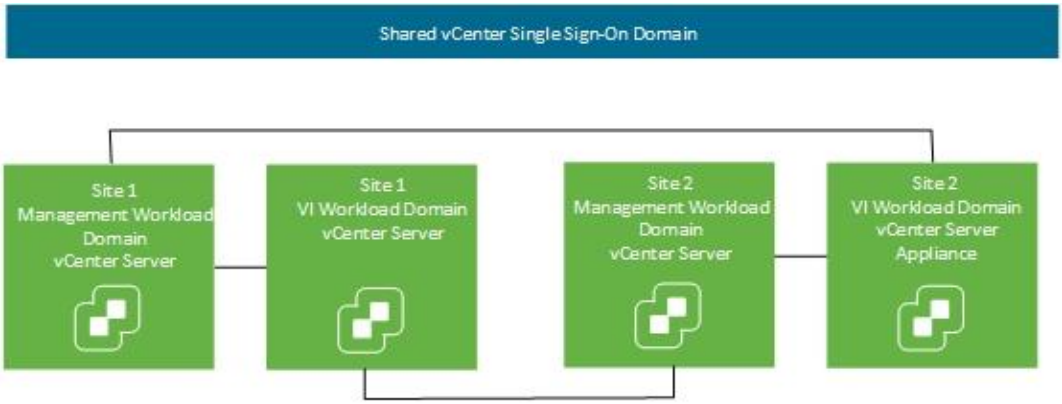
Multi-region routing design

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

LCM considerations

The NSX Global Managers are deployed manually outside of VCF. LCM of these components must be done outside of SDDC Manager because SDDC Manager has no awareness of the Global Managers. The upgrade of the NSX Global Managers must be done using the upgrade coordinator available on the Global Manager appliance. When planning an upgrade of VCF when NSX Federation has been deployed:

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



Future upgrade considerations

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

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

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

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. The Aria Lifecycle Manager VM is deployed from SDDC Manager and used to deploy the Aria suite of components. They are described in more detail in this section.

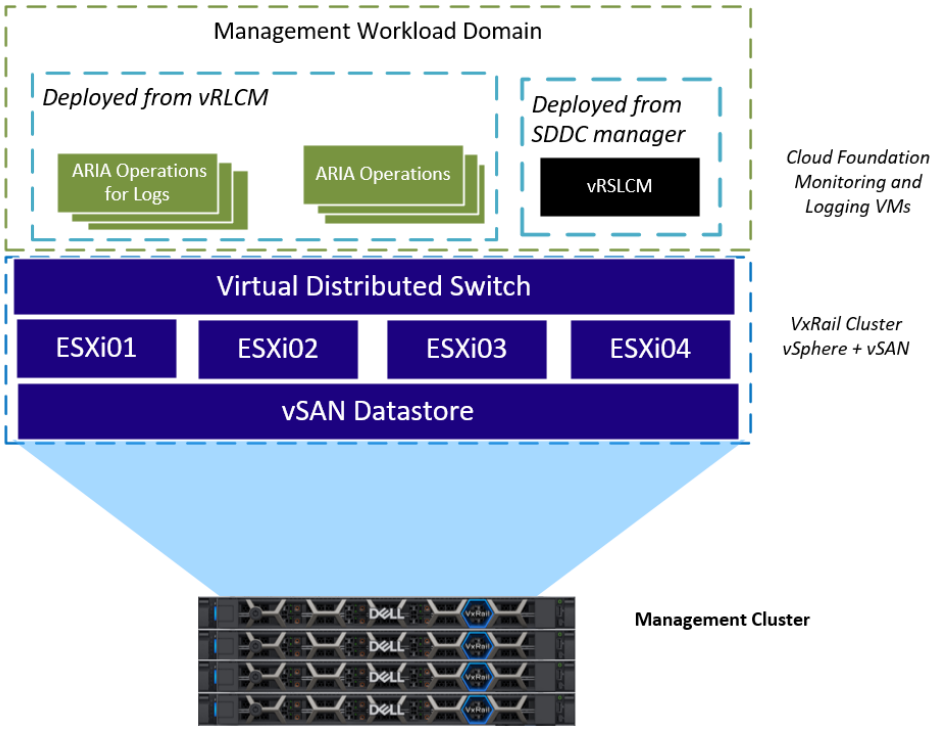


Figure 58. Monitoring and logging operations

VxRail vCenter UI

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

Intelligent Logging and Analytics

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

done through Aria Lifecycle Manager following the VVS deployment guidelines. See Implementation of Intelligent Logging and Analytics for VMware Cloud Foundation.

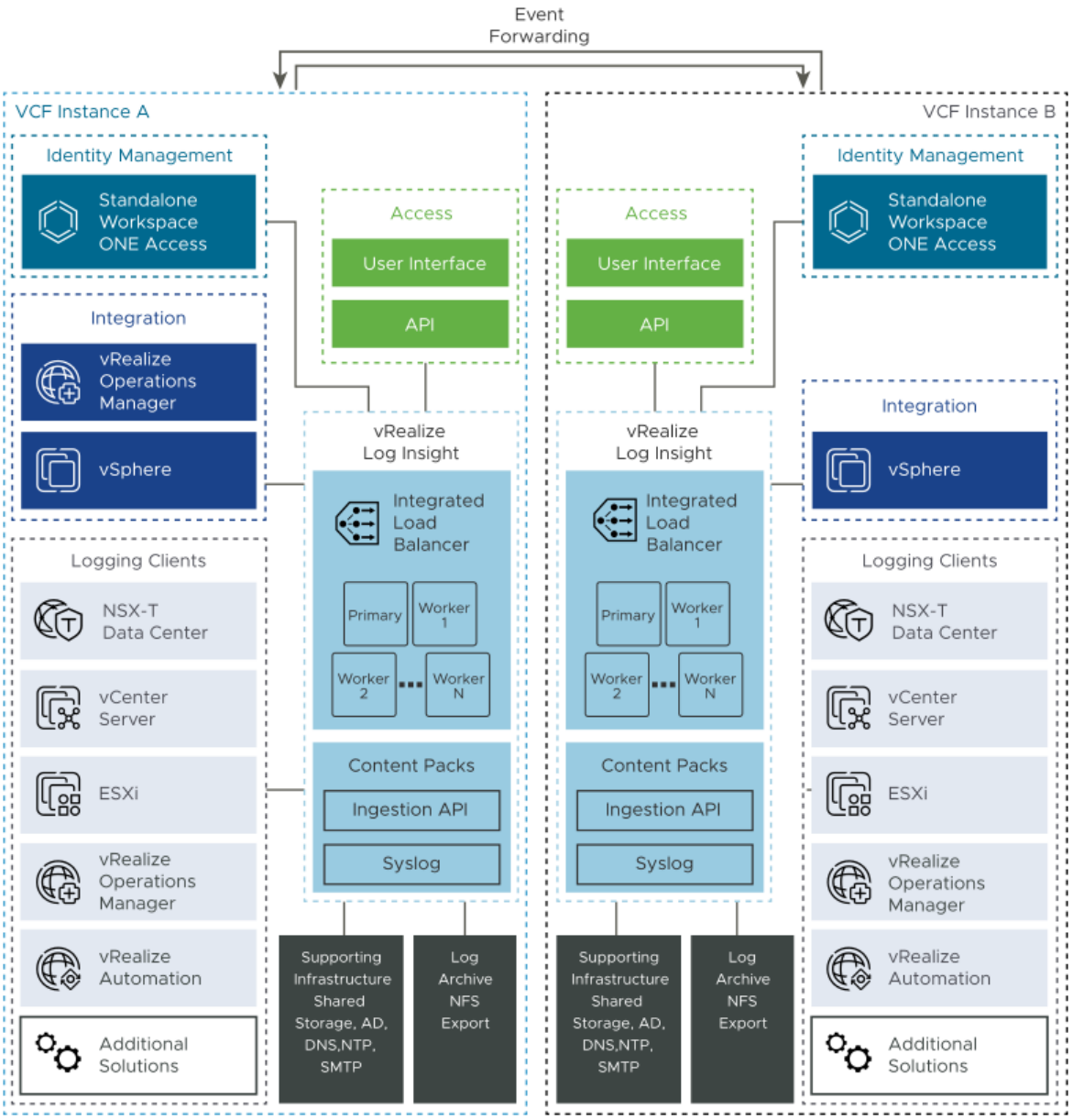


Figure 59. Intelligent Logging and Analytics for VCF

Intelligent Operations Management

The Intelligent Operations Management for VCF validated solution provides centralized monitoring and alerting. It provides the virtual infrastructure or cloud administrator the ability to review and act on events and alerts, through a single interface, to proactively manage system failures. For additional details, see Intelligent Operations Management for VMware Cloud Foundation.

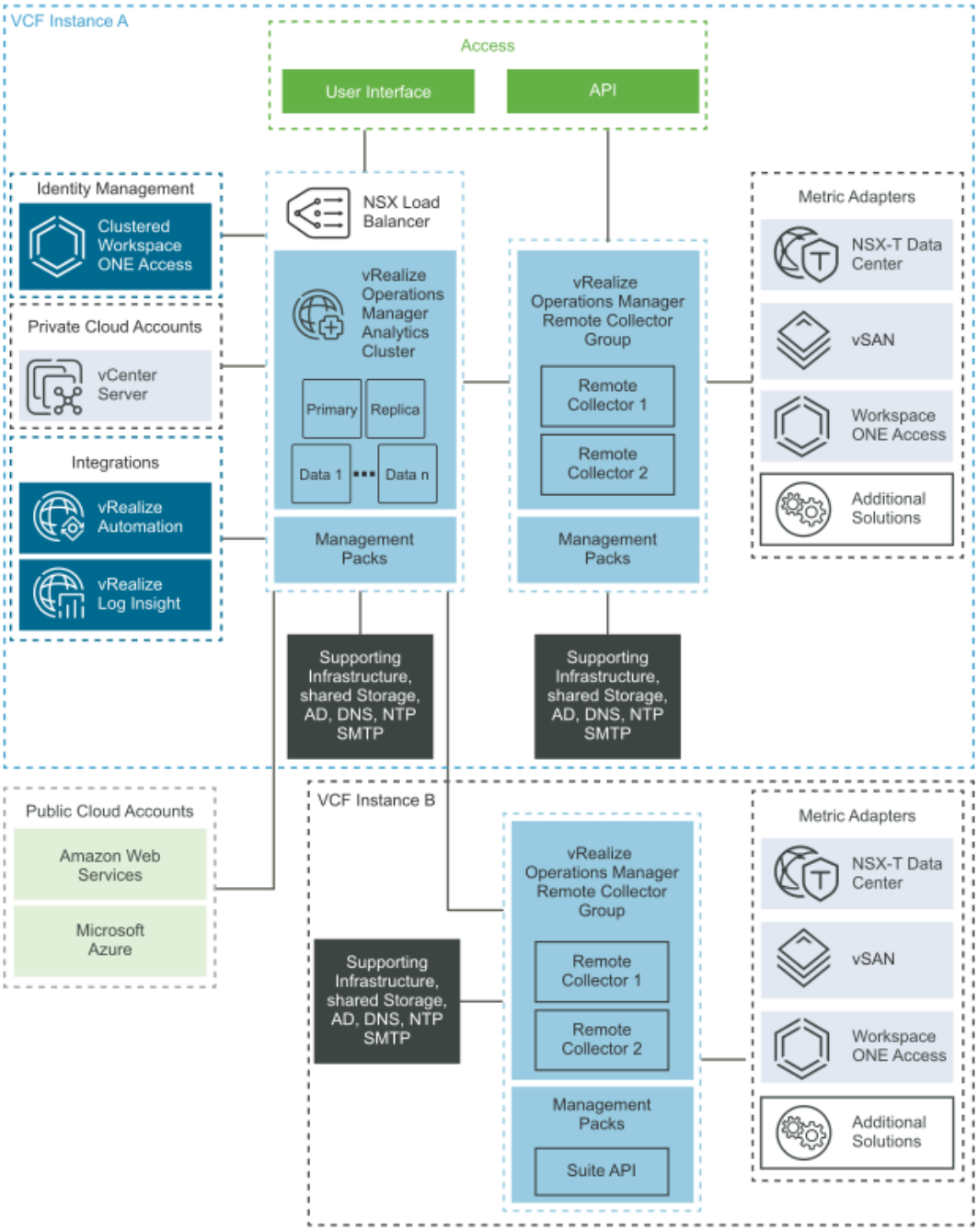


Figure 60. Intelligent Operations Management for VCF

For more details about the design for Aria Operations, see Aria Operations VVD design. The deployment of Aria Operations must be done through Aria Lifecycle Manager in accordance with the VVD deployment guidelines. See Aria Operations Manager Implementation in Region A.

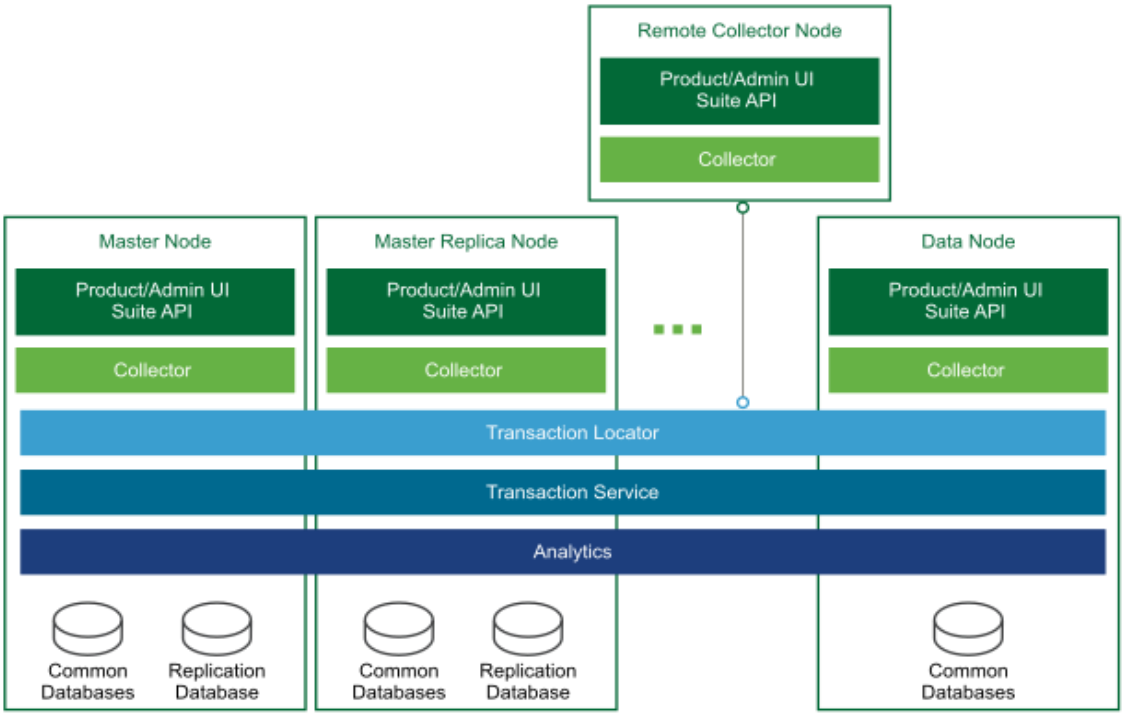


Figure 61. Aria Operations VVS design

Lifecycle management

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

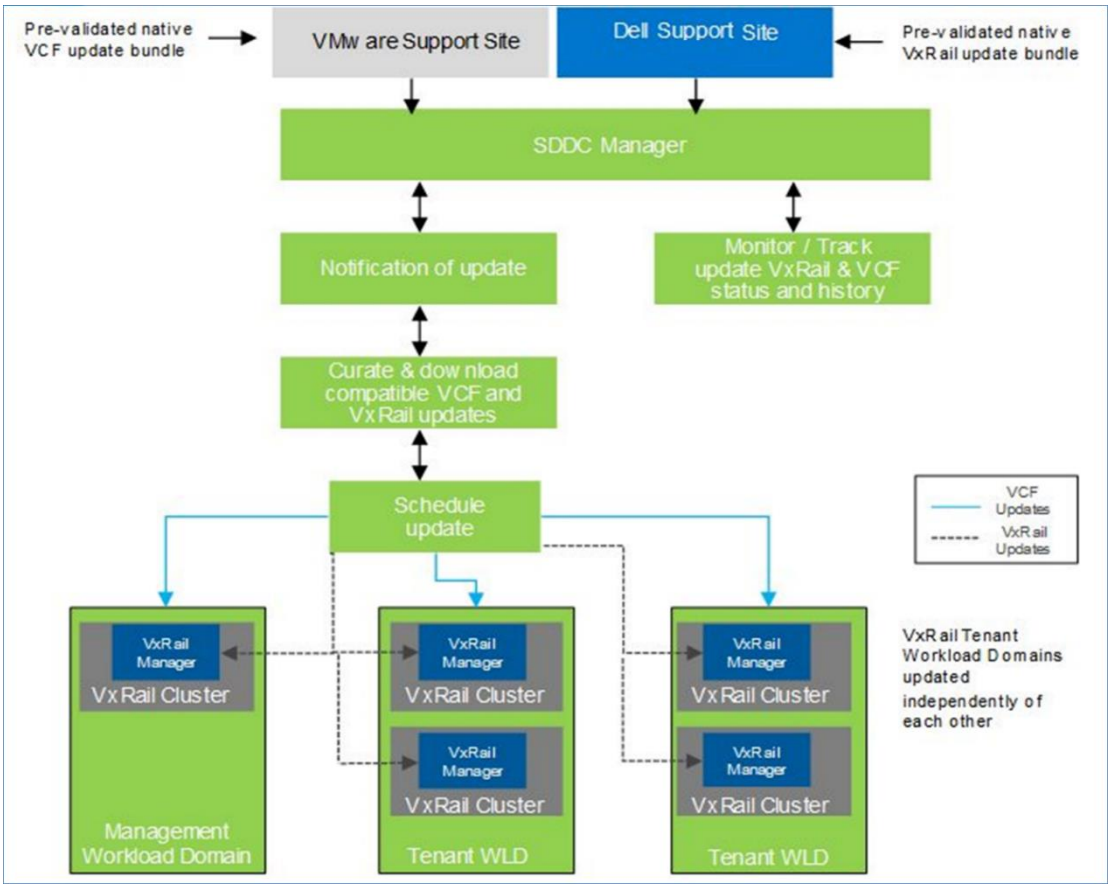


Figure 62. VCF on VxRail LCM components

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

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

Aria Suite Lifecycle Manager

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

Cloud management architecture

Private Cloud Automation for VCF

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

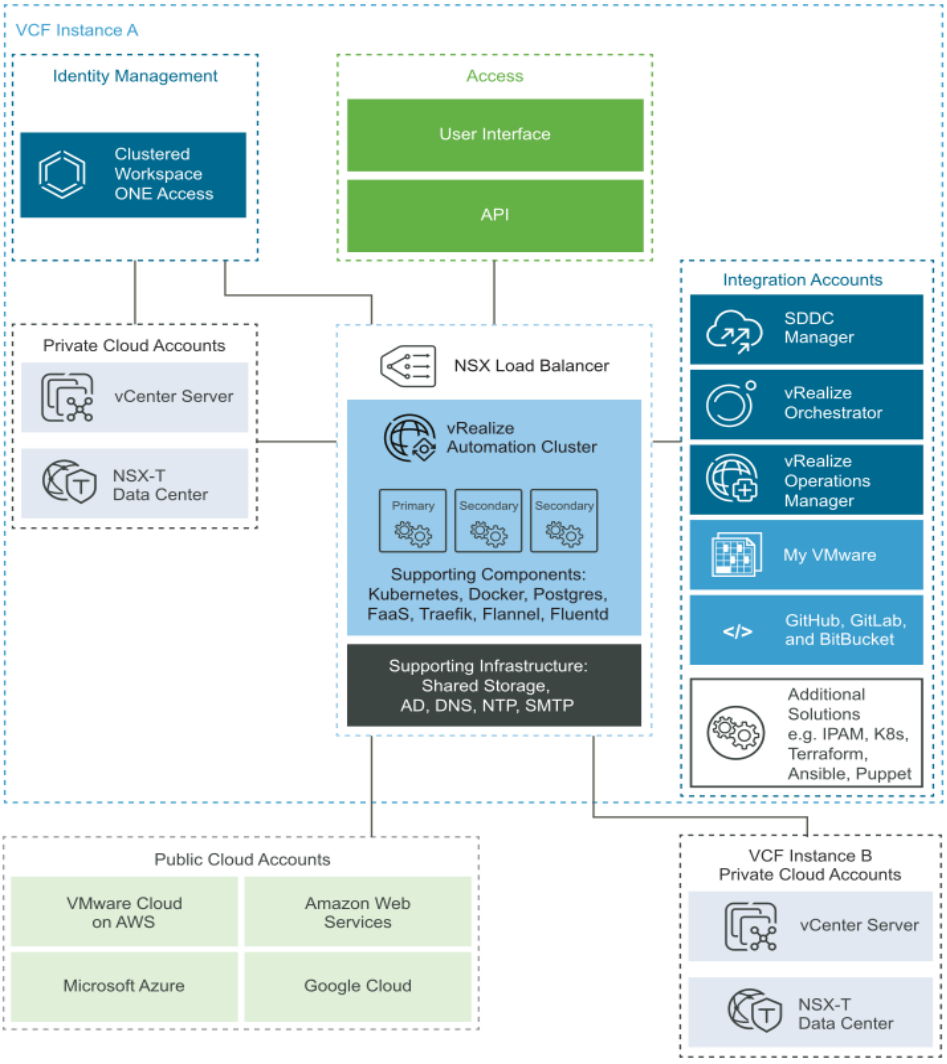


Figure 63. Logical design of Aria Automation

Before you deploy Aria Automation, Aria Lifecycle Manager must be deployed from SDDC Manager. Aria Lifecycle Manager is used to deploy and manage the life cycle of the Aria Suite components.