HITACHI
Inspire the Next

TECHNICAL PAPER

Hitachi Unified Compute Platform 4000 and VMware NSX

Reference Architecture Guide

By Stefan Knobloch, Valentin Hamburger, and Kevin Luksik

August 2016

Hitachi Data Systems

Feedback

Hitachi Data Systems welcomes your feedback. Please share your thoughts by sending an email message to SolutionLab@hds.com. To assist the routing of this message, use the paper number in the subject and the title of this white paper in the text.

Contents

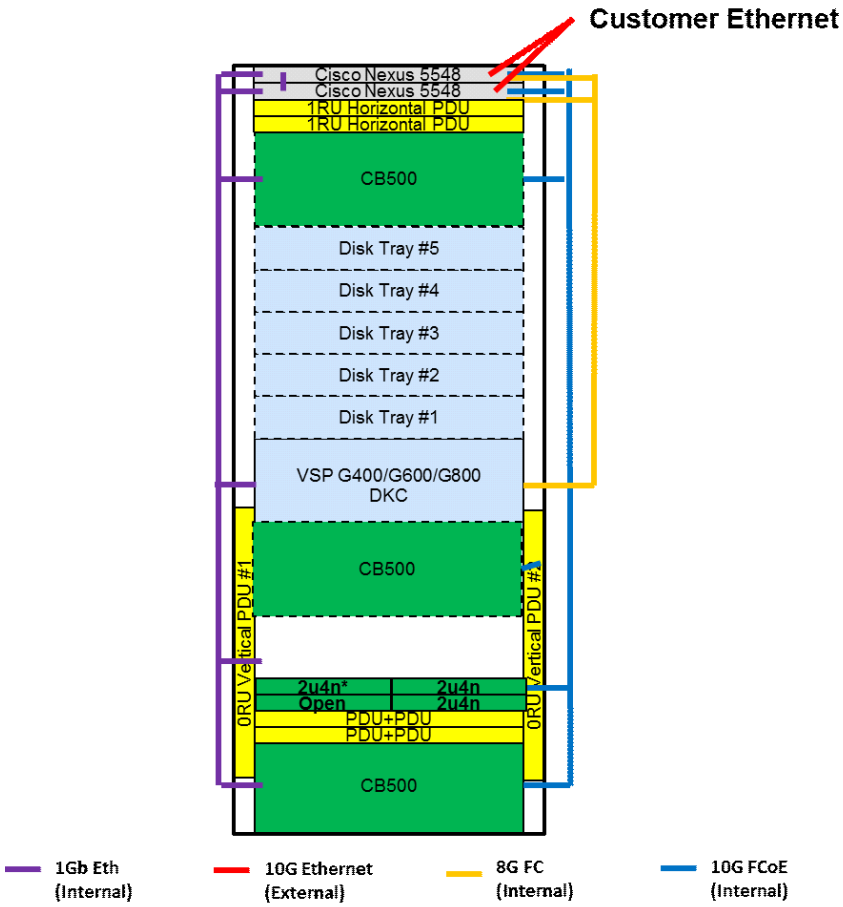
Hitachi Unified Compute Platform 4000 and VMware NSX

Reference Architecture Guide

Hitachi Unified Compute Platform 4000E (UCP 4000E) is a single rack solution that scales up to 24 compute nodes. It is comprised of up to three CB 500 blade chassis, each of which can hold up to eight CB520H server blades with 2×10 Gb/sec NICs.

Special nodes (rack optimized server for solutions, 2U four node) are available to run management VMs. The standard version is equipped with two of these nodes for redundancy, and there is an option to expand the management to three nodes. These nodes come with 2 × 10 Gb/sec NICs.

From a networking perspective, two Cisco Nexus 5548 switches tie all of the nodes together and provide networking services for not just for east-west traffic, but also for north-south traffic. Interfaces dedicated to north-south traffic will server as the link to the customer environment (customer Ethernet).



UCP 4000E with VMware NSX

This section describes vSphere clusters from an NSX perspective.

vSphere Cluster from an NSX Perspective

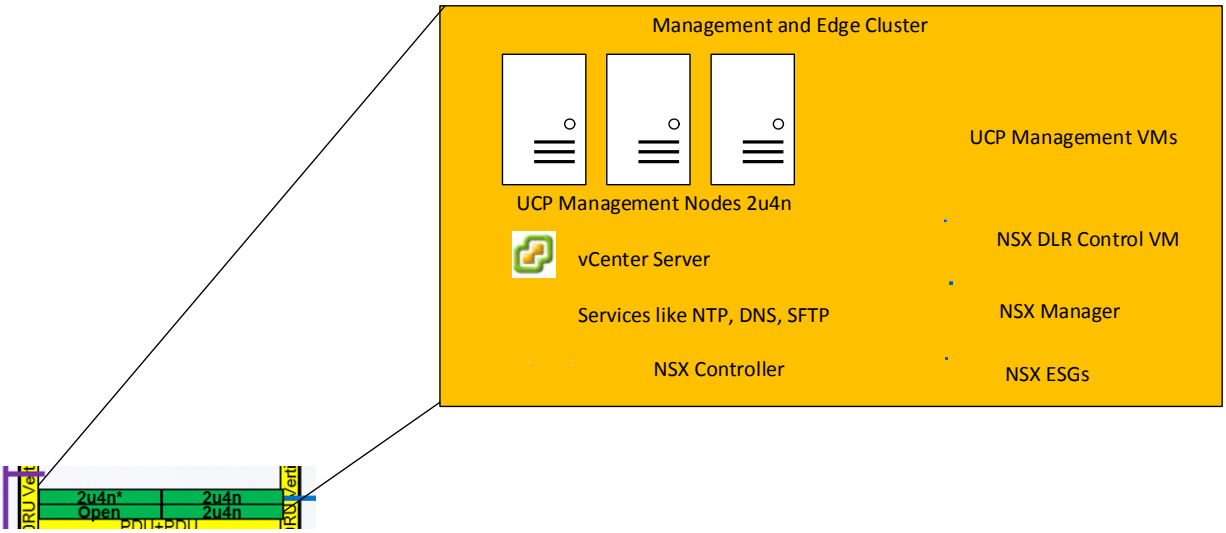
Several clusters exist in NSX that serve different purposes:

- Management Cluster
- NSX Edge Cluster
- NSX Compute Cluster

UCP 4000E management and edge cluster functions can be combined in a single cluster.

UCP Management and Edge Cluster

This design assumes that three UCP management nodes are available.



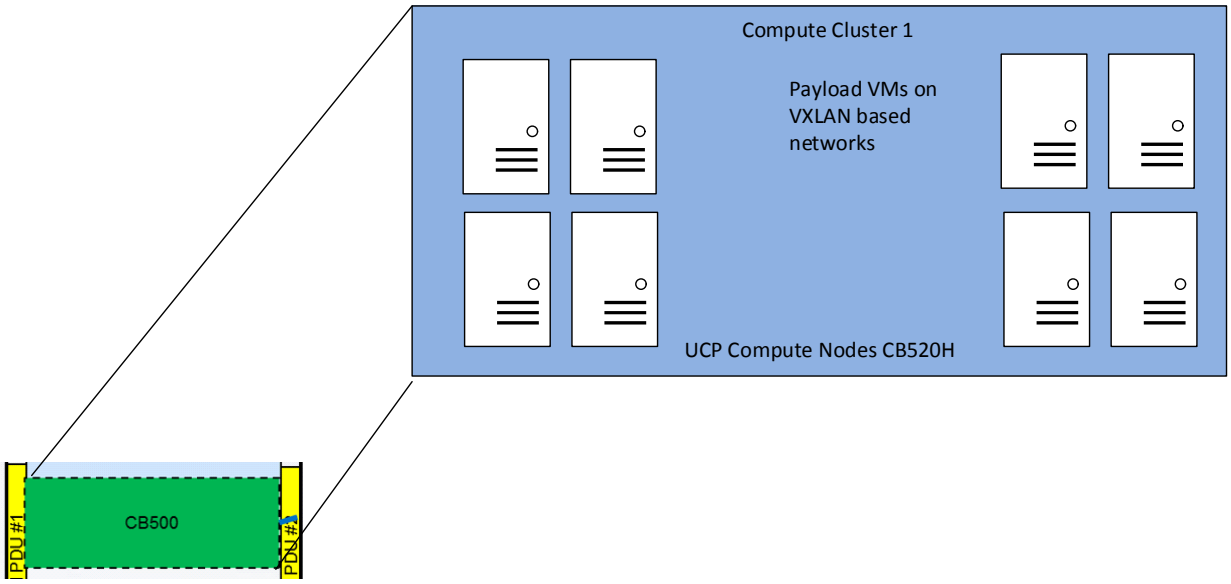
The figure above shows that three management nodes form a vSphere cluster. All management related virtual machines (VMs) are part of this cluster, for example:

- UCP management VMs
- vCenter Server
- Services VMs (NTP, DNS, etc.)
- NSX Manager
- NSX Controller
- NSX ESGs
- NSX DLR Control VMs

UCP Compute Cluster

CB520H server blades are hosted in the CB 500 blade chassis. Up to eight server blades fit into a chassis. UCP 4000E systems scale up to three CB 500 blade chassis with a maximum of 24 blade server systems.

Because NSX management and edge functions are located in UCP management hosts, all hosts in compute clusters can be used for payload systems.



This example shows eight UCP blade servers making one cluster. Clusters in the compute area can be formed according to customer requirements. Thus clusters also can span across hosts in different CB 500 chassis.

vSphere Networking

This section describes ESXi host NICs, vSphere virtual switches, and respective VLANs.

UCP Management Host NIC Driver Settings

UCP management hosts have Emulex NICs installed. For optimal performance, VXLAN offloading should be enabled.

For VXLAN Offload status verification:

- Connect to an ESXi host via SSH
- Enter the following command:

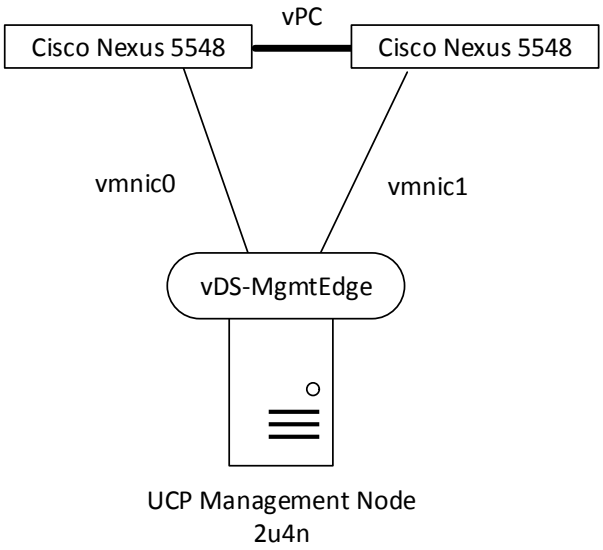
```
# esxcli network nic list
```

- Determine the vmnic# of the NIC to be verified

```
# vsish -e get /net/pNics/vmnic0/stats | grep vxlan
vxlan_offload: true
vxlanUdpPort: 8472
```

vSphere Networking in UCP Management Hosts

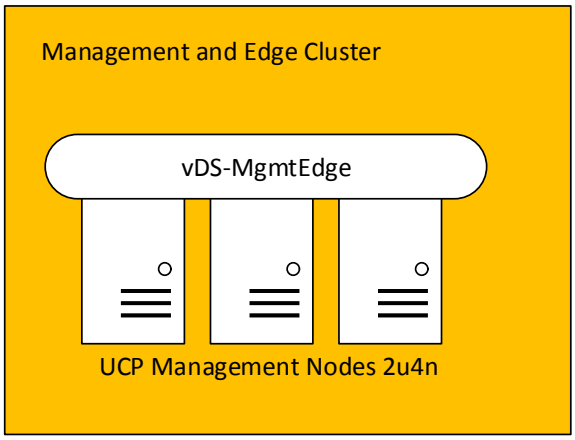
UCP management hosts (rack optimized server for solutions, 2U four node) are equipped with 2 × 10 Gb/sec NICs that are connected to Nexus 5548 top of rack switches.



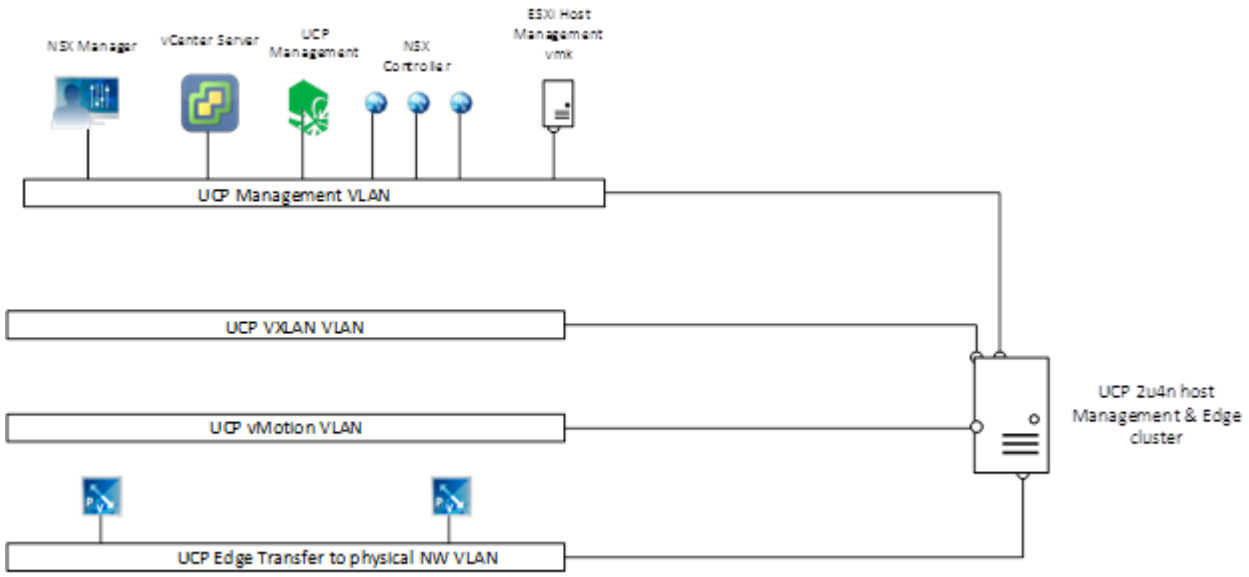
Both vmnics serve as uplinks for the respective virtual distributed switch (vDS). Each UCP management node is connected to the vDS.

On Nexus and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the UCP management VLAN is declared a native VLAN, and therefore traffic is transmitted untagged for this VLAN.

All three UCP management nodes get connected to the same vDS and Nexus 5548, respectively.



These are the typical VLANs that are configured on UCP 4000E management nodes:



UCP management VLAN provides connection for the following:

- UCP management VMs
- All ESXi host management vmkernel ports in all clusters (management and edge cluster as well as compute clusters)
- vCenter Server
- Three NSX Controllers
- NSX Manager

UCP with vMotion VLAN

- All ESXi host vMotion vmkernel ports in all clusters

UCP VXLAN VLAN

- All ESXi host VXLAN vmkernel ports
- Due to the load balancing mode (route based on originating virtual port) two IP addresses are needed per ESXi host.

UCP edge transfer to physical network VLANs

- These VLANs provide connectivity between physical network and logical environments based on VMware NSX.

vSphere Distributed Switch in the Management and Edge Cluster

This section describes settings on the vDS in the management and edge cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

NSX makes use of VXLAN by encapsulating ordinary IP packets in packets with an “outer” header. That is why the size of the original packet increases. VXLAN packets are also IP packets, but the “Don't fragment” bit is set. This is why the MTU has to be increased between ESXi hosts that participate in VXLAN. The physical network has to support the increased MTU and must be configured accordingly.

On the vDS, the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Cisco Nexus switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Cisco Nexus, CDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

Each host has two NICs that can be teamed in different modes that serve different purposes.

Teaming mode is set to “Route based on originating virtual port” so that all NIC resources in an active/active configuration. This way, a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Cisco Nexus switches.

UCP Compute Hosts NIC Driver Settings

UCP compute hosts also have Emulex NICs installed. For optimal performance, VXLAN offloading should be enabled.

For VXLAN Offload status verification:

- Connect to an ESXi host via SSH
- Enter the following command:

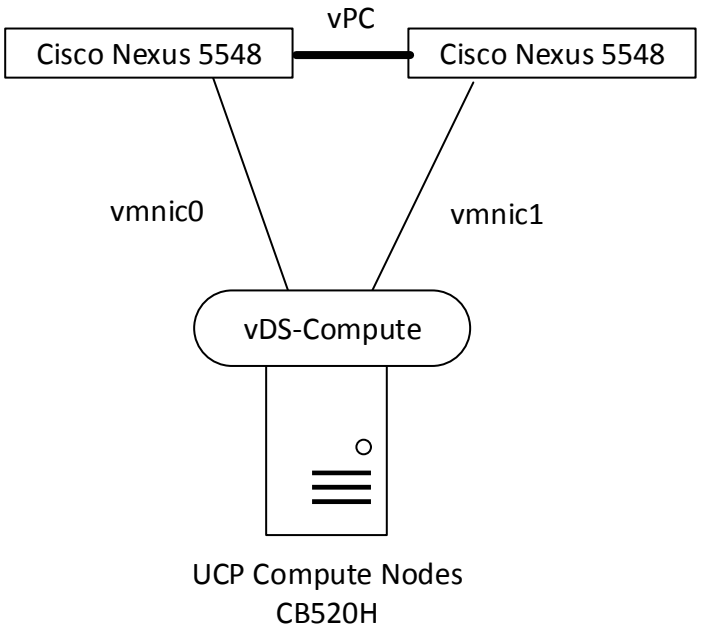
```
# esxcli network nic list
```

Determine the vmnic# of the NIC that is going to be verified

```
# vsish -e get /net/pNics/vmnic0/stats | grep vxlan
vxlan_offload: true
vxlanUdpPort: 8472
```

vSphere Networking in UCP Compute Cluster Hosts

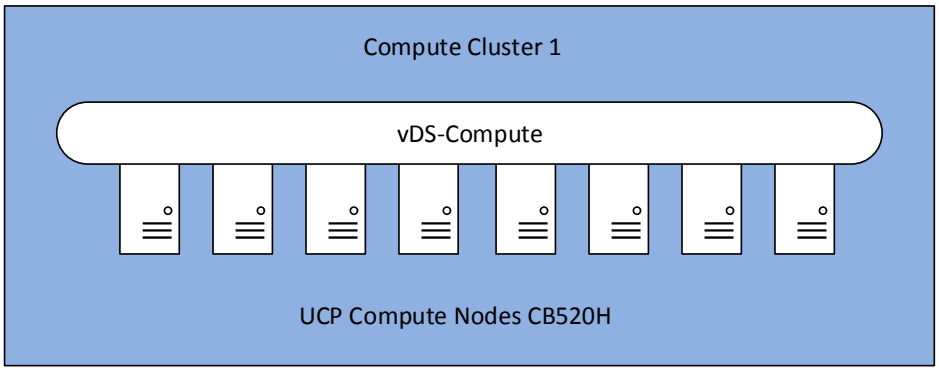
UCP compute cluster hosts (CB520H) are equipped with 2 × 10 Gb/sec NICs that get connected to Nexus 5548 top of rack switches. Pass-Through modules provide the option to connect each host NIC directly with Nexus top of rack switches.



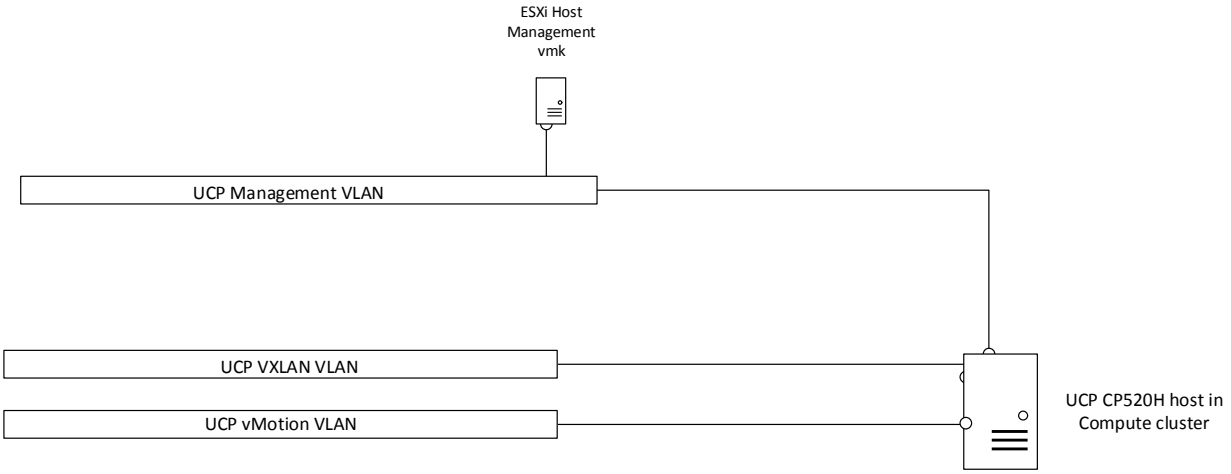
Both vmnics serve as uplinks on the respective virtual distributed switch (vDS). Each UCP compute host gets connected to the vDS.

On Nexus and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the management VLAN is declared a native VLAN, and therefore traffic gets transmitted untagged for this VLAN.

All UCP compute nodes get connected to the same vDS and to Nexus 5548, respectively. The number of compute hosts can scale up to 24. Here just eight hosts are depicted for illustration purposes.



These are the typical VLANs that are configured on UCP 4000E compute nodes:



UCP management VLAN provides connection for the following:

- All ESXi host management vmkernel ports in all clusters (management and edge cluster as well as compute clusters)

UCP with vMotion VLAN

- All ESXi host vMotion vmkernel ports in all clusters

UCP VXLAN VLAN

- All ESXi host VXLAN vmkernel ports
- Due to the load balancing mode (route based on originating virtual port) two IP addresses are needed per ESXi host.

vSphere Distributed Switch in the Compute Cluster

This section describes settings on the vDS in the compute cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

NSX makes use of VXLAN by encapsulating ordinary IP packets in packets with an “outer” header. That is why the size of the original packet increases. VXLAN packets are also IP packets, but the “Don't fragment” bit is set. This is why the MTU has to be increased between ESXi hosts that participate in VXLAN. The physical network has to support the increased MTU and must be configured accordingly.

On the vDS the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Cisco Nexus switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Cisco Nexus, CDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

Each host has two NICs that can be teamed in different modes that serve different purposes.

To ensure that all NIC resources can be used in an active/active manner, teaming mode is set to “Route based on originating virtual port”.

This way a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

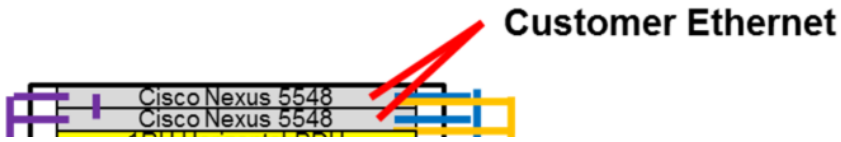
This teaming policy does not require any extra configuration on Cisco Nexus switches.

UCP 4000E and NSX Logical Networking

This section describes a logical networking environment based on NSX, and what it needs to be fully connected to a customer's environment.

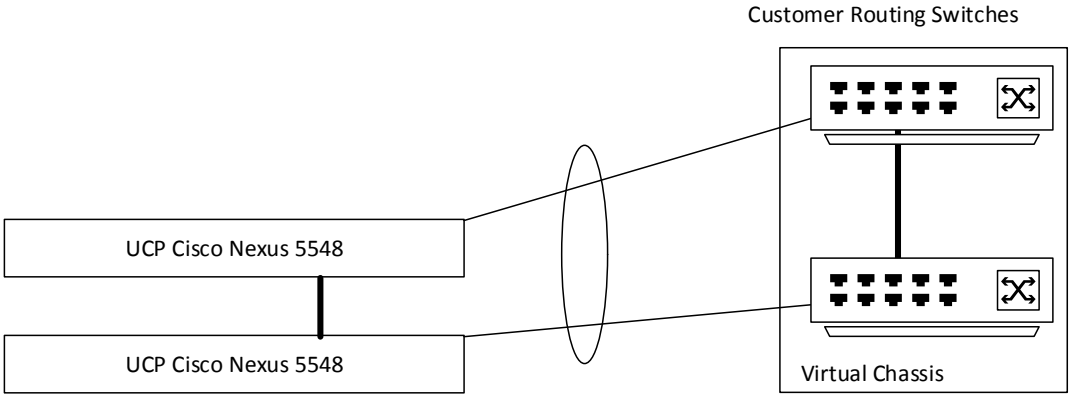
UCP 4000E Connection to Customer Network

The following illustrates the connection between UCP 4000E and the customer network.



The uplinks from Cisco Nexus 5548 need a connection to a customer routing device or devices.

VLANs that have their default gateway on the customer routing switches can be reachable from customer networks. This is mandatory for the UCP management VLAN and Edge transfer VLANs to physical network.



Uplink from Cisco Nexus 5548 to Customer Routing Switches is grouped in an LACP-based (Link Aggregation Control Protocol) PortChannel. This setting has to be agreed upon with the customer.

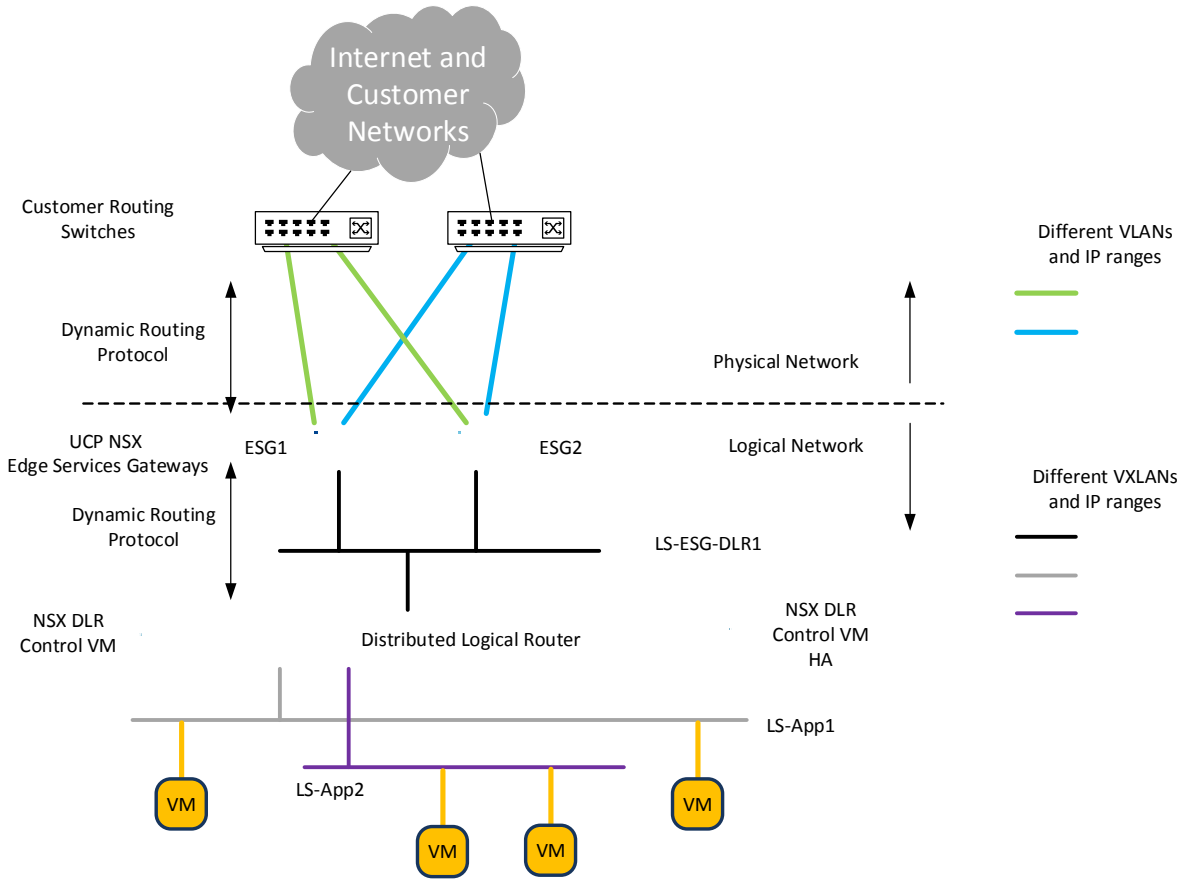
On the transfer VLAN between NSX Edge Service Gateways and the customer routing switch, a dynamic routing protocol can be used to propagate dynamically created networks in NSX.

NSX Logical Networking

It is important to understand that NSX provides a routed environment that needs to be integrated in a customer's given network infrastructure.

UCP 4000E provides layer 2 connectivity to a customer's network. For routing to work smoothly, it needs to be adopted to the customer preferred routing protocol - OSPF or BGP. If the customer switches can provide static routing only, a connection to the logical network of NSX can still be set up. But, of course, the limitations that come along with static routing apply.

The following logical design assumes that dynamic routing is available on customer switches. The logical networking will look like this:



Starting from the top, customer networks and customer routing switches represent the infrastructure that UCP 4000E and NSX need to integrate into.

From a layer 3 perspective, NSX Edge Service Gateways and customer routing switches peer and exchange routing information via a dynamic routing protocol. Ordinary VLANs are the means of transport for the routing protocol. Each NSX ESG will have two interfaces that are used to provide connection to two customer routers.

On a UCP 4000E, two NSX ESGs are deployed and operate in ECMP (Equal Cost Multi Path) mode. This provides for redundancy and high throughput.

Both NSX ESGs will also have an interface connected to a logical switch (LS) that allows for connectivity to an NSX Distributed Logical Router (DLR). ESGs and DLR also use a routing protocol to update each others' routing tables.

Additional logical switches (LS-App1; LS-App2) are attached to the DLR depending on customer needs. Each logical switch has its own IP subnet, and routing between directly attached IP subnets occurs automatically. The DLR acts as the default gateway for virtual machines attached to the respective logical switches.

One major feature of the DLR is the fact that it is distributed across all ESXi hosts. So each host has a vmkernel module that performs the DLR function. That means routing happens on the host in case a VM wants to communicate with another VM on a different logical switch attached to the same DLR.

If a VM wants to communicate to a device in the physical world, packets are routed via DLR and ESG to the customer network.

Building NSX Infrastructure

This section explains what needs to be done to set up a logical environment on a UCP 4000E.

It is assumed that vSphere networking and clustering has been set up already.

Required services such as DNS and NTP are provided by UCP management VMs.

NSX Manager

NSX Manager is deployed into management and edge clusters on UCP rack optimized server for solutions, 2U four node servers. It is attached to the same port group as vCenter and all ESXi host management vmkernel NICs.

It consumes

- 16 GB RAM
- 60 GB of HDD
- 4 vCPUs
- 1 IP address

It is linked to the vCenter and the SSO domain.

NSX Controller

Three NSX Controllers are rolled out to the same port group as the NSX Manager and ESXi hosts are connected to. They come in a VM form factor and are being copied from NSX Manager during rollout.

Each NSX Controller consumes:

- 4 GB RAM
- 20 GB HDD
- 4 vCPUs
- 1 IP address

IP address assignment is configured by setting an IP Pool in NSX Manager that contains 3 IP addresses.

ESXi Host Preparation or VIB Installation

During the ESXi host preparation process VIBs are being copied from NSX Manager and installed on each host. They perform functions to support VXLAN, DLR, and distributed firewall.

The host preparation process works on an ESXi cluster basis. So all hosts in a cluster are being prepared simultaneously.

All clusters (compute and management, and edge) are prepared for NSX and get the necessary VIBs.

VXLAN Configuration

During VXLAN configuration the following is being determined:

- What vDS will be used for VXLAN
- Which VLAN ID should be used for VXLAN traffic
- What is the MTU size
- What method should be used for VTEP IP address assignment
- Which teaming policy shall be used for uplinks

There is a direct link between the teaming policy and the number of IP addresses needed for VTEPs (Virtual Tunnel End Points) per host.

This example has a teaming policy of “route based on originating virtual port” so the amount of IP addresses changes to 2 per ESXi host.

An IP Pool is configured with at least 57 IP addresses (24x CB520H + 3x 2u4n = 27 ESXi hosts. Two IP addresses/host means 57 IP addresses in total).

The IP address planning has to take this into account.

MTU is set to 9000 bytes, which is the maximum for vSphere.

When VXLAN configuration is complete for all clusters, each ESXi hosts will have two vmkernel ports that are used for vxlan only.

VXLAN IDs or Segment IDs or Virtual Network Identifier (VNI)

In the VXLAN header, 24 bits are available for VXLAN IDs. This allows for more than 16 000 000 different IDs. After host preparation is complete and VXLAN has been configured, the Segment ID pool can be defined.

IDs that do not overlap with the IDs of other VXLAN installations in the customer environment are recommended.

VXLAN IDs are similar to VLAN IDs. A different IP subnet can reside in each VXLAN. The number of VXLAN IDs depends on customer requirements.

The UCP 4000E comes with these settings, which allow for 100000 VXLANs:

- Segment ID pool: 100000-199999
- Multicast addresses: 239.40.0.0-239.41.255.255

Multicast addresses are being used within a transport zone with hybrid mode replication. This will be explained in the next section.

For the hybrid mode to work properly, the physical network devices need to support the following:

- IGMP snooping
- IGMP snooping querier

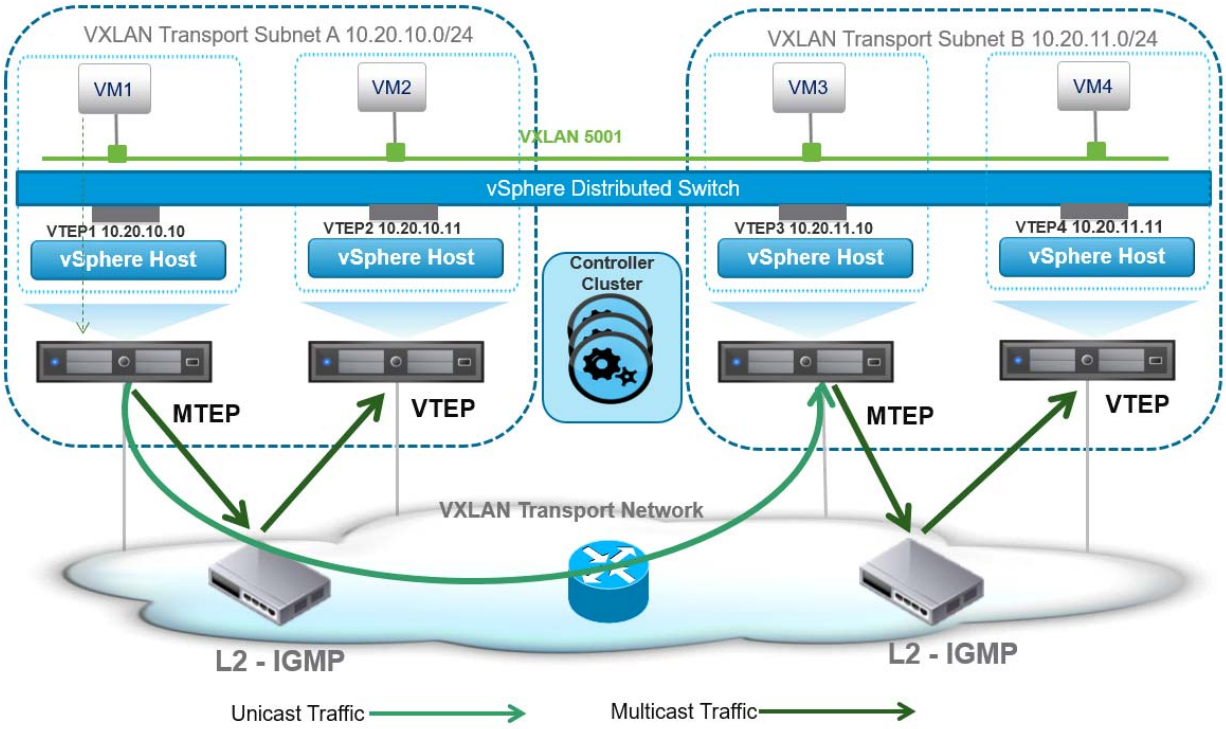
UCP Cisco Nexus devices do so, and are configured accordingly.

VXLAN Transport Zone

A Transport Zone can be considered as groups of hosts that can communicate with each other on given logical switches or VXLAN IDs. Or in other words logical switches are deployed in a transport zone. All ESXi hosts/clusters that are part of this transport zone can use these logical switches to connect VMs to.

Replication of Broadcast, Unknown Unicast, and Multicast (BUM) Traffic

In the figure below, if VM1 sends a broadcast packet it will be delivered to all VMs in the same layer 2 domain, which is the same logical switch.



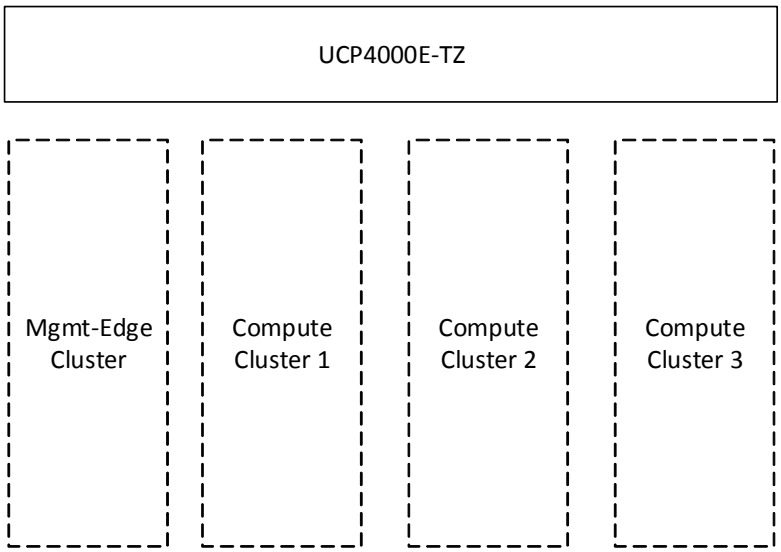
In Hybrid Mode, the ESXi host sends a multicast packet in the VXLAN VLAN. All other ESXi hosts that have VMs attached to the same logical switch have joined this multicast group, receive the packet, and forward it to the respective VMs.

If layer 3 boundaries have to be crossed, a unicast frame is sent to the other IP segment that is also part of VXLAN.

In UCP 4000E there is only a layer 2 VXLAN and therefore no layer 3 boundaries to cross.

Hybrid Mode saves the CPU power of the ESXi host for other tasks in contrast to Unicast Mode. In Unicast Mode the ESXi hosts sends unicast packets to each other host individually in order to ensure that the broadcast of the VM is being delivered.

In UCP 4000E, Hybrid Mode is being used for the transport zone, and only one transport zone includes all clusters.



NSX Logical Switches

NSX Logical Switches create layer 2 domains across all ESXi hosts in UCP 4000E.

There are two purposes for using logical switches:

- Logical Switches for Infrastructure
- Logical Switches for Payload VMs

Logical Switches for Infrastructure

A logical switch that connects NSX DLR and ESGs is used for infrastructure purposes only. No customer VMs are supposed to be connected to this network.

Logical Switches for Payload VMs

A logical switch that connects customer or payload VMs with the network. It is up to the customer to create logical switches according to configuration requirements.

The DLR will get an interface in these logical switches, and it acts as the default gateway for the attached VM.

NSX Distributed Logical Router (DLR)

There is one NSX DLR that provides routing services between directly connected logical switches and the northbound located NSX Edge Services Gateways.

It is comprised of two components:

- The distributed part provides routing services on each ESXi host.
- The part that runs routing protocols is located in a NSX Distributed Router Control VM. This VM is operated in “High Availability” mode, which means there is a second VM that just exchanges heartbeats with the active DLR Control VM. If the active VM fails, the standby VM becomes active.

The DLR Control VM consumes these resources:

- 512 MB RAM
- 512 MD HDD
- 1 x vCPU

NSX Edge Service Gateways (ESGs)

There are two NSX ESGs that provide connectivity between the physical customer infrastructure and NSX logical networks.

These ESGs are operated in Equal Cost Multi Path (ECMP) mode, which is a feature of a dynamic routing protocol. Thus both ESGs are active, and if one ESG fails the remaining ESG will forward the traffic. The customer's routing environment should support ECMP also, in order to maximize throughput.

NSX ESG can be deployed in different sizes, and a large NSX ESG consumes the following resources:

- 1GB RAM
- 512 MB HDD
- 2 x vCPU

Anti-Affinity Rules for DLR and ESGs

These are some guidelines for anti-affinity rules for routing devices:

- NSX DLR Control VM

Active and standby DLR Control VM should be located on different ESXi hosts.

- NSX ESGs

ESGs should be located on different ESXi hosts.

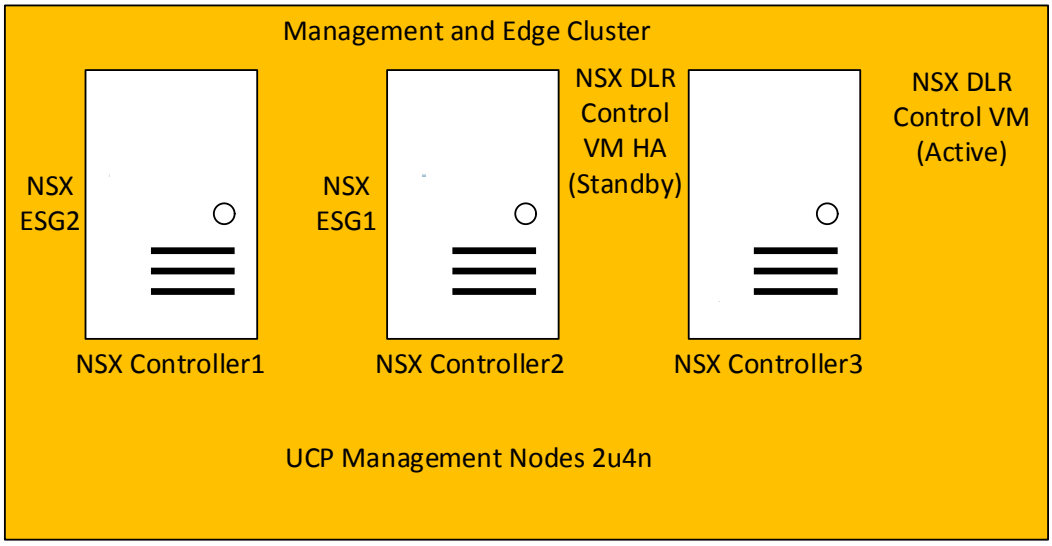
- DLR Control VM and ESG with dynamic routing neighborships

Active DLR Control VM should not be located on the same host as an ESG that it has a routing neighborship with.

There are options for DLR Control VM deployments that can be taken into account for anti-affinity rules.

NSX ESGs and DLR Control VMs in the Management & Edge Cluster

Active devices are distributed across hosts in the management and edge cluster look like this:



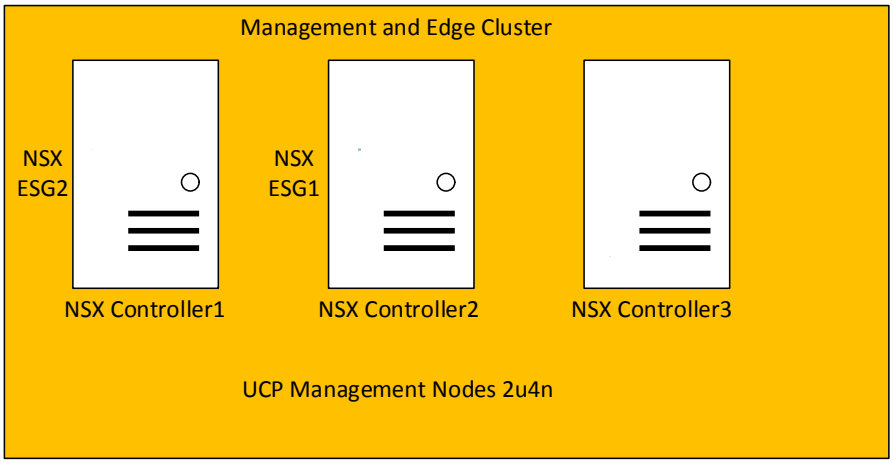
As can be seen in the figure above all three conditions can be met.

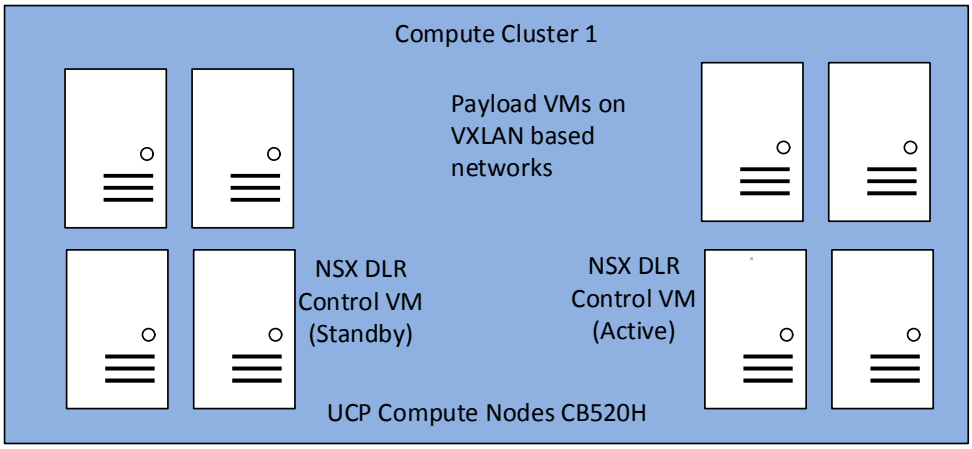
However, if the active and standby DLR control VMs switch functions, they will not automatically switch back.

DLR Control VMs in a Compute Cluster

To avoid the situation mentioned in the previous paragraph, deploy the DLR Control VM in a compute cluster.

This will lead to a separation of VMs as shown below.



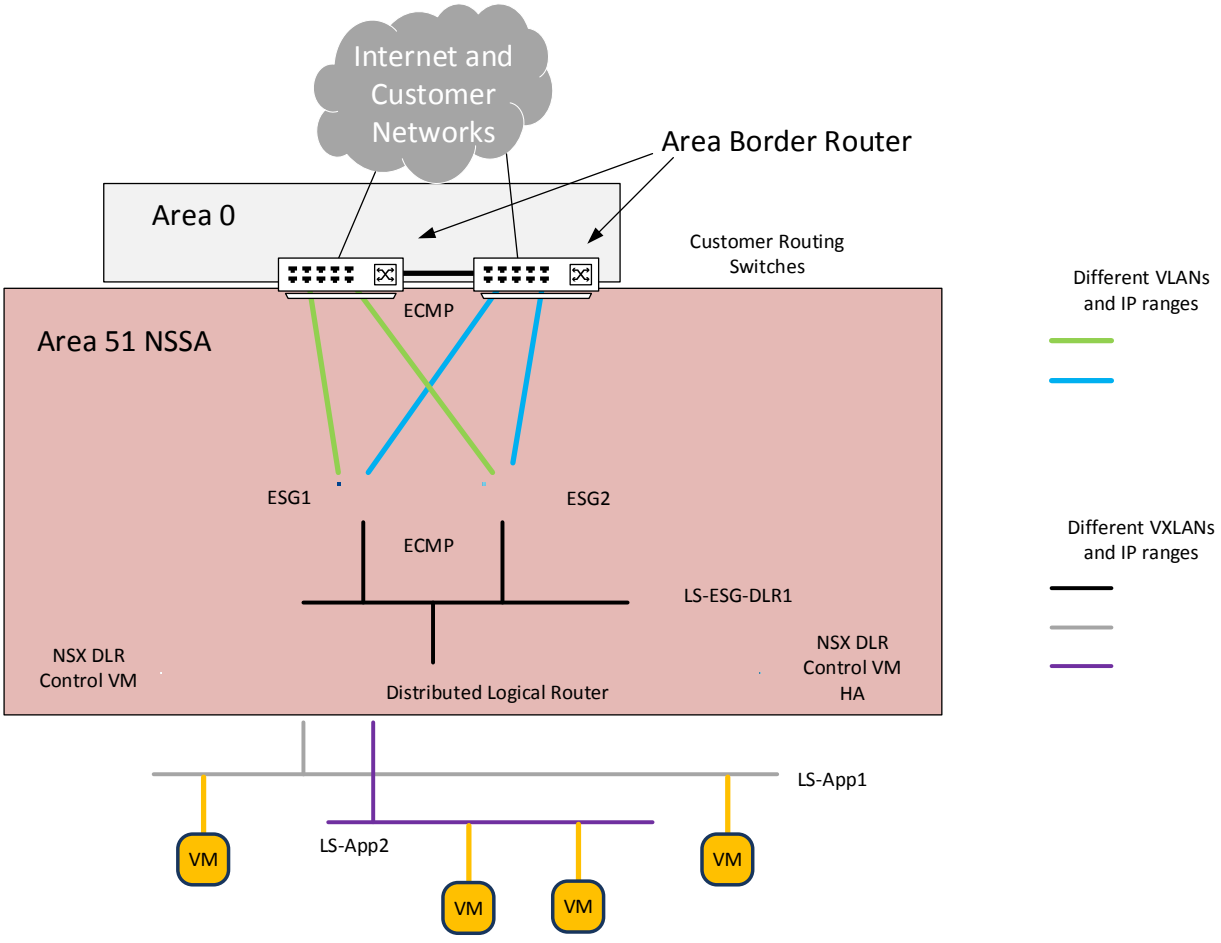


The DLR Control VM only has interfaces in VXLAN-based networks (Logical Switches). Because these networks are available on compute clusters as well, there is nothing to be taken into account from a physical network design.

Dynamic Routing between Customer Switches and NSX

Dynamic routing needs to be agreed upon with customer since it is important to have a common understanding on how it works and what needs to be done to achieve performance goals.

The figure below shows how OSPF areas are being set up and which routing device belongs to which area.



Customer routing switches perform the tasks of Area Border Routers. They are connected to Area 0 (Backbone) and the interfaces pointing towards NSX ESGs belong to area 51 which is a Not So Stubby Area (NSSA).

NSX ESGs and DLR belong to area 51 exclusively.

Networks attached to DLR are propagated via OSPF to ESGs and customer routing switches.

Customer routing switches can inject a default route into area 51.

DLR will have two ways (ESG1 and ESG2) to reach all networks that are not directly attached. Customer routing switches will also have two ways (ESG1 and ESG2) to reach networks connected to DLR.

Both ways have the same associated cost. Thus DLR and customer routing switches can make use of ECMP (Equal Cost Multi Path) and distribute traffic across both ways.

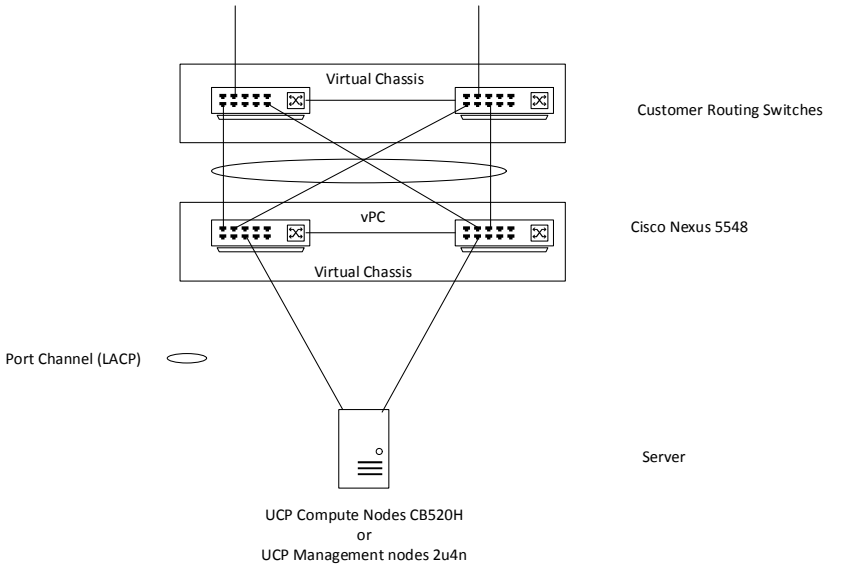
This provides for the following:

- Higher throughput between physical and logical networks
- Quick failover in case one ESG or customer routing switch should go down

Customer Routing Switch and Multi-Chassis Trunking

Prior to implementing the routing environment shown previously, the fact that customer routing switches actually support routing over bundled links that span several chassis has to be confirmed.

For example, Cisco Nexus 7000 switches operated in vPC mode do not support OSPF over vPC.



OSPF Timers in ECMP Mode

In OSPF, routers are supposed to establish neighborships with other routers in the same network. In order to do so, OSPF timers have to be identical in all routers.

In ECMP mode, OSPF timers can be set aggressively (short) in order to quickly react to changes in the network.

Table 1. OSPF Intervals

	Default (seconds)	Supported by NSX (seconds)
OSPF Hello Interval	10	1
OSPF Dead Interval	40	3

NSX Edge Design and Scaling

When it comes to designing a logical infrastructure, the question comes up regarding how much throughput is needed between the physical and logical world.

As a rule of thumb it can be assumed that an NSX ESG provides 7-9 Gb/sec throughput. Scaling up can be achieved by adding additional NSX ESGs. Keep one NSX ESG per ESXi host.

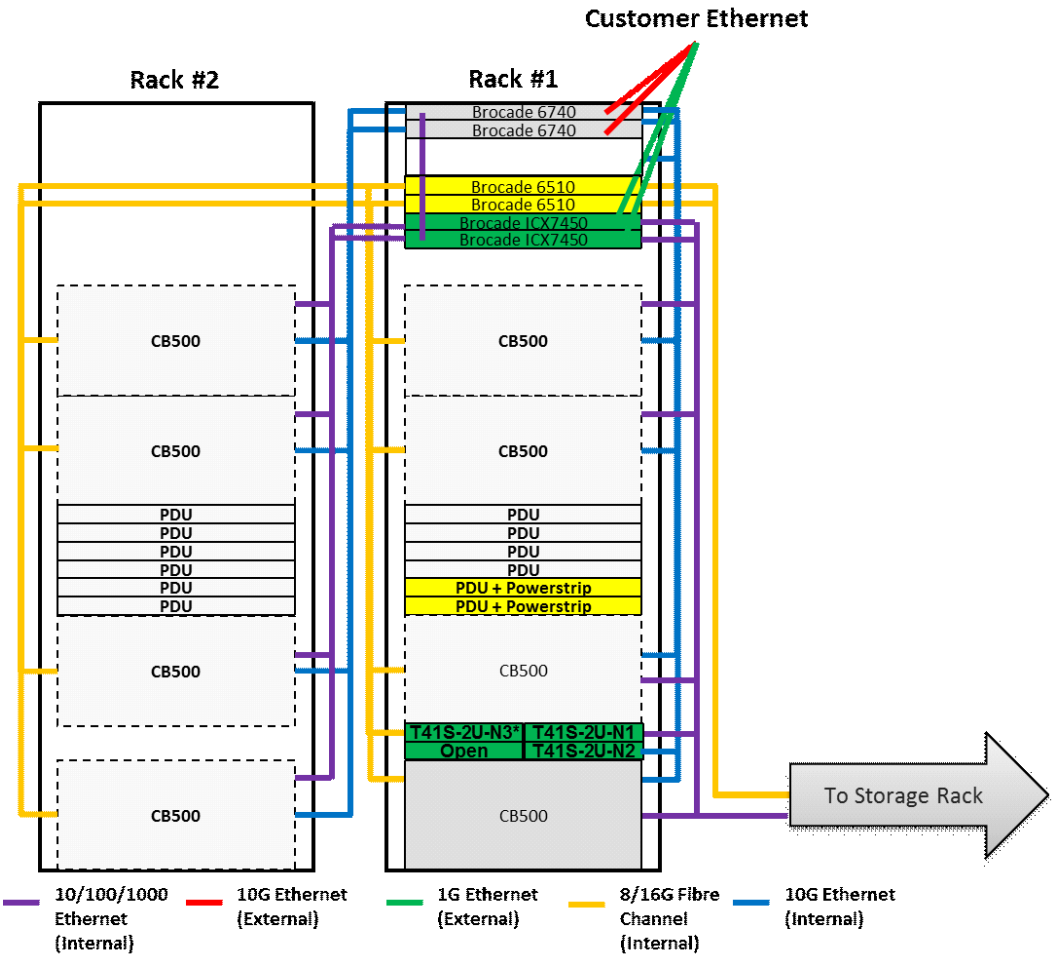
In UCP 4000E two NSX ESGs are present that operate in a load balancing mode by using OSPF ECMP.

A third NSX ESG can be deployed on the third ESXi host in the management cluster.

The management cluster cannot be expanded on a UCP 4000E system. If more throughput or NSX ESGs are needed, the Edge cluster can be setup using CB520H hosts in a compute cluster.

UCP 4000 with Brocade Architecture

The UCP 4000 with Brocade architecture scales up to two racks and 64 compute nodes.



Rack number 1 contains three UCP management nodes (rack optimized server for solutions, 2U four node) equipped with 192 GB RAM and a dual-port Intel 82599 10 GigE Open Compute Project NIC), two Brocade VDX6740 Ethernet Fabric switches for high performance switching, and two Brocade ICX 7450 switches for UCP management.

One rack can hold up to four CB 500 blade chassis, and up to 8 eight CB520H server blades with 2 × 10 Gb/sec NICs can be inserted into a CB 500 chassis.

This way the first rack can scale up to 32 server blades. A second rack with up to 4 blade chassis can be added that scales the solution to up to 64 server blades.

UCP 4000 with Brocade and VMware NSX

This section describes vSphere clusters from an NSX perspective.

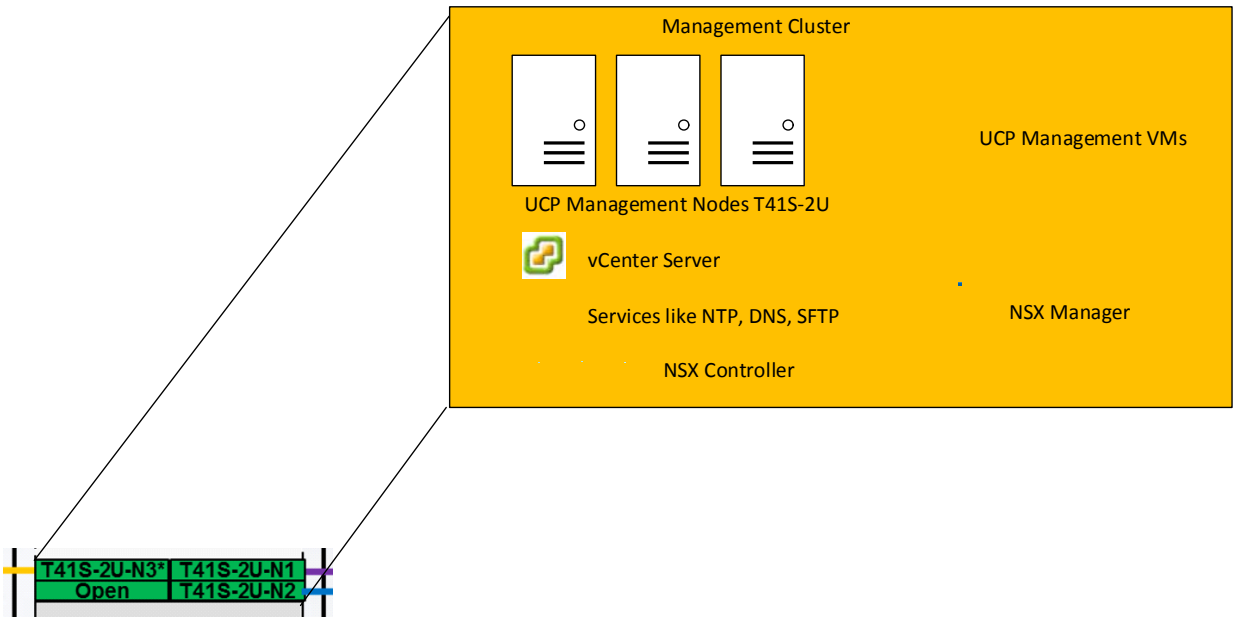
vSphere Cluster from an NSX Perspective

Several clusters exist in NSX that serve different purposes:

- Management Cluster
- NSX Edge Cluster
- NSX Compute Cluster

Brocade management and edge clusters are separate for UCP 4000.

UCP Management Cluster



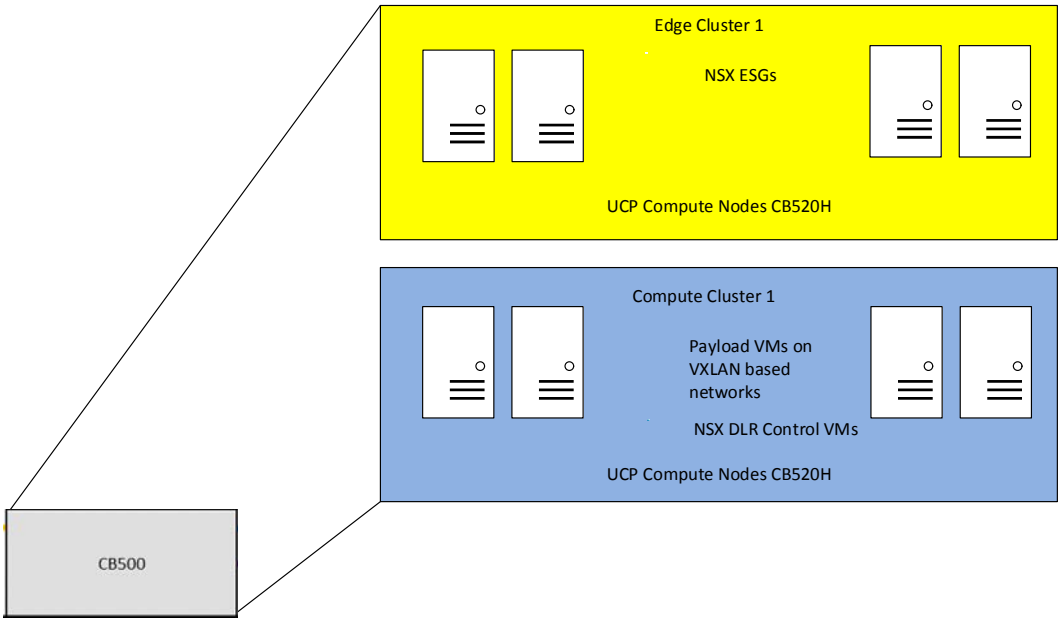
The figure above shows three management nodes that form a vSphere cluster. All management related VMs are part of this cluster, for example:

- UCP management VMs
- vCenter Server
- Services VMs (NTP, DNS, etc.)
- NSX Manager
- NSX Controller

UCP Edge and Compute Cluster

CB520H server blades are hosted in a CB 500 blade chassis. Up to eight servers fit into a chassis. UCP 4000 with Brocade systems scale up to eight CB 500 blade chassis with a maximum of 64 blade server systems.

In order to provide sufficient bandwidth between logical networks and the physical world, the NSX Edge cluster is located in the first CB 500 chassis and is made up of four CB520H hosts.



The remaining four hosts of the first CB 500 chassis can be grouped together to form a compute cluster.

Part of this compute cluster is the NSX DLR Control VM and its HA counterpart.

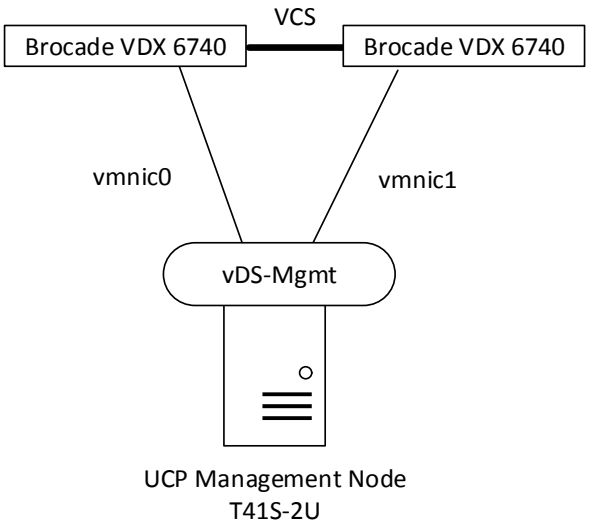
vSphere Networking

UCP Management Hosts NIC Driver Settings

Because there is no VXLAN in the UCP management cluster, no special settings need to be considered.

vSphere Networking in UCP Management Hosts

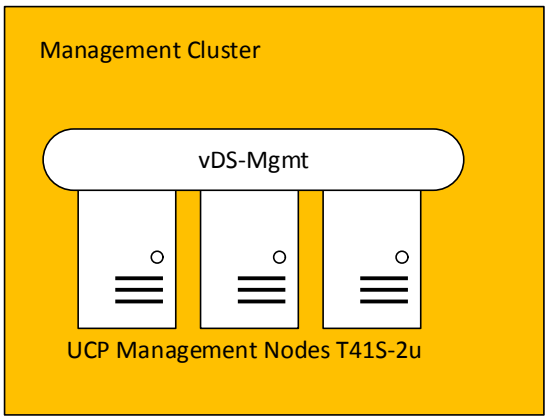
UCP management hosts (rack optimized server for solutions, 2U four node) are equipped with 2 × 10 Gb/sec NICs, that are connected to Brocade VDX6740 top of rack switches.



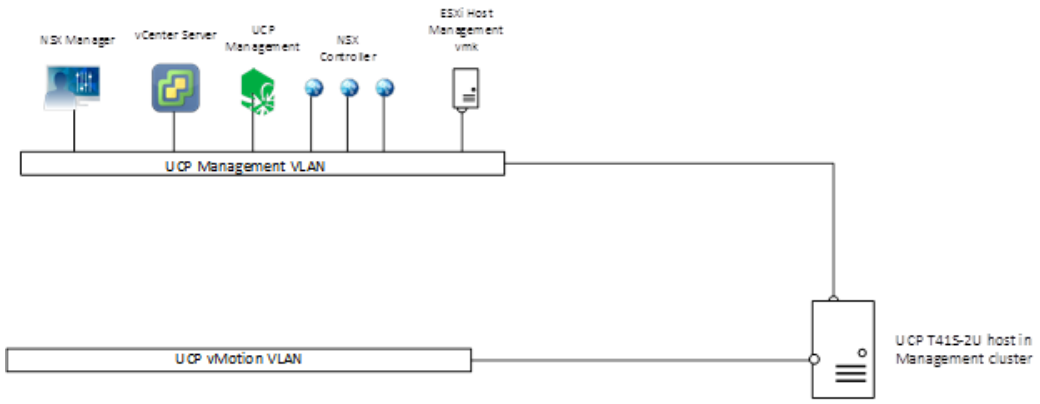
Both vmnics serve as uplinks for the respective virtual distributed switch (vDS). Each UCP management node is connected to the vDS.

The VDX6740 and vDS uplink ports are configured for trunking, which means VLANs get tagged. However, the UCP management VLAN is declared a native VLAN, and therefore traffic gets transmitted untagged for this VLAN.

All three UCP management nodes get connected to the same vDS and to Brocade VDX6740 switches, respectively.



These are the typical VLANs that are configured on UCP 4000 with Brocade management nodes:



UCP management VLAN provides connection for the following:

- UCP management VMs
- All ESXi host management vmkernel ports in all clusters (Management cluster, Edge cluster as well as compute clusters)
- vCenter Server
- Three NSX Controllers
- NSX Manager

UCP with vMotion VLAN

- All ESXi host vMotion vmkernel ports in all clusters

vSphere Distributed Switch in Management Cluster

This section describes settings on the vDS in Management cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

There is no VXLAN in the management cluster. However, since the MTU is set to 9000 for all other vDS (Edge and Compute), it is set to 9000 on the vDS management cluster as well.

On the vDS the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Brocade VDX switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Brocade VDX6740, LLDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

Each host has two NICs that can be teamed in different modes that serve different purposes.

To ensure that all NIC resources can be used in an active/active manner, the teaming mode is set to “Route based on originating virtual port”. This way a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Brocade VDX6740 switches.

UCP Edge Hosts NIC Driver Settings

UCP edge hosts also have Emulex NICs installed in them. For optimal performance, VXLAN offloading should be enabled.

For VXLAN Offload status verification:

- Connect to an ESXi host via SSH
- Enter the following command:

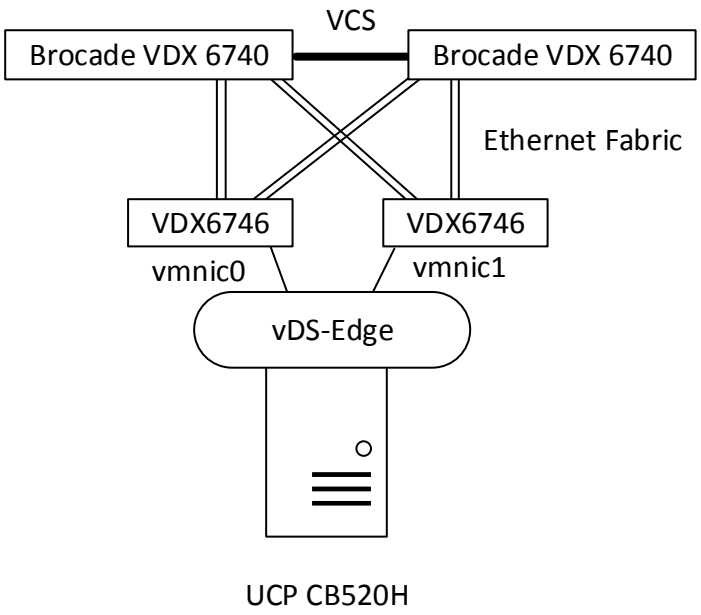
```
# esxcli network nic list
```

- Determine the vmnic# of the NIC that is going to be verified

```
# vsish -e get /net/pNics/vmnic0/stats | grep vxlan
vxlan_offload: true
vxlanUdpPort: 8472
```

vSphere Networking in UCP Edge Cluster Hosts

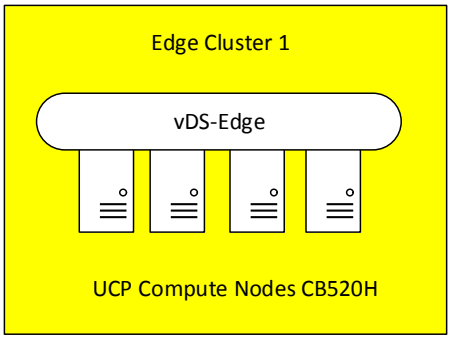
UCP compute cluster hosts (CP520H) are equipped with 2 × 10 Gb/sec NICs, that are connected to Brocade VDX6746 switches. Two Brocade VDX6746 switches are included in each CB 500 chassis, and 4 x10Gb/s interfaces of each VDX6746 are used for uplinks to VDX6740 switches.



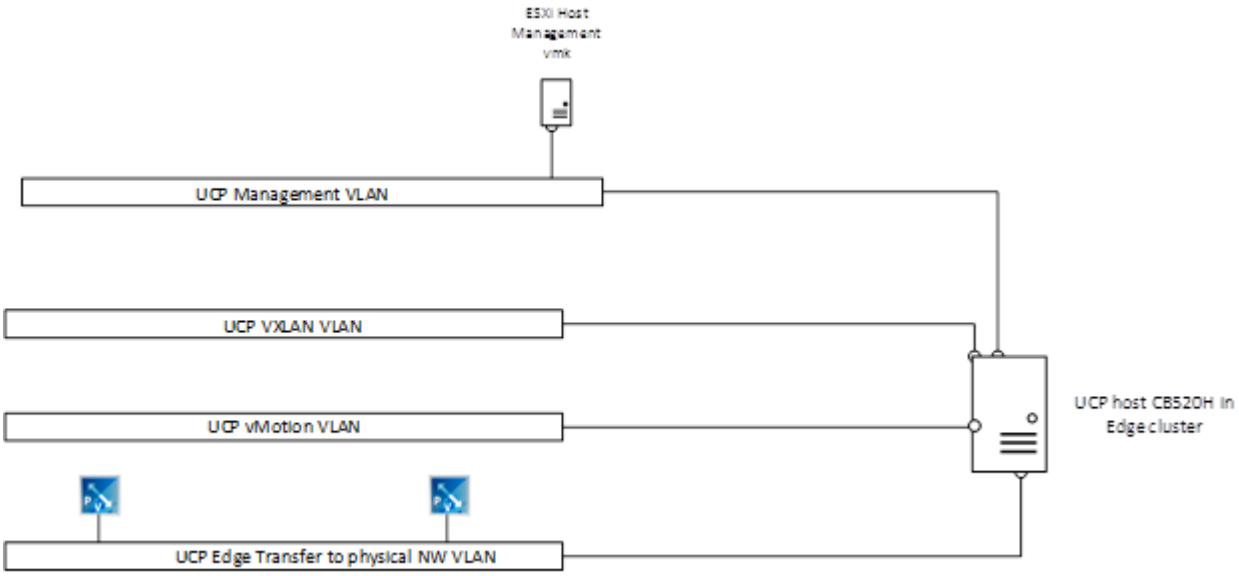
Both vmnics serve as uplinks on the respective virtual distributed switch (vDS). Each UCP compute host is connected to the vDS.

On Brocade VDX switches and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the management VLAN is declared a native VLAN, and therefore traffic is transmitted untagged for this VLAN.

All UCP compute nodes are connected to the same vDS and to Brocade VDX6746 switches, respectively. The number of compute hosts can scale up to 64. Here just four hosts are depicted for illustration purposes.



These are the typical VLANs that are configured on UCP 4000 with Brocade edge cluster nodes:



UCP management VLAN provides connection for the following:

- All ESXi host management vmkernel ports in all clusters

UCP with vMotion VLAN

- All ESXi host vMotion vmkernel ports in all clusters

UCP VXLAN VLAN

- All ESXi host VXLAN vmkernel ports
  - Due to the load balancing mode (route based on originating virtual port) two IP addresses are needed per ESXi host.

UCP edge transfer to physical network VLANs

- These VLANs provide connectivity between physical network and logical environments based on VMware NSX.

vSphere Distributed Switch in the Edge Cluster

This section describes settings on the vDS in the edge cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

NSX makes use of VXLAN by encapsulating ordinary IP packets in packets with an “outer” header. That is why the size of the original packet increases. VXLAN packets are also IP packets, but the “Don't fragment” bit is set. This is why the MTU has to be increased between ESXi hosts that participate in VXLAN. The physical network has to support the increased MTU and must be configured accordingly.

On the vDS, the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Brocade VDX switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Brocade VDX, LLDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

Each host has two NICs that can be teamed in different modes that serve different purposes.

To ensure that all NIC resources can be used in an active/active manner, teaming mode is set to “Route based on originating virtual port”.

This way, a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Brocade VDX switches.

UCP Compute Host NIC Driver Settings

Emulex NICs are installed on UCP compute hosts. For optimal performance, VXLAN offloading should be enabled.

To verify VXLAN Offload status:

- Connect to an ESXi host via SSH
- Enter the following command:

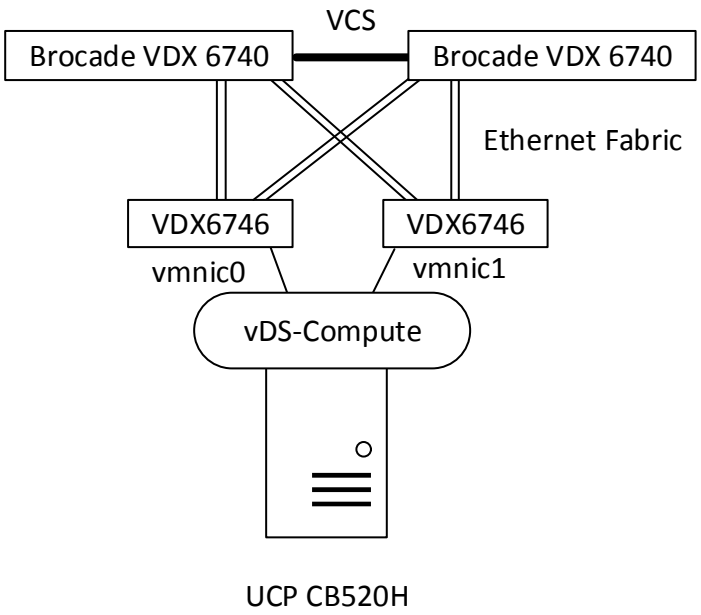
  ```
  # esxcli network nic list
  ```

- Determine the vmnic# of the NIC that is going to be verified

  ```
  # vsish -e get /net/pNics/vmnic0/stats | grep vxlan
  vxlan_offload: true
  vxlanUdpPort: 8472
  ```

vSphere Networking in UCP Compute Cluster Hosts

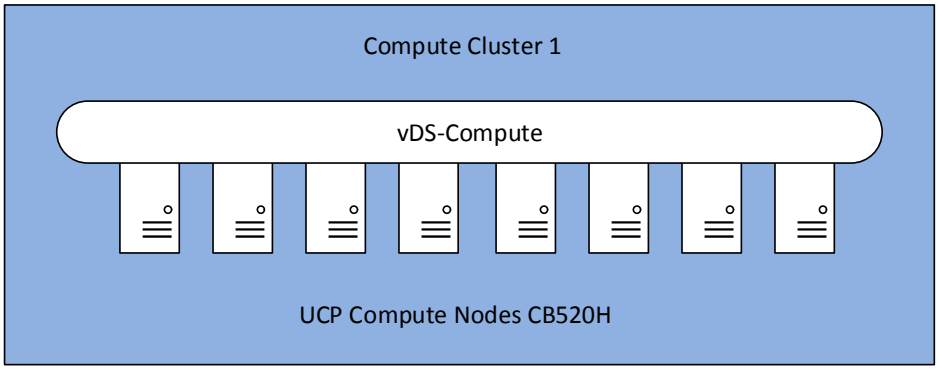
UCP compute cluster hosts (CB520H) are equipped with 2 × 10 Gb/sec NICs, that are connected to Brocade VDX6746 switches. Two Brocade VDX6746 switches are included in each CB 500 chassis, and 4 x10Gb/sec interfaces of each VDX6746 are used for uplinks to the VDX6740 switch.



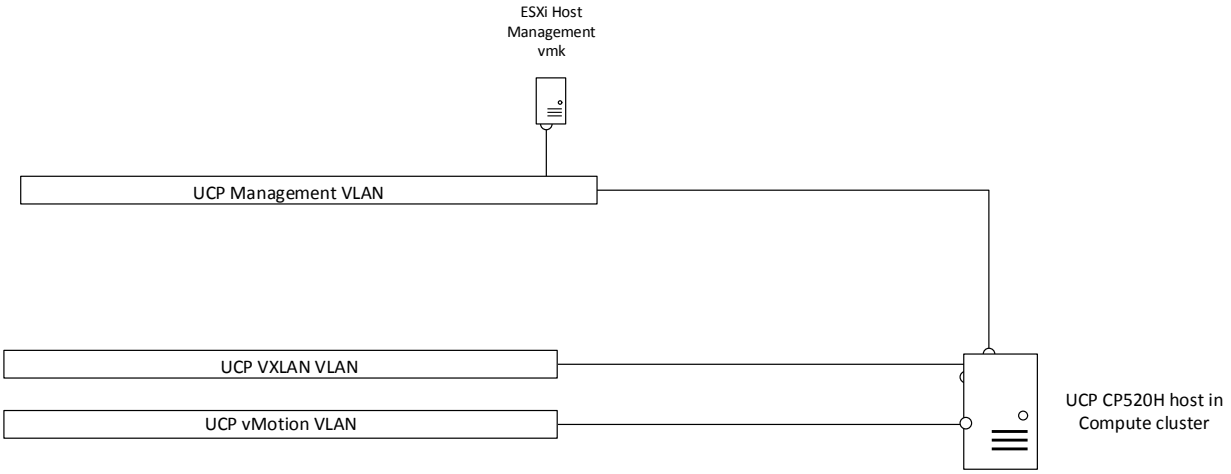
Both vmnics serve as uplinks on the respective virtual distributed switch (vDS). Each UCP compute host is connected to the vDS.

On Brocade VDX switches and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the management VLAN is declared a native VLAN, and therefore traffic is transmitted untagged for this VLAN.

All UCP compute nodes are connected to the same vDS and to Brocade VDX6746 switches, respectively. The number of compute hosts can scale up to 64. Here just eight hosts are depicted for illustration purposes.



These are the typical VLANs that are configured on UCP 4000 with Brocade compute nodes:



UCP management VLAN provides connection for the following:

- All ESXi host management vmkernel ports in all clusters (management and edge cluster, as well as compute clusters)

UCP with vMotion VLAN

- All ESXi host vMotion vmkernel ports in all clusters

UCP VXLAN VLAN

- All ESXi host VXLAN vmkernel ports
  - Due to the load balancing mode (route based on originating virtual port) two IP addresses are needed per ESXi host.

vSphere Distributed Switch in Compute Cluster

This section describes settings on the vDS in compute cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

NSX makes use of VXLAN by encapsulating ordinary IP packets in packets with an “outer” header. That is why the size of the original packet increases. VXLAN packets are also IP packets, but the “Don't fragment” bit is set. This is why the MTU has to be increased between ESXi hosts that participate in VXLAN. The physical network has to support the increased MTU and must be configured accordingly.

On the vDS, the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Cisco Nexus switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Brocade VDX, LLDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

Each host has two NICs that can be teamed in different modes that serve different purposes.

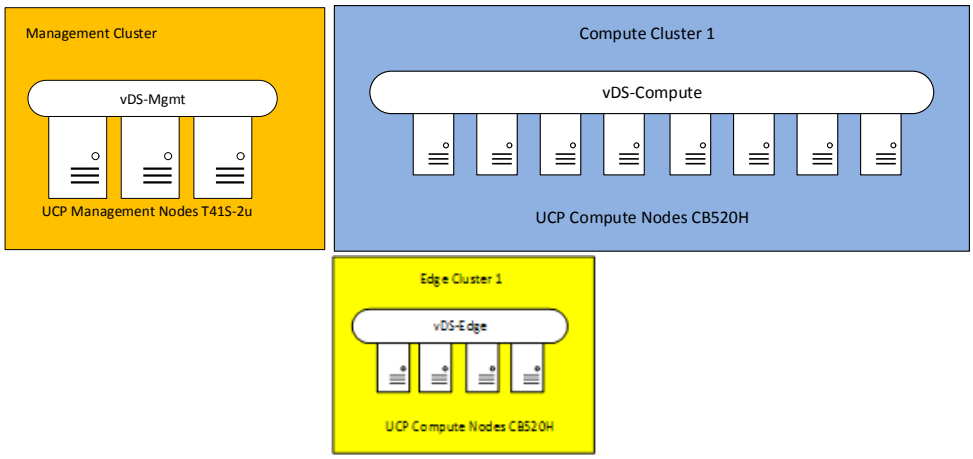
To ensure that all NIC resources can be used in an active/active manner, teaming mode is set to “Route based on originating virtual port”.

This way a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Brocade VDX switches.

vSphere Distributed Switches Summary

There are 3 different vDS that span different clusters:



Only compute clusters and edge clusters have VXLAN VLAN interfaces.

Scaling in vSphere 6.0:

- Maximum number of hosts per distributed switch: 1000
- Maximum number of hosts per cluster: 64

UCP 4000 with Brocade and NSX Logical Networking

The logical designs of NSX in UCP 4000E and UCP 4000 with Brocade are very similar.

This section describes the differences.

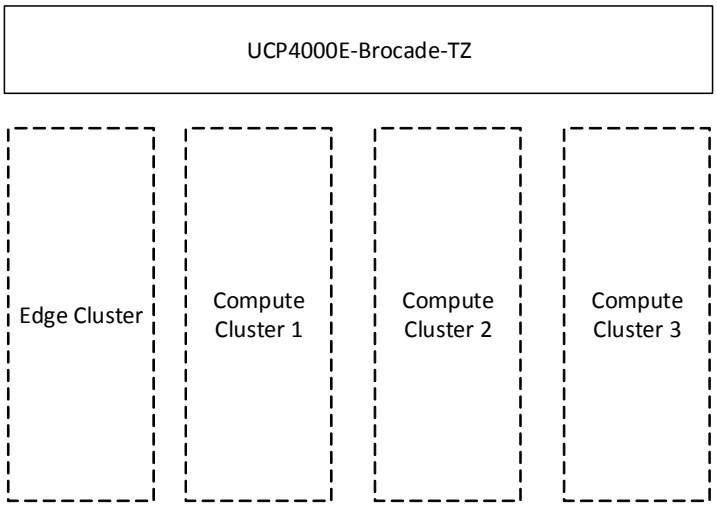
NSX Host Preparation

The management cluster in UCP 4000 with Brocade is not prepared for NSX, which means VIBs are not installed.

All other clusters (compute and edge) are prepared for NSX and have the necessary VIBs.

VXLAN Transport Zone

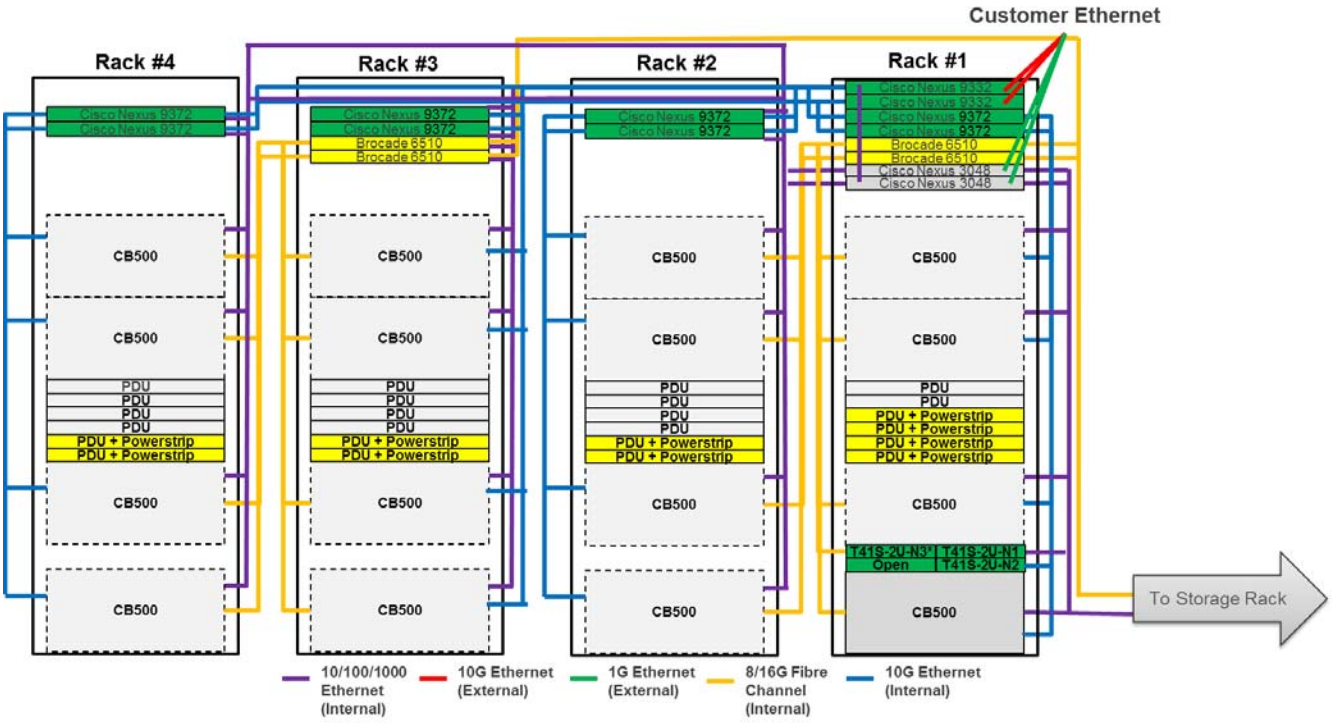
The transport zone is also be a little different:



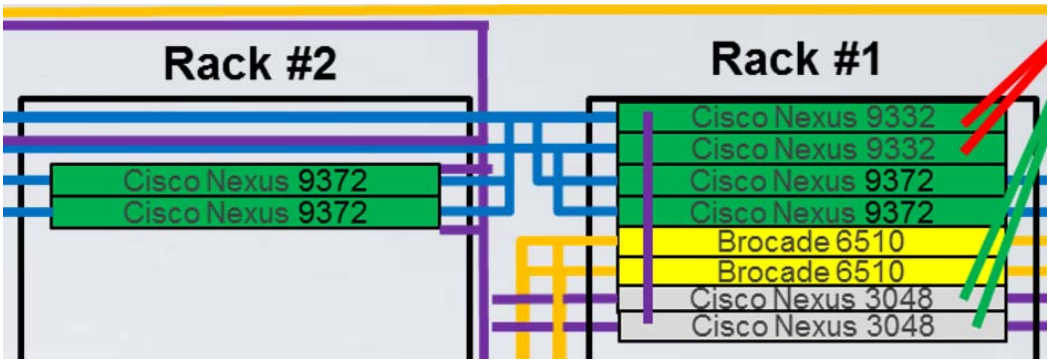
The management cluster is not part of the transport zone.

UCP 4000 with Cisco Architecture

UCP 4000 with Cisco scales up to four racks and 128 compute nodes as shown.



This is a closer view of the switching infrastructure.



Rack number 1 contains three UCP management nodes (rack optimized server for solutions, 2U four node) equipped with 192 GB RAM and a dual-port Intel 82599 10 GigE Open Compute Project NIC), two Cisco Nexus 9372 switches (Leaf) for access switching, and two Cisco Nexus 9332 switches (Spine) for tying all rack switches together and providing uplinks to customer networks.

One rack can hold up to four CB 500 blade chassis and up to 8 eight CB520H server blades with 2 × 10 Gb/sec NICs can be inserted into a CB 500 chassis.

This way, the first rack can scale up to 32 server blades. Each additional rack can bring up to another 32 compute hosts. In total the solution scales to four racks, which means 128 CB520H compute blades.

There is the option to operate the solution in Layer 2 or Layer 3 mode.

UCP 4000 with Cisco (Layer 2 Mode) and VMware NSX

This section describes vSphere clusters from an NSX perspective.

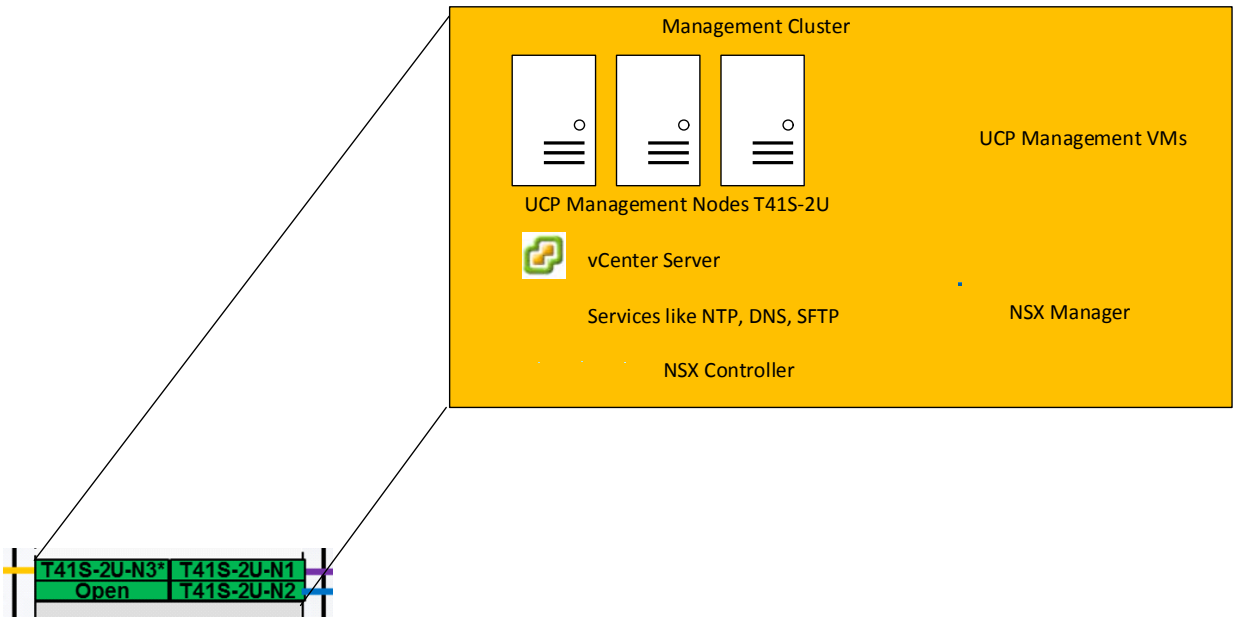
vSphere Clusters from an NSX Perspective

Several clusters in NSX have different purposes:

- Management Cluster
- NSX Edge Cluster
- NSX Compute Cluster

Cisco management and edge clusters are separate for UCP 4000.

UCP Management Cluster



The figure above shows three management nodes forming a vSphere cluster. All management related VMs are part of this cluster, for example:

- UCP management VMs
- vCenter Server
- Services VMs (NTP, DNS, etc.)
- NSX Manager
- NSX Controller

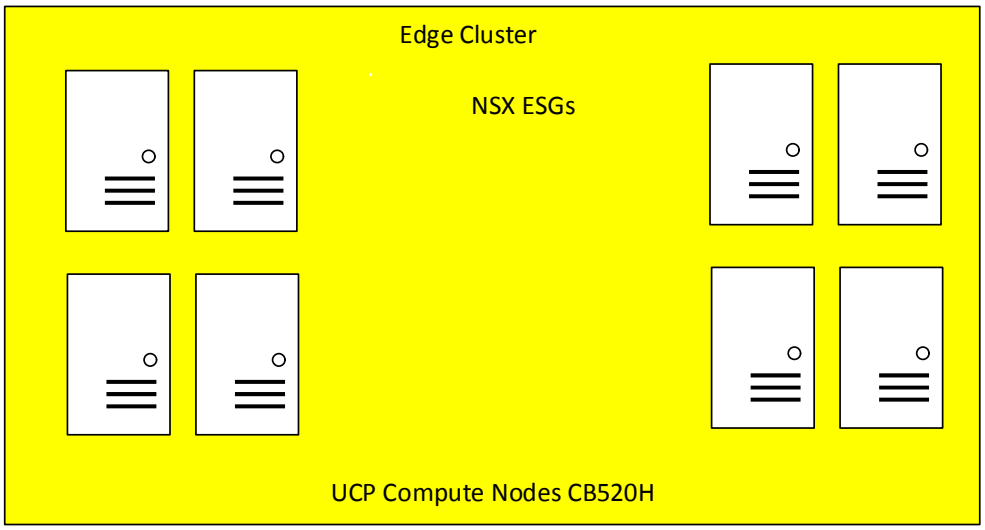
UCP Edge and Compute Cluster

CB520H server blades are hosted in CB 500 blade chassis. Up to eight server fit into a chassis. The UCP 4000 with Cisco solution scales up to 16 CB 500 blade chassis with a maximum of 128 blade server systems.

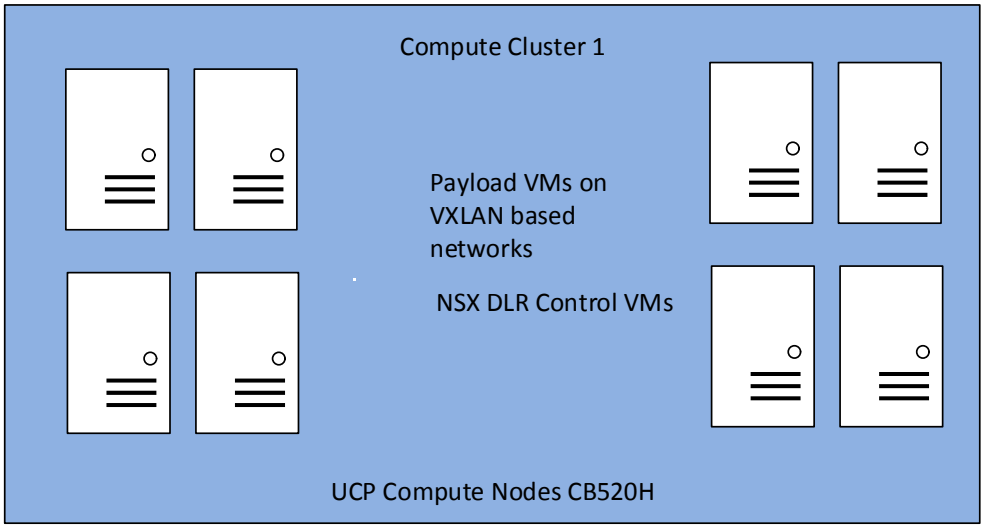
In order to provide sufficient bandwidth between logical networks and the physical world, and to care for an entire rack failure, the NSX Edge cluster can be distributed over different racks and different CB 500 chassis.

Depending on the required throughput between physical and logical networks, ESXi hosts with NSX ESGs can be added to the cluster.

This example shows an Edge cluster with 8 ESXi hosts.



Compute clusters can also be mounted in more than one rack in order to compensate for a rack failure.



vSphere Networking

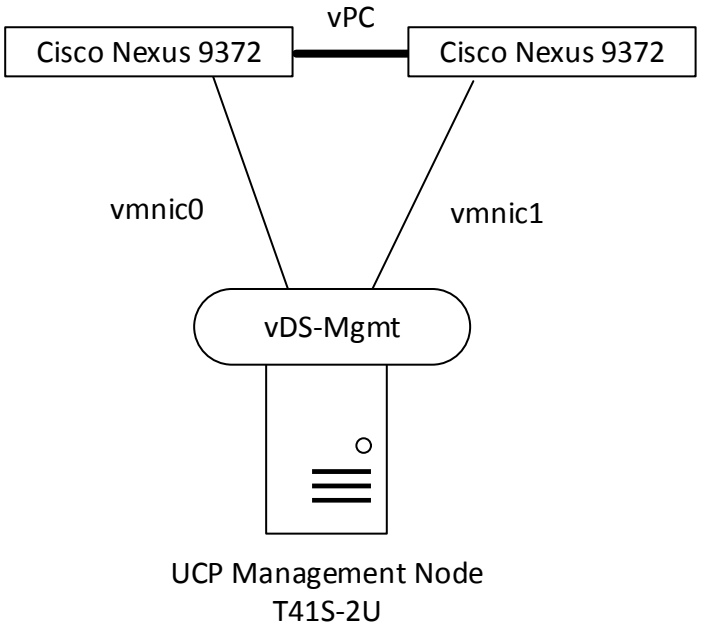
This section describes ESXi host NICs, vSphere virtual switches, and respective VLANs.

UCP Management Hosts NIC Driver Settings

Because there is no VXLAN in the UCP management cluster, no special settings need to be considered.

vSphere Networking in UCP Management Hosts

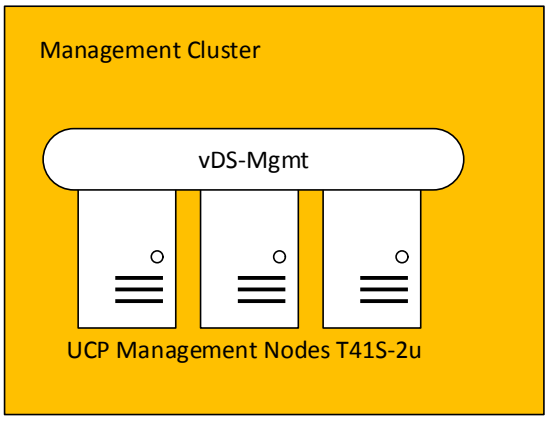
UCP management nodes (rack optimized server for solutions, 2U four node) are equipped with 2 × 10 Gb/sec NICs, that are connected to Nexus 9372 top of rack switches.



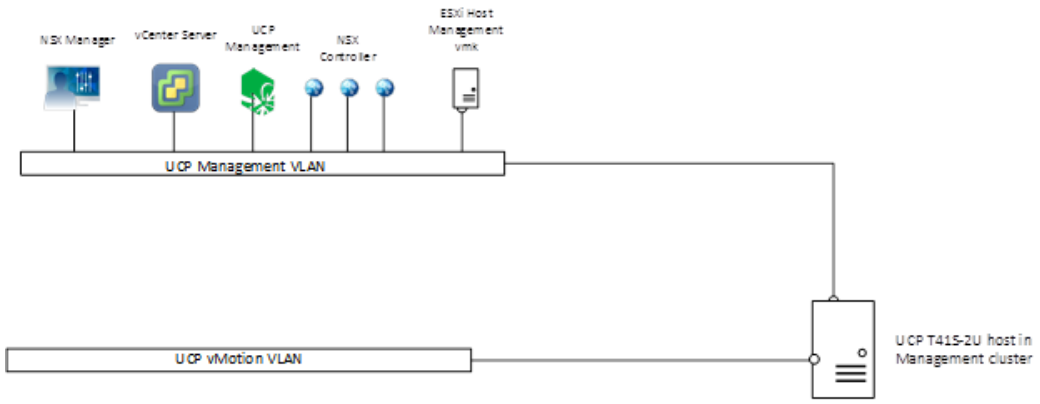
Both vmnics serve as uplinks for the respective virtual distributed switch (vDS). Each UCP management node is connected to the vDS.

On Nexus 9372 and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the UCP management VLAN is declared a native VLAN, and therefore traffic is transmitted untagged for this VLAN.

All three UCP management nodes are connected to the same vDS and to Cisco Nexus 9372 switches, respectively.



These are the typical VLANs that are configured on UCP 4000 with Cisco management nodes:



The UCP management VLAN provides connection for the following:

- UCP management VMs
- All ESXi host management vmkernel ports in all clusters (Management cluster and Edge cluster, as well as compute clusters)
- vCenter Server
- Three NSX Controllers
- NSX Manager

UCP with vMotion VLAN

- All ESXi host vMotion vmkernel ports in all clusters

vSphere Distributed Switch in Management Cluster

This section describes settings on the vDS in Management cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

There is no VXLAN in the management cluster. However, since the MTU is set to 9000 for all other vDS (Edge and Compute), it is set to 9000 on the vDS management cluster as well.

On the vDS, the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Cisco Nexus switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Cisco Nexus, CDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

Each host has two NICs that can be teamed in different modes that serve different purposes.

To ensure that all NIC resources can be used in an active/active manner, the teaming mode is set to “Route based on originating virtual port”. This way a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Cisco Nexus switches.

UCP Edge Hosts NIC Driver Settings

UCP edge hosts also have Emulex NICs installed. For optimal performance, VXLAN offloading should be enabled.

For VXLAN Offload status verification:

- Connect to an ESXi host via SSH
- Enter the following command:

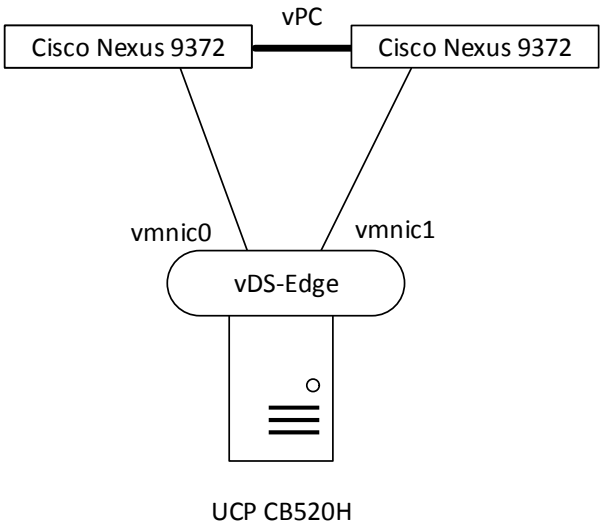
```
# esxcli network nic list
```

- Determine the vmnic# of the NIC that is going to be verified

```
# vsish -e get /net/pNics/vmnic0/stats | grep vxlan
vxlan_offload: true
vxlanUdpPort: 8472
```

vSphere Networking in UCP Edge Cluster Hosts

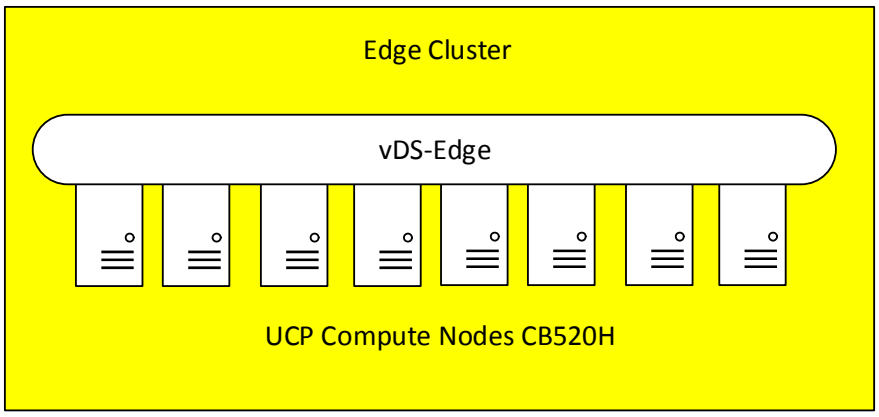
UCP edge cluster hosts (CB520H) are equipped with 2 × 10 Gb/sec NICs, that get connected to Cisco Nexus 9372 switches. Pass-through modules are used in CB 500 for this purpose.



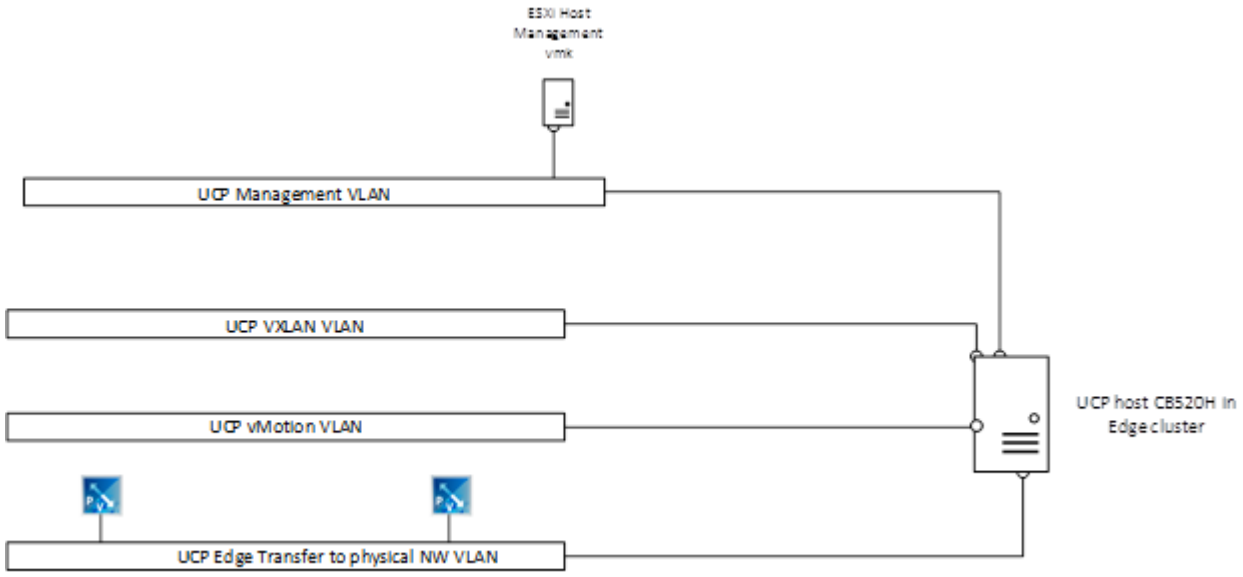
Both vmnics serve as uplinks on the respective virtual distributed switch (vDS). Each UCP compute host is connected to the vDS.

On Cisco Nexus and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the management VLAN is declared a native VLAN, and therefore traffic is transmitted untagged for this VLAN.

All UCP compute nodes are connected to the same vDS and to Cisco Nexus 9372 switches, respectively. The number of compute hosts can scale up to 128. Here eight hosts are depicted for illustration purposes.



These are the typical VLANs that are configured on UCP 4000 with Cisco edge cluster nodes:



UCP management VLAN provides connection for the following:

- All ESXi host management vmkernel ports in all clusters

UCP with vMotion VLAN

- All ESXi host vMotion vmkernel ports in all clusters

UCP VXLAN VLAN

- All ESXi host VXLAN vmkernel ports
- Due to the load balancing mode (route based on originating virtual port), two IP addresses are needed per ESXi host.

UCP edge transfer to physical network VLANs

- These VLANs provide connectivity between physical network and logical environments based on VMware NSX.

vSphere Distributed Switch in the Edge Cluster

This section describes settings on the vDS in the edge cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

NSX makes use of VXLAN by encapsulating ordinary IP packets in packets with an “outer” header. That is why the size of the original packet increases. VXLAN packets are also IP packets, but the “Don't fragment” bit is set. This is why the MTU has to be increased between ESXi hosts that participate in VXLAN. The physical network has to support the increased MTU and must be configured accordingly.

On the vDS, the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Cisco Nexus switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Cisco Nexus, CDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

Each host has two NICs that can be teamed in different modes that serve different purposes.

To ensure that all NIC resources can be used in an active/active manner, the teaming mode is set to “route based on originating virtual port”. This way a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Cisco Nexus switches.

UCP Compute Hosts NIC Driver Settings

UCP compute hosts also have Emulex NICs installed. For optimal performance, VXLAN offloading should be enabled.

For VXLAN Offload status verification:

- Connect to an ESXi host via SSH
- Enter the following command:

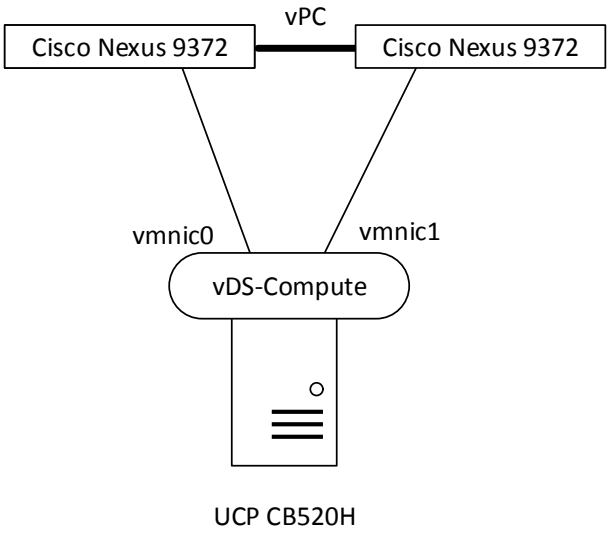
```
# esxcli network nic list
```

- Determine the vmnic# of the NIC that is going to be verified

```
# vsish -e get /net/pNics/vmnic0/stats | grep vxlan
vxlan_offload: true
vxlanUdpPort: 8472
```

vSphere Networking in UCP Compute Cluster Hosts

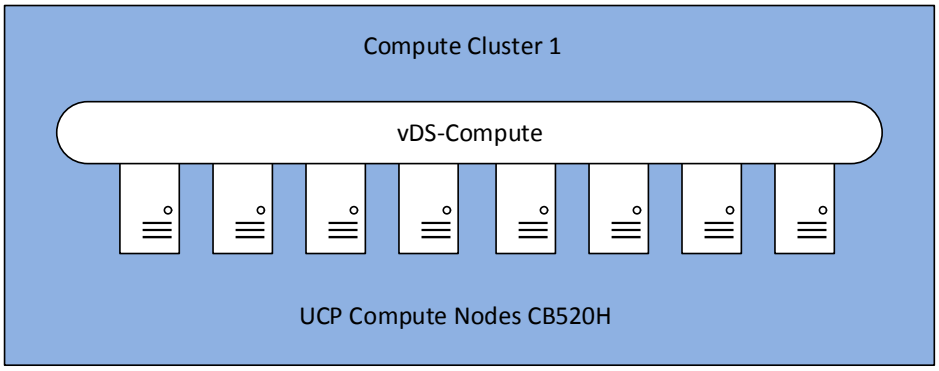
UCP compute cluster hosts (CB520H) are equipped with 2 × 10 Gb/sec NICs, that are connected to Cisco Nexus 9372 switches. Pass-through modules are used in CB 500 for this purpose.



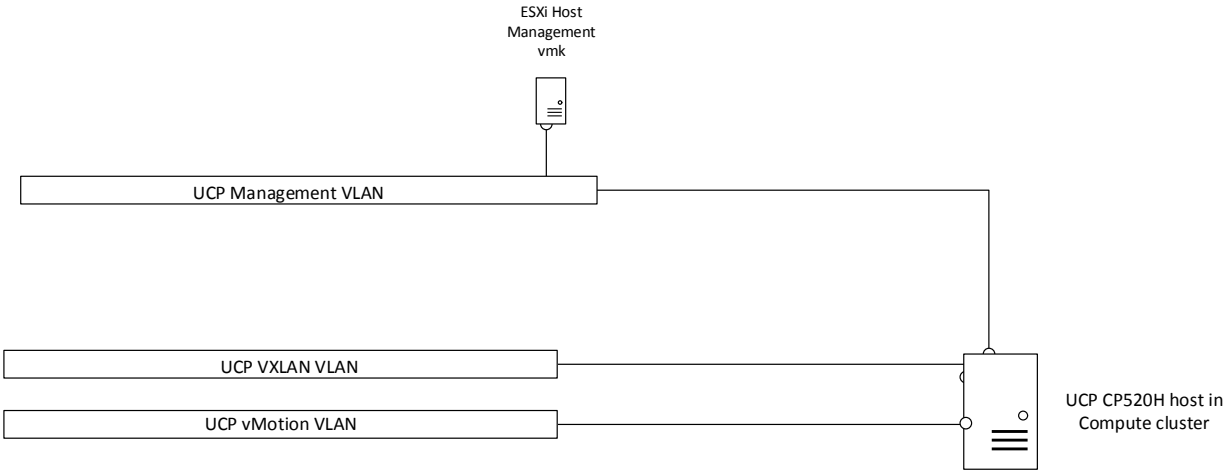
Both vmnics serve as uplinks on the respective virtual distributed switch (vDS). Each UCP compute host is connected to the vDS.

On Cisco Nexus and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the management VLAN is declared a native VLAN, and therefore traffic is transmitted untagged for this VLAN.

All UCP compute nodes are connected to the same vDS and to Cisco Nexus switches, respectively. The number of compute hosts can scale up to 128. Here just eight hosts are depicted for illustration purposes.



These are the typical VLANs that are configured on UCP 4000 with Cisco compute nodes:



UCP management VLAN provides connection for the following:

- All ESXi host management vmkernel ports in all clusters (management and edge cluster as well as compute clusters)

UCP with vMotion VLAN

- All ESXi host vMotion vmkernel ports in all clusters

UCP VXLAN VLAN

- All ESXi host VXLAN vmkernel ports

Due to the load balancing mode (Route based on originating virtual port) two IP addresses are needed per ESXi host.

vSphere Distributed Switch in Compute Cluster

This section describes settings on the vDS in compute cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

NSX makes use of VXLAN by encapsulating ordinary IP packets in packets with an “outer” header. That is why the size of the original packet increases. VXLAN packets are also IP packets, but the “Don't fragment” bit is set. This is why the MTU has to be increased between ESXi hosts that participate in VXLAN. The physical network has to support the increased MTU and must be configured accordingly.

On the vDS, the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Cisco Nexus switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Cisco Nexus, CDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

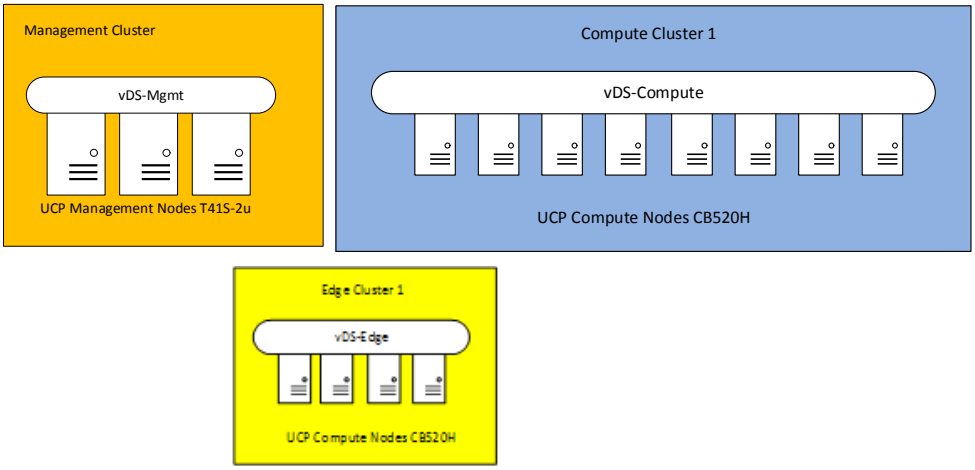
Each host has two NICs that can be teamed in different modes that serve different purposes.

To ensure that all NIC resources can be used in an active/active manner, the teaming mode is set to “route based on originating virtual port”. This way a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Cisco Nexus switches.

vSphere Distributed Switches Summary

There are 3 different vDS that span different clusters:



Only compute clusters and edge clusters will have VXLAN VLAN interface.

Scaling in vSphere 6.0:

- Maximum number of hosts per distributed switch: 1000
- Maximum number of hosts per cluster: 64

That means all ESXi CB 520H compute host can be attached to vDS-Compute, but at least two clusters have to be created.

UCP 4000 with Cisco and NSX Logical Networking

The logical designs of NSX in UCP 4000E and UCP 4000 with Cisco are very similar.

This section illustrates the differences.

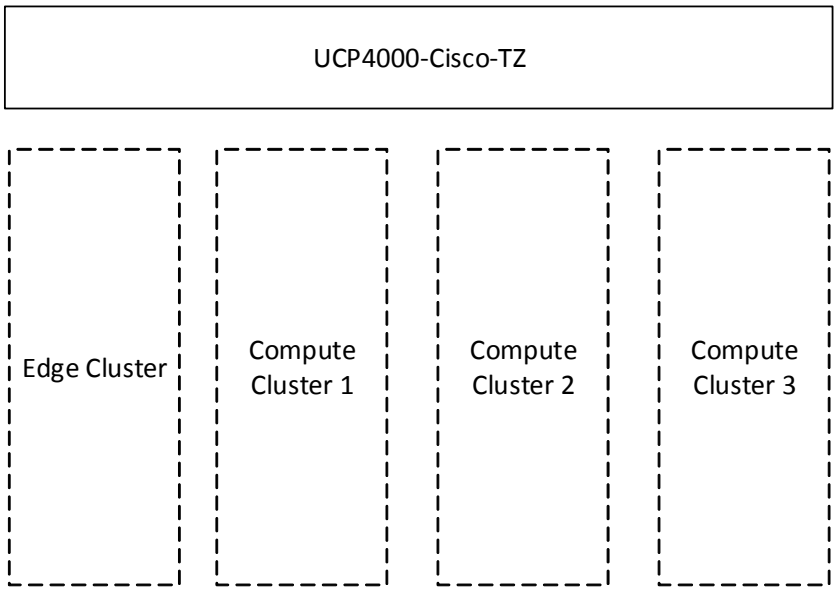
NSX Host Preparation

The management cluster in UCP 4000 with Cisco is not prepared for NSX, which means VIBs are not installed.

All other clusters (compute and edge) are prepared for NSX and get the necessary VIBs.

VXLAN Transport Zone

The transport zone looks like this:



The management cluster is not part of the transport zone.

IP Address Planning Layer 2 Mode

This section provides an overview of the least number of IP addresses that are required for each UCP solution.

Table 2. IP Address Requirements

	UCP 4000E (24 x CB520H hosts)	UCP 4000 with Brocade (64 x CB520H hosts)	UCP 4000 with Cisco (L2) (128 x CB520H hosts)
Hitachi Unified Compute Platform Management VLAN	30	70	135
vMotion	30	70	135
VXLAN	60	140	270

UCP 4000E and NSX Distributed Firewall

In UCP 4000E, the management and edge clusters are part of NSX. That means NSX Distributed Firewall can be activated in these clusters as well.

However there are a lot of VMs for management purposes. Care needs to be taken before activating NSX dFW on this cluster, otherwise the environment can become unmanageable.

It is therefore recommended to include the following VMs in the NSX Exclusion list:

- vCenter
- UCP management VMs

NSX Manager and Controller and Edges/DLRs are automatically on the exclusion list.

UCP 4000 with Cisco in Layer 3 Mode and VMware NSX

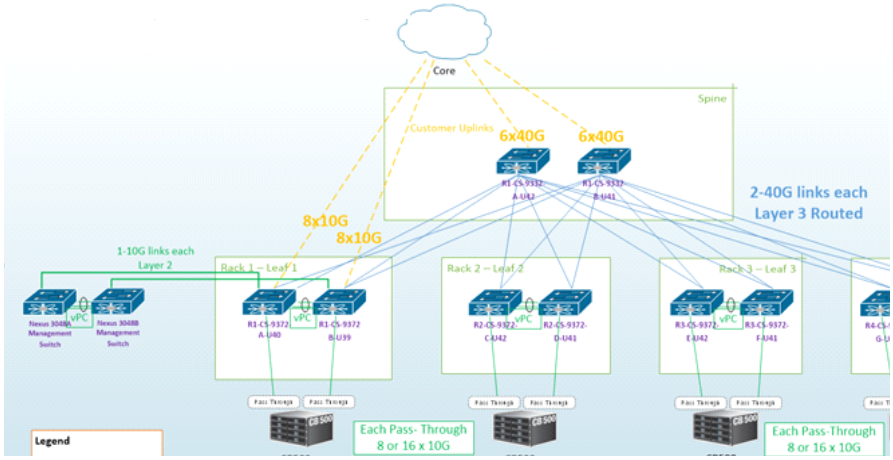
This chapter describes UCP in a fully routed environment with VMware NSX.

UCP 4000 with Cisco Layer 3 is characterized by the location of the default gateway for any device in a rack. The default gateway is a feature of the top-of-rack switches (Cisco Nexus 9372).

As a consequence, IP ranges are local to a rack and no Layer 2 connection is available between racks.

This is valid for all subnets necessary in each rack:

- ESXi management VLAN
- vMotion VLAN
- VXLAN VLAN
- NSX Edge VLANs to the physical world



The links between Cisco Nexus 9372 and Cisco Nexus 9332 switches are routed only. That means a dynamic routing protocol (for example, OSPF) is being used to propagate the networks of each rack to the entire environment.

Another consequence of Layer 3 mode is that VLAN IDs are only local to a rack and can not be reused in other racks to carry IP packets of different subnets. We will make use of this feature as explained further down.

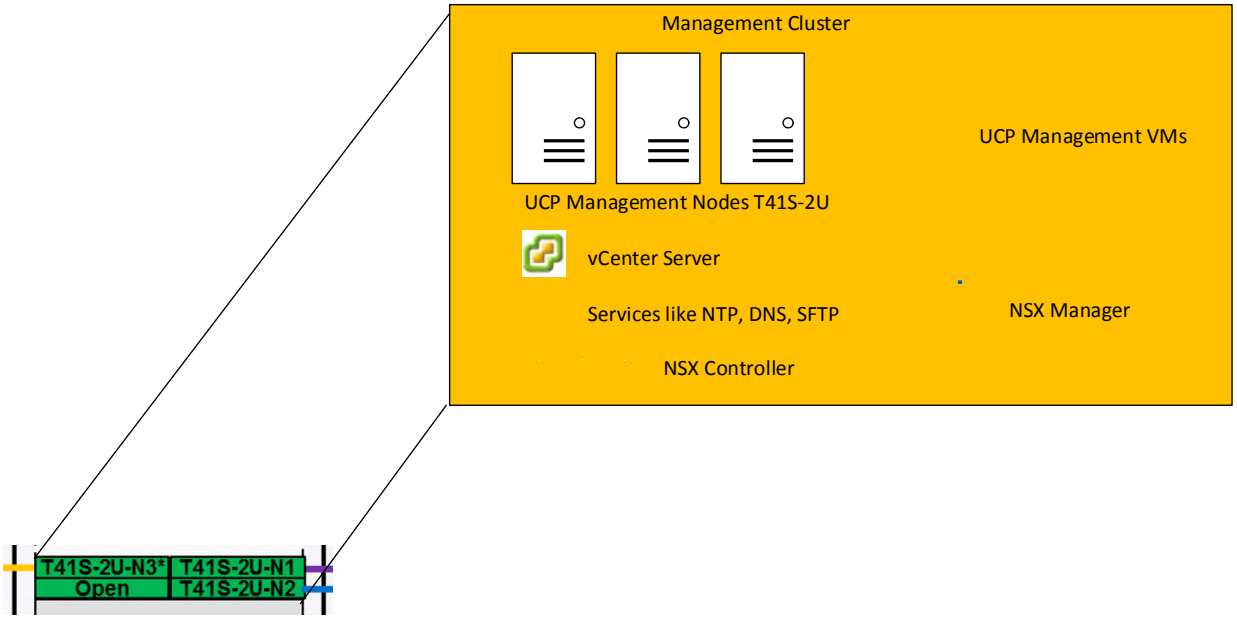
vSphere Cluster from an NSX Perspective

Several clusters exist in NSX that serve different purposes:

- Management Cluster
- NSX Edge Cluster
- NSX Compute Cluster

Cisco management and edge clusters are separate for UCP 4000.

UCP Management Cluster



The figure above shows three management nodes that form a vSphere cluster. All management related VMs are part of this cluster, for example:

- UCP management VMs
- vCenter Server
- Services VMs (NTP, DNS, etc.)
- NSX Manager
- NSX Controller

Clusters in the CB 500 Blade Chassis

This section describes the settings of clusters set up with CB520H hosts.

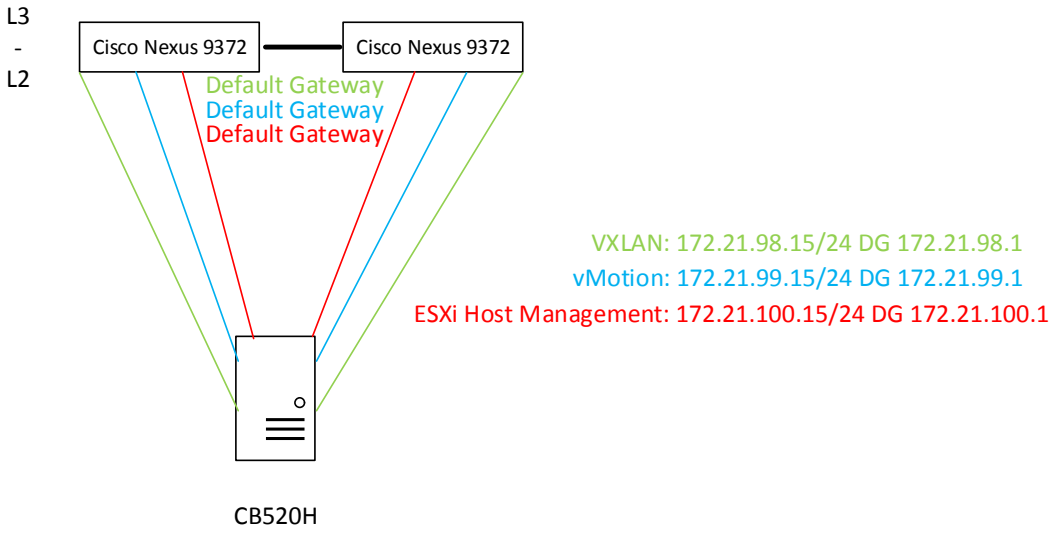
UCP ESXi TCP/IP Stacks in Layer 3

In order to be able to migrate a VM from a host in one rack to a host in a different rack, it is a prerequisite to use the dedicated vMotion TCP/IP Stack on a host. This way it is possible to assign a dedicated Default Gateway for vMotion traffic.

In total there are three independent TCP/IP stacks active on any given host:

- ESXi host management (default TCP/IP stack)
- vMotion (dedicated vMotion TCP/IP stack)
- VXLAN (dedicated VXLAN TCP IP stack)

This is how it looks on a host:



UCP Edge Cluster

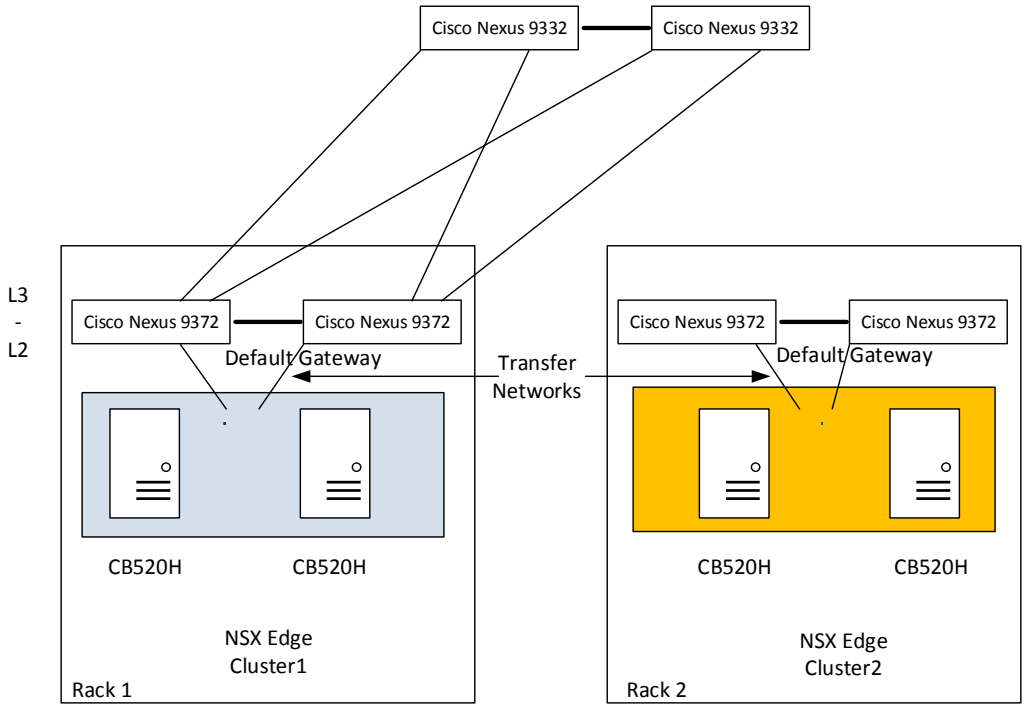
NSX Edges have at least one interface on a VLAN-based port group. These interfaces provide for communication between physical and logical networks. The other interfaces are based on VXLAN. In the layer 3 environment, the VLAN is only local to a rack and the IP range is not available in other racks. This is why NSX Edge VMs can only be migrated between hosts in the same rack, but not between racks (unless special technologies are used to tunnel a VLAN).

Depending on the required throughput between physical and logical networks, ESXi hosts with NSX ESGs can be added to the cluster.

If the failure of a rack has to be taken into account, because of customer requirements, more than one NSX Edge cluster can be configured.

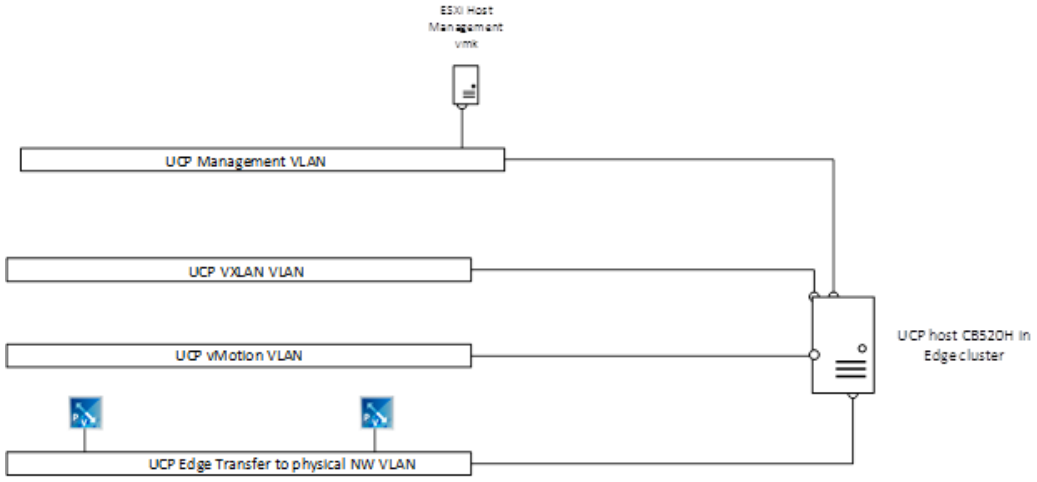
Note - If cloud automation is being used, it has to be confirmed that using several NSX Edge clusters is supported.

This example shows two NSX Edge clusters in different racks:



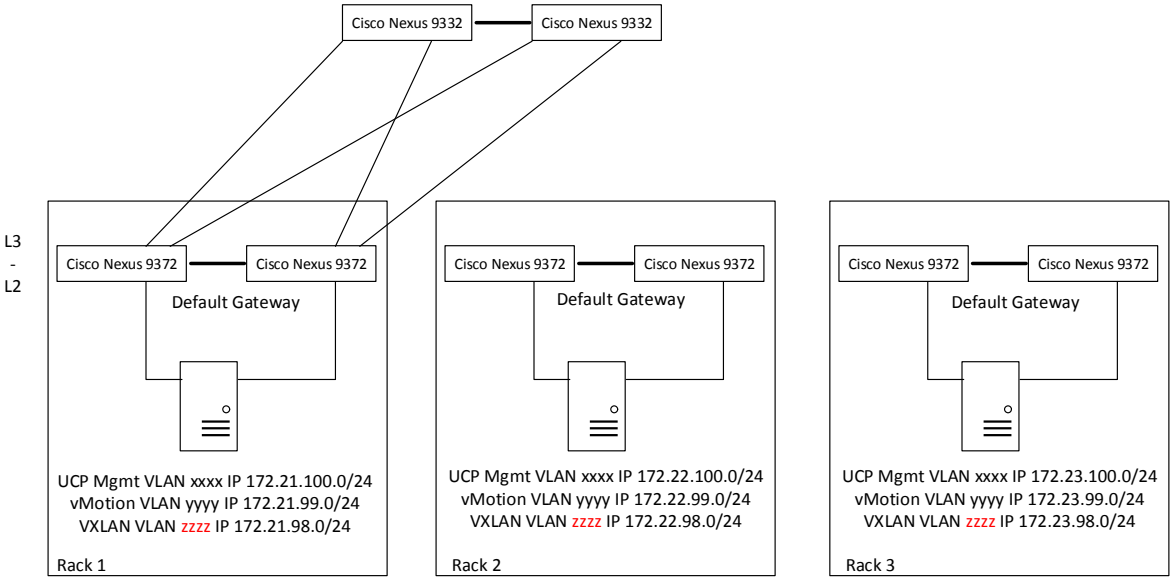
Although the VLAN ID for the transfer network between the NSX Edge and Cisco Nexus is the same in both racks, the IP address range is different. This is why an NSX Edge cannot be migrated from one rack to another.

These are the VLANs that need to be available at each host in an NSX Edge cluster.



Because VLANs are local to a rack, it is recommended that identical VLAN IDs are use for the same-purpose VLANs in different racks.

The following illustration shows how this looks using three racks. It is important to note that the UCP management VLAN and vMotion VLAN IDs can be identical across different racks, but do not have to be. The VXLAN VLAN ID must be identical in all racks, in case clusters shall be distributed across racks.



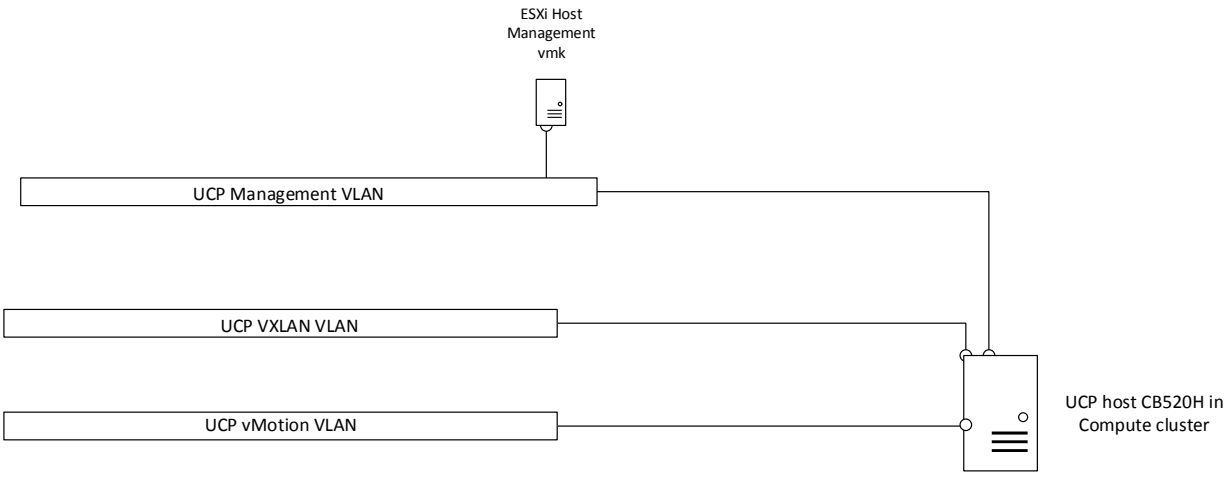
UCP Compute Cluster

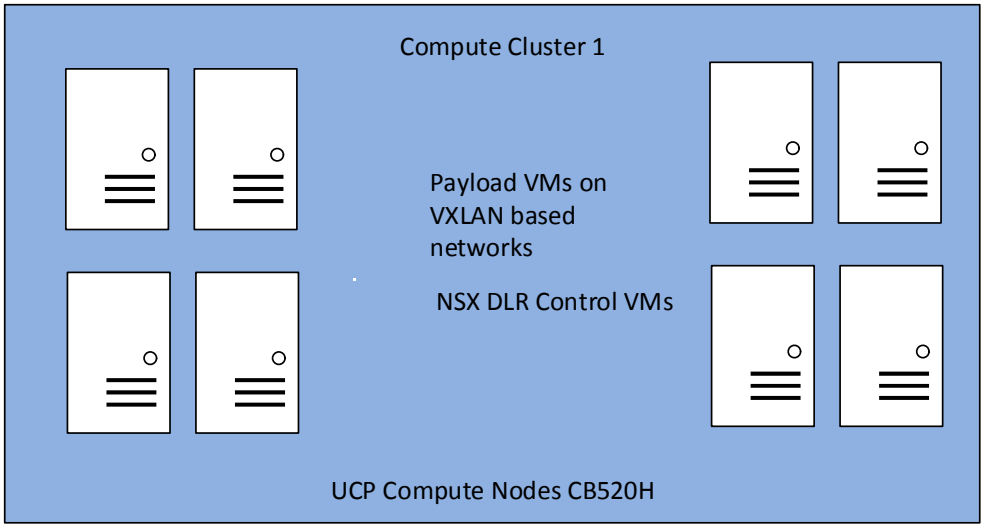
ESXi hosts in compute clusters need access to three VLANs:

- UCP management with all ESXi vmkernel interfaces
- UCP with vMotion
- UCP with VXLAN

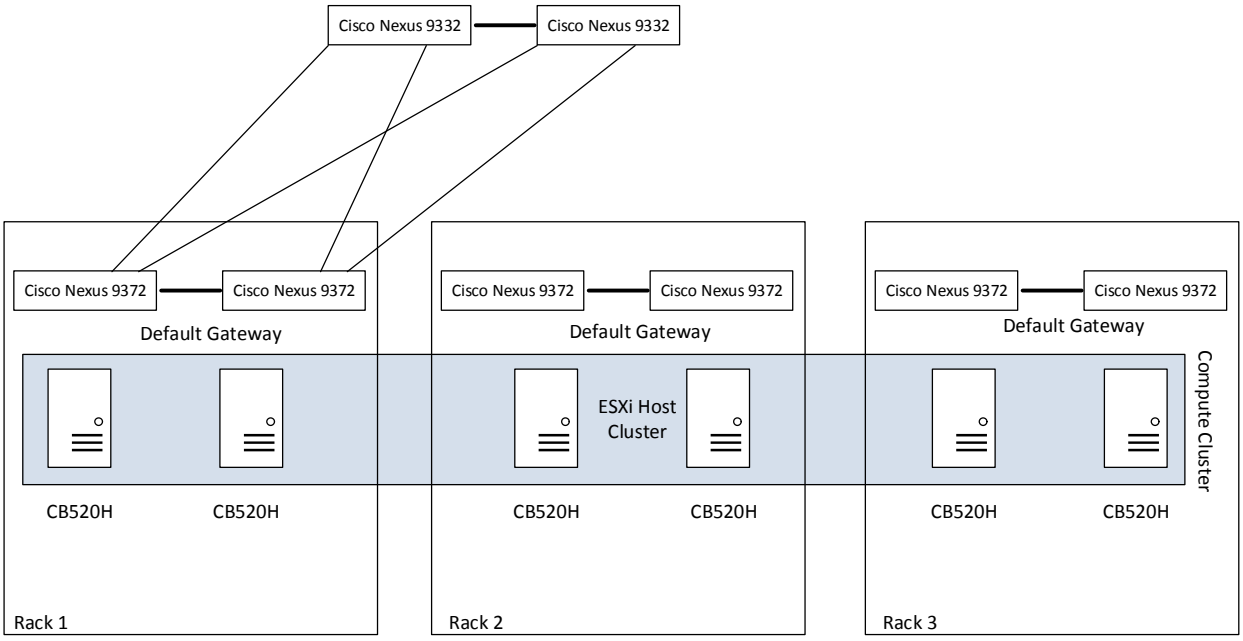
Again, these VLANs are local to a rack and their IP ranges are not available on other racks.

These are the VLANs that ESXi hosts in compute clusters need access to.





Compute clusters can also be mounted in more than one rack in order to compensate for a rack failure. This is illustrated below.



UCP Edge Hosts NIC Driver Settings

UCP edge hosts also have Emulex NICs installed in them. For optimal performance, VXLAN offloading should be enabled.

For VXLAN Offload status verification:

- Connect to an ESXi host via SSH
- Enter the following command:

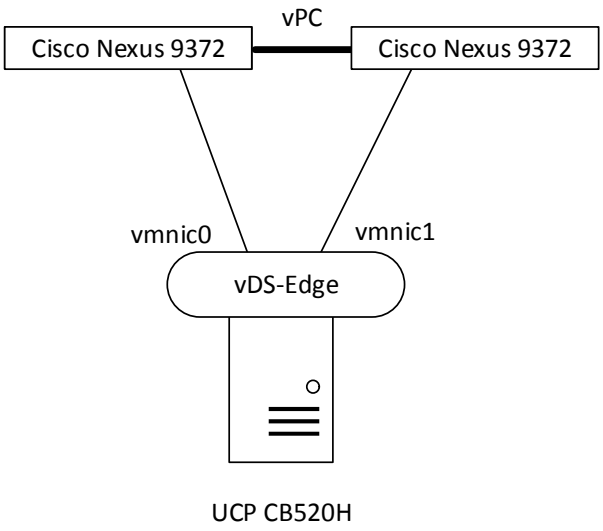
```
# esxcli network nic list
```

- Determine the vmnic# of the NIC that is going to be verified

```
# vsish -e get /net/pNics/vmnic0/stats | grep vxlan
vxlan_offload: true
vxlanUdpPort: 8472
```

vSphere Networking in UCP Edge Cluster Hosts

UCP edge cluster hosts (CB520H) are equipped with 2 × 10 Gb/sec NICs, that are connected to Cisco Nexus 9372 switches. Pass-through modules are used in CB 500 for this purpose.

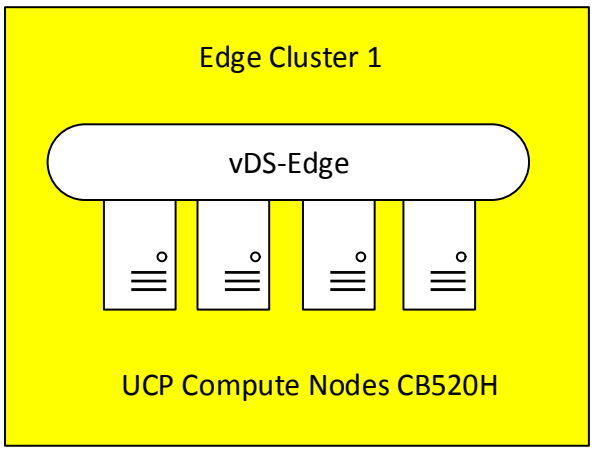


Both vmnics serve as uplinks on the respective virtual distributed switch (vDS). Each UCP compute host is connected to the vDS.

On Cisco Nexus and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the management VLAN is declared a native VLAN, and therefore traffic is transmitted untagged for this VLAN.

All ESXi hosts in an NSX Edge cluster reside in the same rack and are attached to the same logical vDS.

In the figure below, four hosts are depicted for illustration purposes.



vSphere Distributed Switch in an Edge Cluster

This section describes settings on the vDS in an edge cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

NSX makes use of VXLAN by encapsulating ordinary IP packets in packets with an “outer” header. That is why the size of the original packet increases. VXLAN packets are also IP packets, but the “Don't fragment” bit is set. This is why the MTU has to be increased between ESXi hosts that participate in VXLAN. The physical network has to support the increased MTU and must be configured accordingly.

On the vDS, the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Cisco Nexus switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Cisco Nexus, CDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

Each host has two NICs that can be teamed in different modes that serve different purposes.

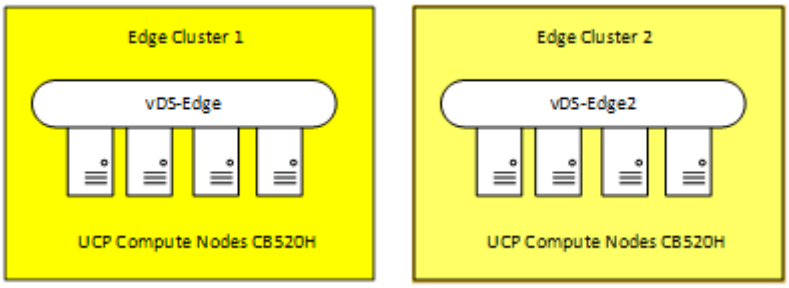
To ensure that all NIC resources can be used in an active/active manner, the teaming mode is set to “route based on originating virtual port”.

This way, a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Cisco Nexus switches.

vDS per NSX Edge Cluster

If several NSX Edge clusters are configured, use a dedicated vDS per cluster. In other words, a vDS for NSX edges should not span across racks.



UCP Compute Host NIC Driver Settings

UCP compute hosts also have Emulex NICs installed in them. For optimal performance, VXLAN offloading should be enabled.

For VXLAN Offload status verification:

- Connect to an ESXi host via SSH
- Enter the following command:

```
# esxcli network nic list
```

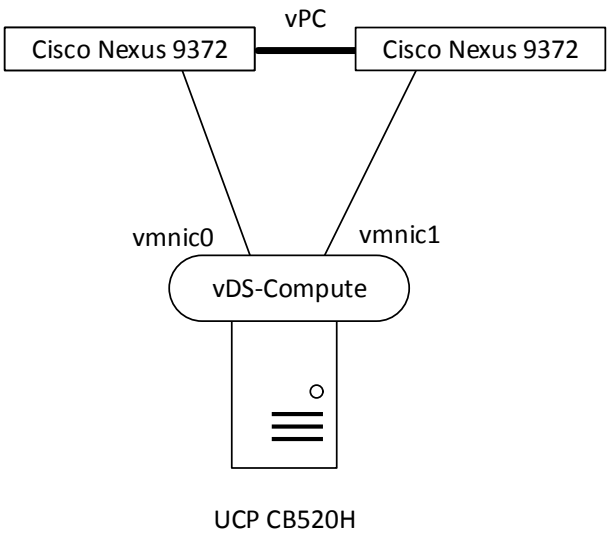
- Determine the vmnic# of the NIC that is going to be verified

```
# vsish -e get /net/pNics/vmnic0/stats | grep vxlan
vxlan_offload: true
vxlanUdpPort: 8472
```

vSphere Networking in UCP Compute Cluster Hosts

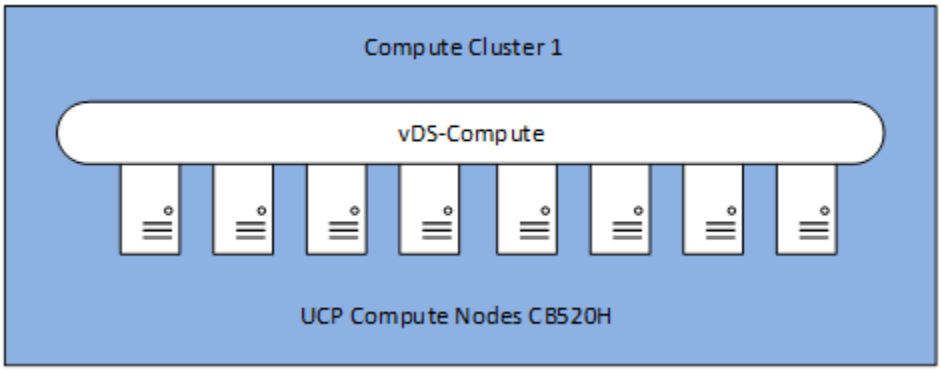
UCP compute cluster hosts (CB520H) are equipped with 2 × 10 Gb/sec NICs, that are connected to Cisco Nexus 9372 switches. Pass-through modules are used in CB 500 for this purpose.

Both vmnics serve as uplinks on the respective virtual distributed switch (vDS). Each UCP compute host is connected to the vDS.

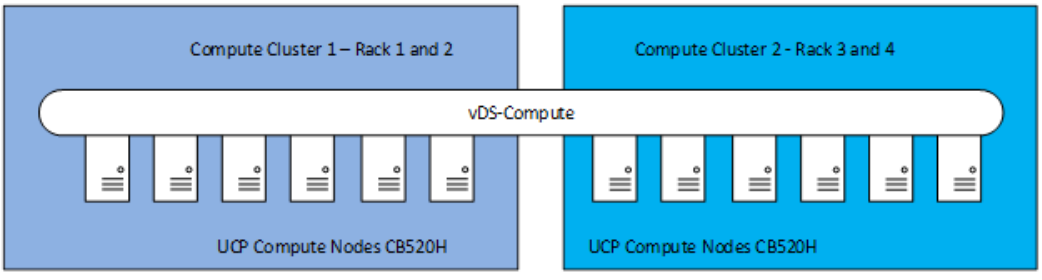


On Cisco Nexus and in vDS, uplink ports are configured for trunking, which means VLANs get tagged. However, the management VLAN is declared a native VLAN, and therefore traffic is transmitted untagged for this VLAN.

All UCP compute cluster hosts are connected to the same vDS and to Cisco Nexus, respectively. The number of compute hosts can scale up to 128. Here just eight hosts are depicted for illustration purposes.



A vDS for compute clusters can be distributed across several racks and clusters.



vSphere Distributed Switch in the Compute Cluster

This section describes settings on the vDS in the compute cluster.

Maximum Transmission Unit (MTU) and Discovery Protocol

NSX makes use of VXLAN by encapsulating ordinary IP packets into packets with an “outer” header. That is why the size of the original packet increases. VXLAN packets are also IP packets, but the “Don't fragment” bit is set. This is why the MTU has to be increased between ESXi hosts that participate in VXLAN. The physical network has to support the increased MTU and must be configured accordingly.

On the vDS, the MTU can be set globally. It is set to 9000 bytes, which is the maximum.

On Cisco Nexus switches, the MTU is set to 9216 bytes.

VMware vDS supports Link Layer Discovery Protocol and Cisco Discovery Protocol. Because physical switches are Cisco Nexus, CDP is enabled on vDS and set to “both” (Advertise and Listen). The protocol provides information on which switch and on which port an ESXi host is connected to.

Port Groups Teaming Policy for Uplink Ports

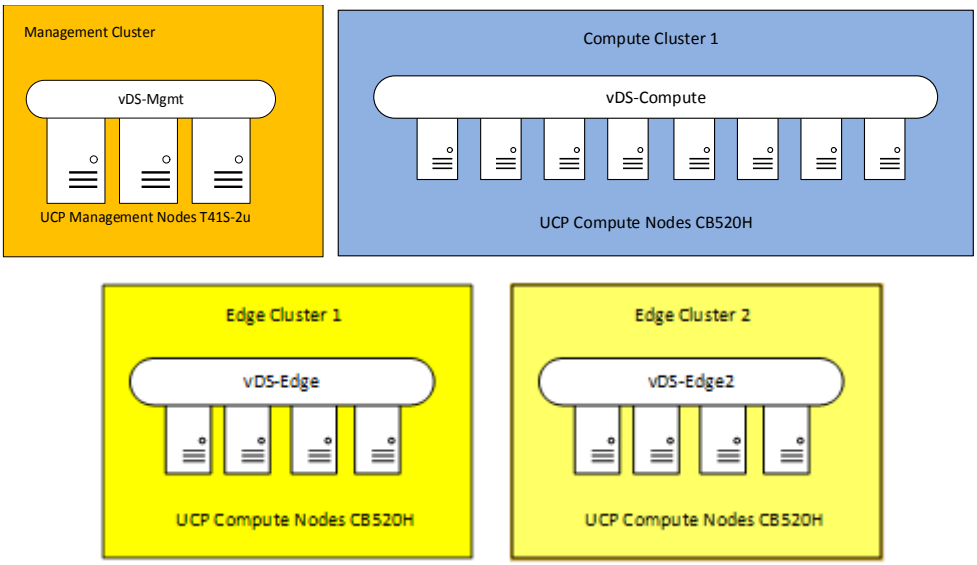
Each host has two NICs that can be teamed in different modes that serve different purposes.

To ensure that all NIC resources can be used in an active/active manner, the teaming mode is set to “route based on originating virtual port”. This way, a VM is bound to one uplink port. If this uplink port fails, the VM is switched to the remaining uplink port.

This teaming policy does not require any extra configuration on Cisco Nexus switches.

vSphere Distributed Switches Summary

There are four different vDS that span different clusters:



Only compute clusters and edge clusters have VXLAN VLAN interfaces. The scaling in vSphere 6.0 is:

- Maximum number of hosts per distributed switch: 1000
- Maximum number of hosts per cluster: 64

This means that all ESXi CB 520H compute hosts can be attached to vDS-Compute, but at least two clusters have to be created.

IP Address Planning for Layer 3 Mode

This table shows the number of IP addresses needed for each VLAN on a rack basis. It does not cover the transfer VLANs/IP addresses needed between Nexus 9372 and Nexus 9332 switches.

Table 3.

	IP Addresses Needed for Rack 1	IP Addresses Needed for Rack 2	IP Addresses Needed for Rack 3	IP Addresses Needed for Rack 4
Hitachi Compute Platform Management VLAN	32	32	32	32
vMotion	32	32	32	32
VXLAN	64	64	64	64
Default Gateway (HSRP)/VLAN	3	3	3	3

VXLAN Configuration

This section describes IP address assignment for virtual tunnel end points) VTEPs. It is assumed that host clusters are able to be distributed across racks.

During VXLAN configuration determined:

- Which vDS will be used for VXLAN
- Which VLAN ID should be used for VXLAN traffic
- What is the MTU size
- What method should be used for VTEP IP address assignment
- Which teaming policy shall be used for uplinks

There is a direct link between the teaming policy and the number of IP addresses needed for VTEPs per host.

This example has a teaming policy of “route based on originating virtual port” so the number of IP addresses changes to 2 per ESXi host.

The MTU is set to 9000 bytes, which is the maximum for vSphere.

IP address assignment via DHCP: Because ESXi host preparation happens on cluster level, a single IP Pool cannot be used to assign IP addresses. This is why a different method needs to be used - dynamic host configuration protocol (DHCP).

DHCP Server for VTEPs

If a Cisco Nexus 9372 can act as a DHCP server, it would be ideal to configure it as one given the fact that it already acts as the default gateway for VXLAN traffic.

If not, another device in the management cluster would have to fulfill this function. In this case it would have to provide all four VTEP ranges with IP addresses.

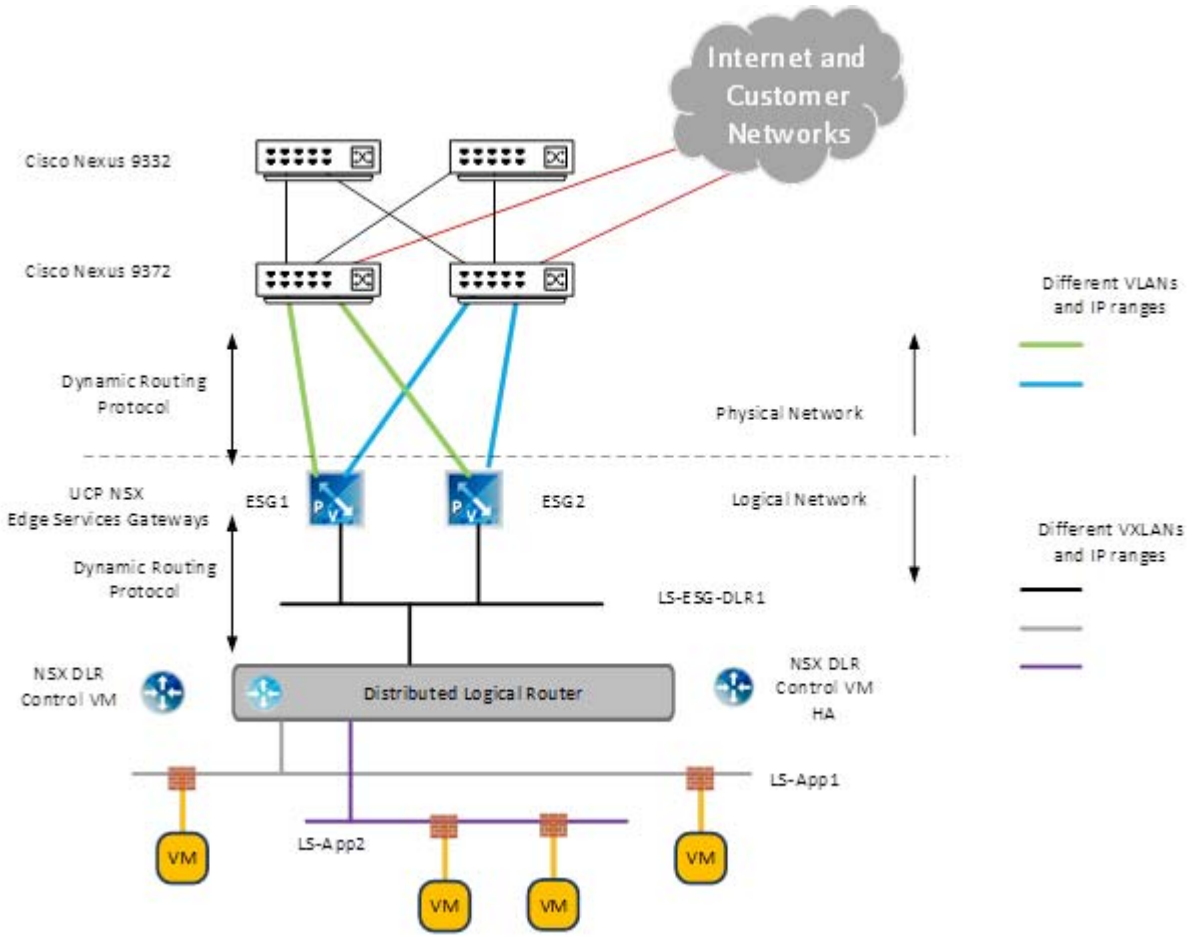
Cisco Nexus 9372 switches would be configured for DHCP relay and forward the DHCP requests to the configured DHCP server.

Logical Networks in Cisco Layer 3

The principle of logical networking remains unchanged. The difference is that the routing peer for NSX ESGs changes compared to Cisco Layer 2 to the top of rack switches, which are Cisco Nexus 9372.

Customers can also peer with their routing devices using Nexus 9372 via 10 Gb/sec interfaces or choose Cisco Nexus 9332 switches with 40 Gb/sec interfaces as peering partners.

Cisco Nexus layer 3 fabric will have to ensure that routing is provided between customer networks and logical networks via NSX ESGs.



For More Information

Hitachi Data Systems Global Services offers experienced storage consultants, proven methodologies and a comprehensive services portfolio to assist you in implementing Hitachi products and solutions in your environment. For more information, see the Hitachi Data Systems Global Services website.

Live and recorded product demonstrations are available for many Hitachi products. To schedule a live demonstration, contact a sales representative. To view a recorded demonstration, see the Hitachi Data Systems Corporate Resources website. Click the **Product Demos** tab for a list of available recorded demonstrations.

Hitachi Data Systems Academy provides best-in-class training on Hitachi products, technology, solutions and certifications. Hitachi Data Systems Academy delivers on-demand web-based training (WBT), classroom-based instructor-led training (ILT) and virtual instructor-led training (vILT) courses. For more information, see the Training and Certification page on HDS.com.

For more information about Hitachi products and services, contact your sales representative or channel partner or visit HDS.com.

Hitachi Data Systems

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 96050-2639 USA
www.HDS.com community.HDS.com

Regional Contact Information
Americas: +1 866 374 5822 or info@hds.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hds.com
Asia Pacific: +852 3189 7900 or hds.marketing.apac@hds.com

AS-527-00 August 2016.