Tableau Breackout a envoyé + confirmation reboot brackout + maj du schéma + remote Leaf template

# Harware Spec

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Platform | Specifications | Access Leaf | Border Gateway | Border MPLS | Spine |
| Nexus 9364D-GX2A | 64 x 400G |  |  |  |  |
| Nexus 9332D-GX2B | 32 x 400G |  |  |  |  |
| Cisco Nexus 9316D-GX | 16 x 400G |  |  |  |  |
| Nexus 93600CD-GX | 28 x 100G & 8 x 400G |  |  |  |  |
| Nexus 10/1 G | TBD |  |  |  |  |

## Leaf to Spine

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Leaf Platform | Specifications | 400G Interfaces | Planned uplink ports | Description |
| Nexus 93600CD-GX | 28 x 100G & 8 x 400G | E1/29-36 | e1/35-36 | Maximum 6 spines |
| Nexus 10/1 G | TBD | TBD | TBD | TBD |

Having 10G Leaf connected as Witness Leaf (via BGW)

## Leaf to Leaf (VPC Peer-link), Leaf to Servers, Border/service leaf to Firewalls and Gateways Connection

First 400G port to be used as peerlink

|  |  |  |  |
| --- | --- | --- | --- |
| Leaf Platform | Specifications | 400G Interfaces | Planned vPC ports |
| Nexus 93600CD-GX | 28 x 100G & 8 x 400G | e1/29-36 | e1/29-30 |
| Nexus 10/1 G | TBD | TBD | TBD |
| Nexus 93108TC-FX | 48 x 100M/1/10GBASE-T & 6 x 40/100 | e1/49 - 54 | e1/53 - 54 |
| Nexus 9336C-FX2 | 36 x 25/40/100G | e1/1 - 36 | e1/35 - 36 |

## Breakout for Nexus

<https://www.cisco.com/c/en/us/td/docs/dcn/nx-os/nexus9000/101x/configuration/interfaces/cisco-nexus-9000-nx-os-interfaces-configuration-guide-101x/m-overview.html#concept_855836F7B4C24198891FC6A5FACF6DE7>

Please note that we need to reboot the module (leaf) to apply the configuration

 The module is reloaded and the configuration for the interface is removed when the command is executed.

We need to read : Cisco Nexus N9K-C93600CD-GX breakout considerations:

# Underlay

# IP addressing :

We use /30 subnet in the whole fabric, another option is to use /31 if don’t have sufficient network space

***Configuration -* P2P Configuration**

**Prefer to Use unnumbered if no constrains found during setup**

**! Configuration on Leaf Interface**

LEAF1(config)#Interface e1/35

LEAF1(config-if)#Description \_Uplink\_To\_Spine

LEAF1(config-if)#ip address 10.10.10.1/30

**! Configuration on Spine Interface**

SPINE1(config)#Interface e1/1

SPINE1(config-if)#Description \_Link\_To\_Leaf

SPINE1(config-if)#ip address 10.10.10.2/30

# MTU :

For all Underlay traffic (VTEP to VTEP) interfaces need to be at MTU 9100

interface x/x

description UNDERLAY interface

mtu 9100

# MTU for External connections:

Depending on the device (PE MPLS, ESX, BG, Client switch…) MTU need to be set between 1500-9000

interface x/x

description OVERLAY interface toward servers/PE/Internet …

mtu *<1500-9000>*

!

interface vlan 1000  
description OVERLAY interface toward servers

mtu *9000*

!

# TCP MSS on all switches

In order to avoid fragmentation in BGP packet, it’s a best practice to rise the default MSS value on the switch from 536 to 8960

!  
configuration terminal

ip tcp mss 8960

!

# ISIS :

We use ISIS Level-1 topology as it is the only supported topology for Nexus VXLAN Fabric.  
The overload bit is used by Nexus to signal other devices not to use the switch as an intermediate hop in their shortest path first (SPF) calculations (on startup)

feature isis

router isis UNDERLAY

log-adjacency-changes

net 49.0001.0010.0100.1001.00

is-type level-1

set-overload-bit on-startup 60

For each interface from Leaf to Spine we use this sample to add isis in the routing interface

The same configuration should be implemented between elements:

* Access Leaves and Border Leaves (PE, Internet) to Spines
* BGW – Spines,

interface Ethernet 1/35

description Link to Spine S1

mtu 9100

ip address 10.10.10.1/31

ip router isis UNDERLAY

The loopback0 is used as router id for isis

interface loopback 0

ip address 10.20.10.1/32

ip router isis UNDERLAY

# ECMP

By default, isis load balance between 8 uplink, if we need in the future more then 8 spine we will use

router isis UNDERLAY

maximum-paths <number>

# Multi-destination traffic

We have two options for BUM traffic, Unicast mode (Ingress replication ) or Multicast mode, The solution is based on Multicast mode

In multicast mode each VNI is mapped to a Mcast GRP based on PIM ASM (AnySourceMcast)

feature pim

!

interface Ethernetx/y

description Link to Spine/Leaf/BGW

ip pim sparse-mode

# RP Placement

Need to discuss with NERIM the best way

|  |  |
| --- | --- |
| Fabric options | Anycast RPs Placement |
| Option 1 | 2 RPs configured on 2 Spines (one in each site) |
| Option 2 | 2 RPs configured on 2 Spines in each site |

Template for 1 site to be replacted (with another IP scheme) in the second site

**! Loopback Interface Configuration on each RP, enable PIM on Lo0 IGP Interface**

interface loopback0

description IGP Loopback Router\_ID

ip address 10.10.10.x/32

ip pim sparse-mode

**! Loopback Interface Configuration (Anycast RP) on all RPs (Spine)**

interface loopback 254

description Anycast RP

ip address 10.10.10.254/32

ip pim sparse-mode

**! Anycast-RP Configuration on all RPs (Spine / site)**

ip pim rp-address 10.10.10.254 group-list 224.0.0.0/8

ip pim anycast-rp 10.10.10.254 10.10.10.1 (Spine-1 IP)

ip pim anycast-rp 10.10.10.254 10.10.10.2 (Spine-2 IP)

**! Configure RP for mcast group on all Leafs**

ip pim rp-address 10.10.10.254 group-list 224.0.0.0/8

# Mulicast grouping for VXLAN:

There are three main approaches to map L2VNIs to Multicast groups:

— One-to-One L2VNI to Multicast Group mapping

— VRF-based L2VNI to Multicast Group mapping

— Odd & Even L2VNI to Multicast Group mapping

The less scale and most simple is One to One, we can make as well VRF based or ODD & Even in order to scale more and limit the number of MCast groupe spread over the Fabric

*One-to-One L2VNI to Multicast Group mapping*

interfcae nve 1

member vni 10011

mcast-group 225.1.1.11

member vni 10012

mcast-group 225.1.1.12

*VRF-based L2VNI to Multicast Group mapping*

interfcae nve 1

member vni 2000201

mcast-group 225.1.1.12

member vni 2000401

mcast-group 225.1.1.12

member vni 2000402

mcast-group 225.1.1.12

*Odd & Even L2VNI to Multicast Group mapping*

interfcae nve 1

member vni 2000201

mcast-group 225.1.1.11

member vni 2000401

mcast-group 225.1.1.11

member vni 2000402

mcast-group 225.1.1.12

We can have a mix of approach depending on the infra (Servers, ESX cluster, Switch client …)

# BFD

TBD, we do not recommend BFD for underlay (only direct back to back links) and we do it for Overlay (BGP connection for BGW inter-site)

2 choice:

Tune eBGP evpn and ISIS (underlay)

Activate BFD to NSX (Via Breackout) to double check that no bug are present

# VXLAN BGP EVPN

RR interconnection to be discussed 🡪 No

Two route reflectors in each site (in all templates) will be configured. RRs will be on two spine switches in each site.

EVPN BGP neighborship (to be discussed)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| BGP Peer A | BGP Peer B | Site | DC1 | DC2 |
| Leafs & BGWs | Spines | Same site |  |  |
| RR Spines Site 1 ? | RR Spines Site n ? | All sites |  |  |
| BGW Site 1 | BGW Site n | All sites |  |  |

# BGP EVPN on leaves

**Template for VxLAN and EVPN + VLAN and VXLAN VNI**

**! Enable VLAN-based VXLAN**

feature vn-segment-vlan-based

**! Enable VXLAN**

feature nv overlay

**! Enable the EVPN control plane for VXLAN**

nv overlay evpn

vlan <vlan-number>

vn-segment <vn-number>

!

evpn

vni <20000+VLANID> l2

rd *RID:L2VNI*

route-target both *ASN:L2VNI*

!

Note : VNI, RD and RT could be adapted to NERM needs

**Template for VRF for VXLAN Routing**

**To Add Strategy of Import/Export for NSX 🡪 To discuss site 2 prepending and rewrite feature in BGW (RT of NSX)**

vrf context <context-name>

vni <*40000+vrf-id*>

rd *RID:L3VNI*

address-family ipv4 unicast

route-target both *ASN:L3VNI*

route-target both *ASN:L3VNI* evpn

address-family ipv6 unicast

route-target both *ASN:L3VNI*

route-target both *ASN:L3VNI* evpn

!

VNI, RD and RT could be adapted to NERIM needs

**Template for SVI hosts for VXLAN routing**

vlan <vlan-number>

vn-segment <vn-number>

!

interface vlan <vlan-number>

vrf member <vrf-name>

ip address <ip-address> <netmask>

!

Template for VRF Overlay VLAN for VXLAN routing

**! Create the VRF overlay VLAN and configure the vn-segment**

vlan <vlan-number>

vn-segment <vn-segment>

!

**! Configure VRF overlay VLAN/SVI for the VRF**

interface VlanX

no shutdown

vrf member <”overlay vrf name”>

ip forward

**!The system vlan nve-overlay id command should be used for a VRF or a Layer-3 VNI (L3VNI) only. Do not use this command for regular VLANs or Layer-2 VNIs (L2VNI).**

system vlan nve-overlay id <>

**! Create VRF and configure VNI**

vrf context CUSTOMER-VRF

vni <vni-number>

Anycast Gateway for VxLAN Routing: Each Fabric (site) should have individual MAC address. This MAC address is shared by the anycast gateway for all edge devices of the VxLAN fabric.

Template for configuring the anycast gateway (To reconfirm)

**! Configure distributed gateway virtual MAC address**

**! All VTEPs should have the same virtual MAC address**

fabric forwarding anycast-gateway-mac <mac-address>

!

interface vlan <vlan-number>

fabric forwarding mode anycast-gateway

!

Template for configuring NVE interface and VNIs

interface nve-interface

**! This defines BGP as the mechanism for host reachability advertisement**

host-reachability protocol bgp

**! Add Layer-3 VNIs, one per tenant VRF, to the overlay**

member vni <vni-number> associate-vrf

**! Add Layer 2 VNIs to the tunnel interface**

member vni <vni-number>

**! Configure the mcast group on a per-VNI basis (depending on MCAST strategy**

mcast-group <multicast-group-address>

**! advertise virtual rmac with advertise pip in bgp**

advertise virtual-rmac

**! use Loopback1 (new loopback for VTEP peering**

source-interface loopback1

!

Template for configuring BGP on the leaves and BGW

feature bgp

**! Need to discuss Public AS vs Private AS (use template-peer)**

router bgp <ASN-number>

log-neighbor-changes

address-family l2vpn evpn

**! Requires advertise virtual-rmac in NVE**

advertise-pip

**! Recommended to provision the same as IP address of interface loopback 0 for IGP**

router-id <address>

**! Provision Spines as RR**

**! ASN-number shuld remain the same and consistent within each site**

**! Using TemplateBGP**

template peer iBGP-Spines

remote-as <ASN-number>

**! Use loopback 0 as source interface for BGP peering**

update-source loopback0

**! Configure address family Layer 2 VPN EVPN under the BGP neighbor.**

address-family l2vpn evpn

send-community

send-community extended

**! Applying template to RR**

neighbor <RR1-address>

inherit peer iBGP-Spines

description Spine-1 Loopback0

neighbor <RR2-address>

inherit peer iBGP-Spines

description Spine-2 Loopback0

**!!! VRF client**

vrf <vrf-name>

address-family ipv4 unicast

**! Configure iBGP Multipath Load Sharing (Leaf will select multiple iBGP paths as the best paths to a destination.The best paths or multipaths are then installed in the IP routing table)**

maximum-paths ibgp 2

advertise l2vpn evpn

**! In case you're willing to advertise Type 5, you need network/redistribute statements. This is for Static routing (to be avoided)**

network <ip-network/subnet>

!

address-family ipv6 unicast

advertise l2vpn evpn

!

!

# BGP EVPN on Spines

Template for configuring BGP on the spines without VTEP functionality

feature bgp

**! Need to discuss Public AS vs Private AS (use template-peer)**

router bgp <ASN-number>

log-neighbor-changes

address-family l2vpn evpn

**! Recommended to provision the same as IP address of interface loopback 0 for IGP**

router-id <address>

**! Use template for RR client**

template peer iBGP-Leafs

remote-as <ASN-number>

update-source loopback0

address-family l2vpn evpn

send-community

send-community extended

route-reflector-client

**! Applying template to Leaf**

neighbor <LEAF1-address>

inherit peer iBGP-Leafs

description LEAF-1 Loopback0

!

# Multisite

# Multisite Limitation

VXLAN EVPN Multi-Site has the following configuration guidelines and limitations:

* The Multi-Site border gateway allows the co-existence of Multi-Site extensions (Layer 2 unicast/multicast and Layer 3 unicast) as well as Layer 3 unicast and multicast external connectivity.
* Anycast mode can support up to six border gateways per site.
* Beginning with Cisco NX-OS Release 9.2(1), Border Gateways (BGWs) in a vPC topology are supported.
* Support for Multicast Flood Domain between inter-site/fabric border gateways is not supported.
* Multicast Underlay between sites is not supported.
* iBGP EVPN Peering between border gateways of different fabrics/sites is not supported.
* Anycast mode can only support Layer 3 services that are attached to local interfaces.
* In Anycast mode, BUM is replicated to each border-leaf and DF election, between Border Leafs of a particular site decides which border leaf would forward the traffic inter-site traffic (Fabric to DCI and vice versa) for that site.
* In Anycast mode, all the Layer 3 services are advertised in BGP via EVPN Type-5 routes with their physical IP as the next hop.
* vPC mode can support only two border gateways.
* vPC mode can support both Layer 2 hosts and Layer 3 services on local interfaces.
* In vPC mode, BUM is replicated to either of the border-gateway’s for traffic coming from external site and hence both the border gateways are forwarders for site external to site internal (DCI to Fabric) direction.
* In vPC mode, BUM is replicated to either of the border gateways for traffic coming from the local site leaf for a VLAN using Ingress Replication (IR) underlay. Both border gateways are forwarders for site internal to site external (Fabric to DCI) direction for VLANs using the IR underlay.
* In vPC mode, BUM is replicated to both border gateways for traffic coming from the local site leaf for a VLAN using the multicast underlay. Therefore, a decapper/forwarder election happens and the decapsulation winner/forwarder only forwards the site-local traffic to external site border-gateways for VLANs using the multicast underlay.
* In vPC mode, all the Layer 3 services/attachments are advertised in BGP via EVPN Type-5 routes with their virtual IP as next hop. If the VIP/PIP feature is configured, they are advertised with PIP as the next hop.
* If different Anycast Gateway MAC addresses are configured across sites, ARP suppression must be enabled for all VLANs that have been extended.
* Bind NVE to a loopback address that is separate from loopback addresses that are required by Layer 3 protocols. A best practice is to use a dedicated loopback address for the NVE source interface (PIP VTEP) and Multi-Site source interface (anycast and virtual IP VTEP).

## Enabling VXLAN EVPN Multisite

Enable VxLAN and EVPN on BGW

**! Enable VLAN-based VXLAN**

feature vn-segment-vlan-based

**! Enable VXLAN**

feature nv overlay

**! Enable the EVPN control plane for VXLAN**

nv overlay evpn

**Enable VxLAN EVPN multi-site**

The following configuration enables VxLAN EVPN Multi-site Feature. Multi-site is enabled on BGW only. Site-id must be the same on all border gateways of the site.

3 Loopback on BGW : Looback0 Router-ID

Loopback1 VTEP Source (Public IP)

Loopback 2 : VTEP MultiSite (Public IP)

For example

evpn multisite border-gateway <site-id>

**! Delay restore for advertisement of Anycast BGW IP**

delay-restore time 180

!

interface Loopback 1

description VTEP Source

address <a.b.c.d>/<mm> secondary

!

interface Loopback 2

description Multisite VTEP

ip address <a.b.c.d>/<mm>

ip pim sparse-mode

!

interface nve1

source-interface loopback 1

host-reachability protocol bgp

multisite border-gateway interface loopback 2

member vni <vni-id>

suppress-arp

multisite ingress-replication

mcast-group <multicast-group-address>

!

!

router bgp <ASN-number>

neighbor <address-BGW >

remote-as <ASN-number-remote>

description Site-External Peer (Other Site BGW, PGW or RS or Witness)

update-source loopback1

ebgp-multihop 2 (5 for Witness)

password 0 <password>

**! Fabric-external configuration, only for site-external peers**

peer-type fabric-external

address-family l2vpn evpn

send-community

send-community extended

rewrite-evpn-rt-asn

!

!

neighbor <address-Spine>

remote-as <ASN-number-internal>

description Site-Internal Peer (Site Internal Spine)

update-source loopback0 (kept as other Internal Leaf)

password 0 <password>

address-family l2vpn evpn

send-community

send-community extended

rewrite-evpn-rt-asn

!

!

!

**Configuring Fabric/DCI Link Tracking**

interface Ethernet<x>/<y>

description Site-Internal

evpn multisite fabric-tracking (Only for DC1 <> DC2)

!

interface Ethernet<x+1>/<y+1>

description Site-Internal

evpn multisite dci-tracking (Only for DC1 <> DC2)

!

## BGW BUM Traffic Enforcement

This is to allow the vPC BGW level to limit the propagation of BUM traffic types in aggregate toward the remote sites

evpn storm-control broadcast level 0-100 (0,1)

evpn storm-control multicast level 0-100 (0,1)

evpn storm-control unicast level 0-100 (0,1 To be confirmed)

## Consideration for adding more then two sites

EVPN Multi-Site architecture requires every BGW from a local site to peer with every BGW at remote sites. We have two option, full-mesh and Route Servers spread across sites within DCI Underlay

## ARP Suppression

Configuring the hardware access-list tcam region arp-ether <size> double-wide is not required on Cisco Nexus 9200, 9300-EX, 9300-FX, 9300-GX, and 9300-FX2 platform switches

interface nve 1

**!enable arp suppression globally on all L2VNIs**

global suppress-arp

member vni <vni-id>

!

!

NGOAM

It is recommended to enable VXLAN operations, administration, and maintenance (OAM) on all Switches in the fabric to facilitate installing, monitoring, and troubleshooting Ethernet networks to enhance management in VXLAN based overlay networks.

Similar to ping and traceroute utilities that allow quick determination of problems in the IP networks, VXLAN OAM provides very similar ping, pathtrace, and traceroute utilities to diagnose the reachability to the hosts and the VTEPs in a VXLAN network. The OAM channel is used to identify the type of the VXLAN payload that is present in these OAM packets.

There are two types of payloads supported:

— Conventional ICMP packet to the destination to be tracked

— Special NVO3 draft Tissa OAM header which is used in pathtrace and carries useful information about ingress and egress interface and interface statistics on a per-hop basis.

feature ngoam

ngoam install acl

ngoam profile 1

oam-channel 2

ngoam authentication-key <key-string>

Example:

ping nve ip <a.b.c.d> vrf <vrf-name> source <a.b.c.d> verbose

pathtrace nve ip <a.b.c.d> profile <number> vrf <vrf-name> vni <nvid> req-stats verbose

VPC

Virtual Port-Channel (vPC) VTEPs combines two technologies, VPC and VXLAN, to provide device-level redundancy for VTEPs. A pair of vPC switches share the same VTEP address (anycast VTEP address)

VPC Best Practices.

* **vPC Delay Restore** – After vPC peer reload, traffic might be black-holed, before L3 connectivity is reestablished. vPC links bring up can be delayed allowing Underlay and Overlay Convergence. Default timer is 30 seconds. Relaxing this timer to 150 seconds does provide better convergence time by allowing more time for the Underlay and Overlay convergence before bringing up the VPCs.
* **SVI Delay Restore** – this feature delays SVI bring-up to allow Underlay and Overlay converge after vPC peer reload. After reloading, traffic might be black-holed. Default timer is 10 seconds.
* **NVE Hold-Down timer** – this timer suspends advertisements of NVE loopback interface until overlay has converged after vPC peer reload. After vPC peer reload, traffic going to Anycast VTEP hashed to the peer will be black-holed. NVE loopback interface bring up can be delayed using this hold-down timer. For proper overlay convergence, hold-down time needs to be longer than delay restore time. Default timer is 180 seconds. A timer of 220 seconds provides more time for the vPC port members to come up.
* **vPC Peer-Switch** – this feature makes the vPC peer devices to appear as single STP root and BPDUs processed by the logical STP root formed by the 2 vPC peer devices.
* **Unique Domain ID’s** – use unique Domain ID’s for vPC-to-vPC for all vPC pairs defined in a contiguous layer 2 domain. The vPC peer devices use the vPC domain ID to automatically assign a unique vPC system MAC address which must be unique. This system MAC is used in STP BPDU, LACP BPDU, and IGMP advertisements.
* **vPC ARP/ND synchronization** – this feature suspends SVIs when peer device goes down or peer link goes down. After restoring of the peer device, or peer link, ARP table is empty and traffic black-holed. Before bringing up SVI, peer devices synchronize ARP table over CFS service.
* **The vPC peer link should be resilient**. A recommended practice is for the peer-link port-channel to contain links at least of two separate ports if possible.
* **vPC Infrastructure VLANs** – these VLANs are used for backup routing path and present on Peer Link and used in case of failure of uplinks on vPC peer or failure of vPC port members. Also, it is used for BUM traffic transfer.
* **Bridge Assurance** should be configured on the vPC peer link ends. Entering the vPC peer link command does this automatically. ISSU requires having no Bridge Assurance on vPC member links.
* **Peer-gateway** allows a vPC peer device to act as the active gateway for packets addressed to the other peer device router MAC. It keeps the forwarding of traffic local to the vPC peer device and avoids use of the peer-link.
* **Auto-recovery** restores vPC services when its peer fails to come online. You must save this setting in the startup configuration. On reload, if the peer link is down and three consecutive peer-keepalive messages are lost, the secondary device assumes the primary STP role and the primary LACP role. The software reinitializes the vPCs, bringing up its local ports. Because there are no peers, the consistency check is bypassed for the local vPC ports. The device elects itself to be the STP primary regardless of its role priority and also acts as the master for LACP port roles (default value 240 sec). It is recommended to use an extended timer of 360 sec which is more secure in case of both VPC peers reloading at the same time.
* **LACP vpc-convergence:** when configured the switch waits until all the VLANs are initialized and programmed and then send LACP sync PDU, which will start sending traffic to the VPC domain without drops. You may configure the lacp vpc-convergence command in a VXLAN and non-VXLAN environments that have vPC port-channels to hosts that support LACP. This command must be enabled on both vPC peer switches. This command must be configured only on PortFast/Edge ports (vPC port channels on which the spanning-tree port type edge [trunk] is enabled).
* **vPC orphan-port suspend**: to suspend a nonvirtual port channel (vPC) port when the peer link of a vPC secondary goes down, use the vpc orphan-port suspend command. This command will reduce convergence time on Border Leaf with non vPC L3 routed ports connection to Edge Router. This command can also be used on all non vPC ports (except Fabric Uplinks to the Spines), especially in the scenario where servers are using active-standby type of teaming, forcing a physical shutdown on the link for the teaming software to change the traffic to flow on the other link connected to the primary.

VPC & VXLAN consideration.

* For VPC, the loopback interface has 2 IP addresses: the primary IP address and the secondary IP address. The primary IP address is unique and is used by Layer 3 protocols. The secondary IP address on the loopback is necessary because the interface NVE uses it for the VTEP IP address, particularly for all VxLAN traffic that includes multicast and unicasts encapsulated traffic. The secondary IP address must be the same on both vPC peers.
* Each VPC peer needs to have separate BGP sessions to the spine.
* VPC peers must have identical configurations.
  + Consistent VLAN to VN-segment mapping
  + Consistent NVE1 binding to the same loopback interface
  + Using the same Secondary IP address
  + Using different primary IP addresses
  + Consistent VNI to group mapping
  + The VRF overlay VLAN should be a member of the peer-link port-channel
* VPC “peer-gateway” feature must be enabled on both peers.
* When Nexus 9000 switches are configured as VXLAN Leaf Switches also known as VXLAN Tunnel End Points (VTEP) in virtual Port Channel (vPC) domain, you must have a backup Layer 3 Routing adjacency in between them over the vPC peer-link with the use of an interface vlan. This VLAN must be local to the switches, not stretched across the VXLAN fabric and belong to the Default VRF (Global Routing Table).
* It is worth noting that once a pair of VXLAN switches is configured as part of a vPC domain, the anycast VTEP is always used as next-hop for all the EVPN advertisements relative to directly connected endpoints. This is valid also for local end points connected in Active/Standby fashion. The consequence is that roughly half of the flows destined to those devices may be delivered from the spines to the VTEP device connected to the standby end points (the spines have two equal cost paths to reach the Any cast VTEP IP address); the traffic would hence have to take an extra hop across the peer-link in order to be delivered to the active interface of the endpoint. This suboptimal behavior can be avoided by grouping endpoints based on the types of connectivity (Active/Standby vs LACP) and connecting them to separate sets of leaf switches.

feature lacp

feature vpc

!

interface mgmt0

vrf member management

ip address <a.a.a.a>/<netmask>

!

vpc domain 100

peer-switch

**! Priority has to be provisioned different on pair of Leafs**

role priority {4096 | 8192}

peer-keepalive destination <b.b.b.b> source <a.a.a.a> vrf management eq 3200

auto-recovery

delay restore 150

delay restore interface-vlan 10

auto-recovery reload-delay 360

peer-gateway

**! Configure layer3 peer-router under VPC domain, this command is needed as dynamic routing is configured on SVI 3967 over VPC peerlink**

layer3 peer-router

ip arp synchronize

!

interface nve1

source-interface hold-down-time 220

!

**! Minimum two interfaces should be part of vPC PL**

interface ethernet <number>/<x - x+1>

description vPC PL

switchport

channel-group <id> mode active

no shutdown

!

interface port-channel1

description vPC PL

switchport

switchport mode trunk

switchport trunk native vlan 2

switchport trunk allowed vlan <range>

spanning-tree port type network

vpc peer-link

interface port-channel <number>

switchport

switchport mode trunk

switchport native vlan 2

switchport trunk allowed vlan <range>

! **If not connected to another switch**

spanning-tree port type edge trunk

lacp vpc-convergence

vpc <number>

!

interface loopback1

ip address x.x.x.x/32 secondary

!

**Infra VLAN Leaf01**

!  
vlan 3967

name BACKUP\_VLAN\_ROUTING\_NVE\_INFRA

!

interface Vlan3967

no shutdown

no ip redirects

ip address 10.1.2.1/24

no ipv6 redirects

mtu 9100

**ip router isis UNDERLAY (Why we need to annonce it into the underlay)**

**ip pim sparse-mode**

!

**system nve infra-vlans** 3967

!

**Infra VLAN Leaf02**

vlan 3967

name BACKUP\_VLAN\_ROUTING\_NVE\_INFRA

!

interface Vlan3967

no shutdown

no ip redirects

ip address 10.1.2.2/24

no ipv6 redirects

mtu 9100

**ip router isis UNDERLAY (Why we need to annonce it into the underlay)**

**ip pim sparse-mode**

!

**system nve infra-vlans** 3967

!

Classical Ethernet & MST

Only when needed

spanning-tree mst configuration

name *<name>*

revision *<vlaue>*

instance 1 vlan *<range>*

spanning-tree mode mst

External Connectivity with eBGP Routing

TBD

**!Inside Interface connecting to BareMetal**

Interface ethernet x/y

Switchport

mtu *<1500-9000>*

switchport mode trunk

spanning-tree port type edge trunk

switchport trunk native vlan 2

switchport trunk allowed vlan 10,20

channel-group x mode active

Interface port-channel x

switchport

switchport mode trunk

switchport trunk native vlan 2

switchport trunk allowed vlan 10,20

spanning-tree port type edge trunk

lacp vpc-convergence

vpc x

**using route-map to filter Inside / Outside routes**

route-map RM-OUTSIDE-Client1--Static-to-BGP permit

match ip address perfix-list PL-OUTSIDE-B1-STATIC

**!Create prefix-list for Filtering Host Routes advertisement to external router**

ip prefix-list PL-HOST-ROUTE seq 5 permit 0.0.0.0/0 eq 32

route-map RM-EXT-HOST-ROUTE-FILTER deny 10

match ip address prefix-list PL-HOST-ROUTE

route-map RM-EXT-HOST-ROUTE-FILTER permit 1000

router bgp 65000

vrf Client-1

address-family ipv4 unicast

redistribute static route-map RM-OUTSIDE-Client1-Static-to-BGP

redistribute direct route-map RM-DIRECT

neighbor 10.10.10.x

remote-as 650001

address-family ipv4 unicast

route-map RM-EXT-HOST-ROUTE-FILTER out

vrf Client-2

address-family ipv4 unicast

redistribute static route-map RM-OUTSIDE-B2-Static-to-BGP

redistribute direct route-map RM-DIRECT

neighbor 10.10.11.y

remote-as 65002

address-family ipv4 unicast

route-map RM-EXT-HOST-ROUTE-FILTER out

QinVNI

TBD

OOB

TBD

NX-OS Programmability & Telemetry (YANG, AAA, SNMP, Logging, NYP, NGNI…)

TBD

Security (ISIS authen, NTP Authen, BGP Auth, Management)

**!Preventing VLAN Hopping**

vlan dot1Q tag native

vlan 2

shutdown

name NATIVE

vlan <vlan-number>

vn-segment <vn-number>

Controle Plan (DataPlane, Storm Control, Spanning-Tree)

Security (ACL for BareMetal)

TBD

RemoteLeaf (Witness Template)

TBD