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 8 spines |
| Nexus 10/1 G | TBD | TBD | TBD | TBD |

## 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 |

# 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**

**! 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 225.1.1.11

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)

# VXLAN BGP EVPN