Low Level Design

Datacentre Network Refresh

University of Luxembourg

Multi-Site BGP EVPN VXLAN Fabric

Authors:

Manager:

11 May 2024 | Document Version 1.0

Table of Contents

1. History 5

2. Introduction 6

3. Project Overview 7
3.1. Current High Level Design Overview 7
3.2. Target High Level Design 7

4. Physical Design 11
4.1. BGP EVPN VXLAN Solution components 11
4.2. Datacentre Interconnect Back-to-back vs ISN (Core) 11
4.3. Spine Switches and Core Switches 12
4.4. Leaves 17
4.5. Nexus Dashboard 31
4.6. Connectivity Matrix 33

5. Naming Convention 34

6. Logical EVPN VXLAN Fabric Design 36
6.1. Design Key Points 36
6.2. VXLAN 37
6.3. Underlay Network 38
6.4. Control Plane 41
6.5. VXLAN BGP EVPN Enhancements 47
6.6. Overlay Data Forwarding 51
6.7. Management 54
6.8. Guidelines and Limitations for VXLAN EVPN Multi-Site (version 10.4(x)) 54

List of Figures

Figure 1. EVPN vs ACI 7
Figure 3. Multi-Site with ISN 12
Figure 4. Cisco Nexus 9504 Physical Layout 13
Figure 5. Cisco N9K-X9736C-FX Line Card 13
Figure 6. Cisco N9K-C9504-FM-G Fabric Module 15
Figure 7. Cisco N9K-SUP-A+ Supervisor Module 16
Figure 8. Cisco N9K-C9336C-FX2 Leaf Switch 17

Figure 9. Cisco QSFP-100G-SR4-S 100G Transceiver 20
Figure 10. Cisco QSFP-40G-SR4-S Transceiver 20
Figure 11. Cisco N9K-C93180YC-FX3 Leaf Switch 21
Figure 12. Cisco QSFP-100G-SR4-S 100G Transceiver 23
Figure 13. Cisco QSFP-40G-LR4-S Transceiver 24
Figure 14. Cisco QSFP-100G-SR4-S 100G Transciever 24
Figure 15. Cisco N9K-C93180YC-FX Leaf Switch 25
Figure 16. Cisco QSFP-40G-SR4-S Transceiver 27
Figure 17. Cisco SFP-10G-T-X Transceiver 27
Figure 18. Cisco N9K-C93180YC-EX Leaf Switch 28
Figure 19. Cisco N9K-C93108TC-EX Leaf Switch 28
Figure 20. Cisco QSFP-40G-SR4-S Transceiver 30
Figure 21. Cisco QSFP-40G-LR4-S Transceiver 31
Figure 22. Cisco SFP-10G-T-X Transceiver 31
Figure 23. VXLAN Tunnel 37
Figure 24. VXLAN Encapsulation 38
Figure 25. Site-Internal Fabric 38
Figure 26. Site-Internal VTEPs 39
Figure 27. Site-External Example 40
Figure 28. Multi-Site with Two Sites 41
Figure 29. Site-Internal Underlay CP 41
Figure 30. Site-External Underlay CP 43
Figure 31. Multi-Site Underlay CP 44
Figure 32. Site-Internal Overlay CP 45
Figure 33. Site-External Overlay CP 47
Figure 34. Multi-Site Overlay CP 47
Figure 35. IRB Symmetric 48
Figure 36. IRB-Symmetric Intra-VRF Routing 49
Figure 37. PIP vs VIP for VPC 50
Figure 38. Loop Detection and Mitigation 53

List of Tables

Table 1. MPLS, VRF-Lite and VXLAN BGP EVPN Multi-Site at Core Comparison 9
Table 2. Back-to-Back and ISN Comparison 12
Table 3. N9K-X9736C-FX 100-Gigabit Ethernet Line Card Performance and Scale 14
Table 4. Cisco Nexus 9504 Switch supports fabric module N9K-C9504-FM-G with following characteristics 15
Table 5. Cisco N9K-SUP-A+ supervisor module specification 16
Table 6. Cisco Nexus N9K-C9336C-FX2 switch specification 18
Table 7. Cisco Nexus N9K-C9336C-FX2 switch Hardware performance and scalability specifications 18
Table 8. Cisco Nexus N9K-C93180YC-FX3 switch specification 21

Table 9. Cisco Nexus N9K-C93180YC-FX3 switch Hardware performance and scalability specifications 22
Table 10. Cisco Nexus N9K-C93180YC-FX switch specification Table: Cisco Nexus N9K-C93180YC-FX switch specification 25
Table 11. Cisco Nexus N9K-C93180YC-FX switch Hardware performance and scalability specifications 25
Table 12. Cisco Nexus N9K-C93180YC-EX and N9K-C93108TC-EX switches specifications 28
Table 13. Cisco Nexus N9K-C93180YC-EX switch Hardware performance and scalability 29
Table 14. Nexus Dashboard VM Requirements 32

1. History

2. Introduction

3. Project Overview

3.1. Current High Level Design Overview

3.1.1 Current High-Level Design

3.1.1.1 L1-L2

3.1.1.2 L3

3.2. Target High Level Design

During the pre-sale phase, several potential options for new CompanyA infrastructure design were discussed.

3.2.1 EVPN vs ACI

Table below provides high level comparison between to different SDN solutions for datacentres: ACI and BGP EVPN VXLAN Fabric.

Feature	EVPN VXLAN	ACI
Solution Type	Open Industry-standard	Vendor Proprietary
Model Type	Network Centric Model	Application Centric Model
Flexibility	Can create multiple vendor Fabric	Only Cisco Fabric
Automation	Ansible, YANG, any open standard	Native REST API using XML or JSON, Python
Controller and Management	Can be done by Nexus Dashboard/Fabric Controller	APIC
Learning Curve	Easy to Medium	Medium
Scalability	Easy to integrate other vendor fabric	Can only integrate Cisco Fabric
Operations and Analysis	Manual Scripting	Day – 2 Operations Automation
Use Case	Open Standard Static Network centric DC	Proprietary Application centric with requirement of huge scalability and hundreds of racks
Cisco Validated Design	Yes	Yes

Figure 1. EVPN vs ACI

When comparing two technologies, it's important to focus on their key differences and similarities. Here's a succinct summary:

Underlying Architecture:

- **BGP EVPN VXLAN:** Based on Open Standard.
- **ACI:** ACI is a proprietary SDN solution developed by Cisco.

Deployment Model

- **BGP EVPN VXLAN:** more modular and flexible manner, allowing best-of-breed approach by using hardware and software components from different vendors.
- **ACI:** ACI is a more integrated solution that typically requires organizations to deploy Cisco's Nexus switches and ACI fabric.

Management Approach

- **BGP EVPN VXLAN:** offers a more decentralized management approach, with network policies configured and enforced directly on the network devices.

- **ACI:** offers a centralized management approach where network policies are defined and enforced centrally through the APIC controller.

Management Approach

- **BGP EVPN VXLAN:** offers a more decentralized management approach, with network policies configured and enforced directly on the network devices.
- **ACI:** offers a centralized management approach where network policies are defined and enforced centrally through the APIC controller.

The CompanyA's decision in favor of ACI was strongly influenced by the fact that BGP EVPN VXLAN is an open standard solution, allowing seamless integration with other vendor solutions.

3.2.2 Multi-Site vs Multi-Pod

Another topic of discussion was related to multi-site vs multi-pod solution.

In the context of VXLAN BGP EVPN, both multi-Pod and multi-site architectures enable the interconnection of distinct VXLAN BGP EVPN fabrics within a single overlay domain.

- **Multi-Pod**: This architecture establishes a unified data plane across multiple pods. Traffic between these pods is encapsulated within VXLAN tunnels, which are created between VTEPs (VXLAN Tunnel Endpoints) located in different datacentres. Notably, there is no intermediate decapsulation or encapsulation when traffic crosses datacentre boundaries. In scenarios involving only two datacentres, a back-to-back design is feasible. Connections between pods can be facilitated through either spine or leaf switches.
- **Multi-Site:** This architecture involves dividing the data planes across different sites, resulting in distinct fault domains. This separation of fault domains is a significant advantage over multi-pod configurations.

In the context of VXLAN BGP EVPN multi-site, Cisco has introduced a new leaf role called the Border Gateway (BGW). By leveraging BGWs, disparate fabrics can be smoothly integrated into the multi-site architecture. For redundancy and load balancing, Cisco recommends deploying multiple BGWs per site (up to a maximum of four) in anycast mode.

While both solutions are still supported by Cisco, the vendor strongly recommends using a multi-site approach, which has been considered best practice for some time.

In case of 2 sites, Cisco recommends 4 different topologies:

- BGW-to-cloud model
- BGW back-to-back model
- Model with BGW between spine and superspine
- BGW-on-spine model

Model with BGW between spine and superspine has been chosen.

3.2.3 VRF-Lite vs MPLS

Core replacement is essential part of the project. In current design all sights are interconnect via MPLS.

It has been determined during initial workshops that long-term goal for CompanyA of Luxembourg is to migrate all Campus sites to VXLAN BGP EVPN fabrics.

Redesign of campus sites is outside the scope of this project.

Even though campus re-design is outside this project, Core replacement is important part of the project.

During initial design phase it was determined that there is no route leaking configured between VRFs. All Inter-VRF communication happens through Firewalls. In addition to that, no VPLS or L2VPNs are used by CompanyA of Luxembourg over MPLS cloud. It was determined that it is technically possible to replace MPLS with VRF-Lite technology. Another solution is to use VXLAN BGP EVPN Multi-Site technology to interconnect campus sites and datacentre.

Below table compares MPLS with VRF-Lite and VXLAN BGP EVPN Multi-Site deployment in the Core.

Table 1. MPLS, VRF-Lite and VXLAN BGP EVPN Multi-Site at Core Comparison

Benefits & Drawbacks	MPLS	VRF-Lite	VXLAN BGP EVPN Multi-Site
Benefits	Small number of ISIS and BGP Peers. A little bit more complex configuration. Less changes to current design.	Less technologies involved in the solution Straightforward configuration	Less technologies involved in the solution No need in ISIS Small number of BGP Peers
Drawbacks	New campus devices must support MPLS	High number of ISIS and/or BGP Sessions. More changes to current design.	Each campus site must have devices that support Border Gateway Functionality

After discussion with CompanyA of Luxembourg and consideration of all pros and cons of either solution, it was decided to keep MPLS in the Core as temporal solution. Refer to Logical Core Design Section for more details on each approach.

Target of the project is to simplify current MPLS core by decommissioning soon End of Life Cisco Nexus 7k switches acting as MPLS P routers and Cisco ASR1001-X routers acting as Route Reflectors. Route Reflector roles will be transferred to the new Cisco Nexus 9504 Core devices. There is no need in P routers as campus MPLS PE routers will be directly connected to the new Cisco Nexus 9504 Core devices via dark fibers. Logical MPLS Design remains unchanged.

3.2.4 Nexus Dashboard Fabric Controller

Cisco Nexus Dashboard is a central management console for multiple datacentre sites and a common platform for hosting Cisco datacentre operation services, such as Insights, Orchestrator, and Fabric Controller.

Although it is not planned to use Network Director (ND) for operations after deployment, it appears that NDFC still offers benefits during the implementation and testing phases.

4. Physical Design

The following sections describe the platform choices for CompanyA of Luxembourg.

4.1. BGP EVPN VXLAN Solution components

CompanyA of Luxembourg BGP EVPN VXLAN fabric overlay network comprised of Cisco Nexus Series Switches playing following roles in the fabric:

- Spine Switch
- Leaf Switch
- Border Leaf and Border Gateway Switch
- Core Switch

BGP EVPN VXLAN fabric consists of 2 sites interconnected via Border Gateway Leaves through Core Switches.

DC1 fabric consists of 2 Spines, 2 switches that combine Border Leaf and Border Gateway roles and 65 Leaves. DC2 fabric consists of 2 Spines, 2 switches that combine Border Leaf and Border Gateway roles and 18 Leaves.

Each datacentre also hosts 1 Core Switch for interconnection between datacentres and campuses.

4.2. Datacentre Interconnect Back-to-back vs ISN (Core)

VXLAN EVPN Multi-Site architecture is a design for VXLAN BGP EVPN–based overlay networks. It allows interconnection of multiple distinct VXLAN BGP EVPN fabrics or overlay domains, and it allows new approaches to fabric scaling, compartmentalization, and DCI. VXLAN EVPN Multi-Site architecture provides integrated inter-connectivity that doesn't require additional technology for Layer 2 and Layer 3 extension. It thus offers the possibility of seamless extension between compartments and fabrics. It also allows you to control what can be extended. In addition to defining which VLAN or Virtual Routing and Forwarding (VRF) instance is extended, within the Layer 2 extensions you can also control broadcast, unknown unicast, and multicast (BUM) traffic to limit the ripple effect of a failure in one datacentre fabric.

In a two-site scenario, this can be achieved through a back-to-back approach or via an inter-site network (ISN).

The diagram below presents the case with ISN:

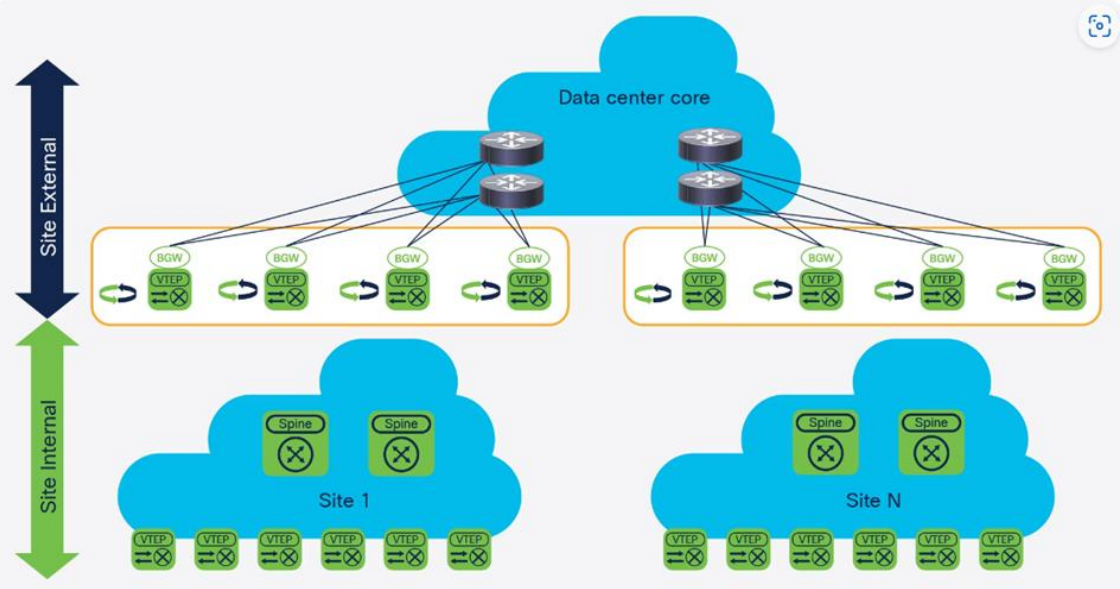

Figure 2. Multi-Site with ISN

During design workshop, pros and cons of both solution were discussed and reviewed and are summarized in the table below:

Table 2. Back-to-Back and ISN Comparison

Feature	Back-to-Back	ISN	Comments
Fault domains separation	+		In case of ISN (Core) inter-DC traffic and inter-campus traffic follows via Core.
Flexibility		++	Theoretically, L3/L2 VNIs can be stretched between DCs and campuses
Scalability		+	Other DCs can be easily added
Deployment	=	=	
Operations / Automation	=	=	
Price	=	=	

After careful consideration it was decided to proceed with ISN design. Each datacentre will have one Cisco Nexus 9504 Core Switch. Hardware redundancy will be achieved by redundant Supervisor Modules on each switch. Network redundancy will be achieved via dual-homing each Border Gateway Leaf with both Core Switches. Details about Cisco Nexus 9504 switches and connectivity between Border Gateway Leaves and Core Switches can be found in following sections.

4.3. Spine Switches and Core Switches

Spine switches in a BGP EVPN VXLAN fabric act as the connecting nodes between all the Leaves or VTEPs. They form the backbone of the EVPN VXLAN network and forward traffic between the Leaves. Each leaf switch is connected to each spine switch in the network. Spine switches enable redundancy within the network and provide multiple paths for VTEPs to forward traffic to each other.

Spine switches in an EVPN VXLAN network are part of the underlay network and transport the VXLAN-encapsulated packets.

Modular Nexus 9504 Series Switches operating in NX-OS mode were chosen as Fabric Spine switches for BGP EVPN VXLAN fabric deployment at CompanyA of Luxembourg.

Same model has been chosen to play role of Core Switch. Details about interconnection of Core Switches with Border Gateway Switches is covered in Leaf Chapter below.

4.3.1 Nexus 9500 Series

Cisco Nexus 9500 Series switches provide the capability to use foundational layer 2/3 technologies, as well as modern technologies such as VXLAN, with a Border Gateway Protocol–Ethernet VPN (BGP-EVPN) control plane, Segment routing, Multiprotocol Label Switching (MPLS), and automation via NX-APIs.

The Cisco Nexus 9500 Series modular switches support a comprehensive selection of cloud-scale line cards and fabric modules that provide 1-, 10-, 25-, 40-, 50-, 100-, 200-, and 400-Gigabit Ethernet interfaces. A cloud scale line card provides up to 6.4 Terabits per second (Tbps) per slot and each cloud scale fabric module provides up to 1.6 Tbps to each line card slot. Each Cisco Nexus 9500 Series Chassis supports up to six fabric modules, which plug in vertically at the back of the chassis behind the fan trays.

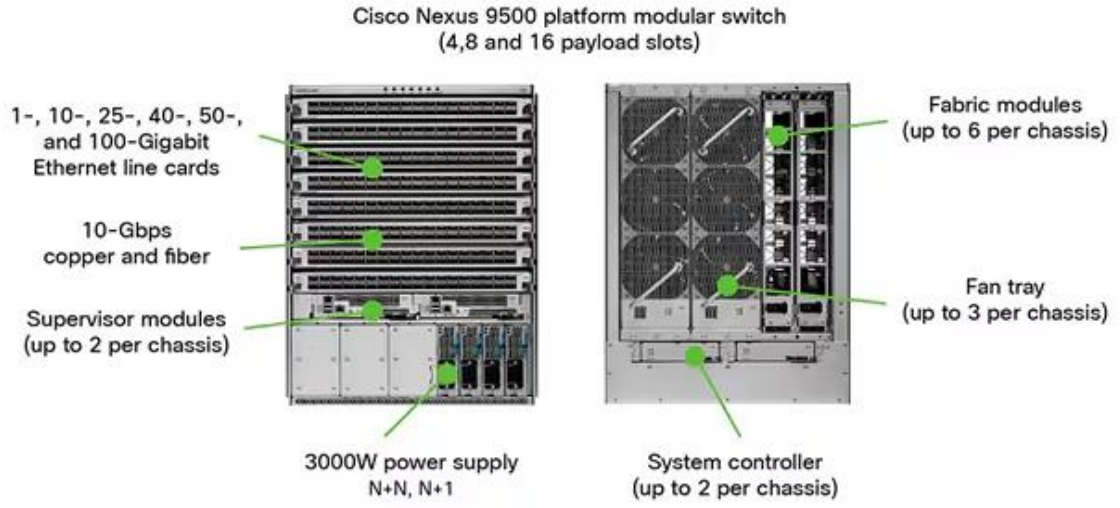

Figure 3. Cisco Nexus 9504 Physical Layout

Each Cisco Nexus 9500 Series chassis in CompanyA of Luxembourg deployment will be equipped with 4x N9K-X9736C-FX 100-Gigabit Ethernet Line Cards.

Figure 4. Cisco N9K-X9736C-FX Line Card

Those linecards have following characteristics:

- 36-port 100-Gigabit Ethernet Quad Small Form-Factor Pluggable 28 (QSFP28) line card

- Every port is 1x100-, 2x50-, 1x40-, 4x25-, 4x10-, and 1x10-Gigabit Ethernet breakout capable
- Ports 1 – 28 support 1 Gigabit Ethernet

Table 3. N9K-X9736C-FX 100-Gigabit Ethernet Line Card Performance and Scale

Parameter	N9K-X9736C-FX
Throughput	Up to 6.4Tbps when deployed with 4 Fabric Modules
Packet Buffer size	160MB
Number of Longest Prefix Match (LPM) route entries	IPv4: Up to 2 million, IPv6: 1,900 to 1 million (based on prefix length)
Number of IP host entries	IPv4: up to 2 million, IPv6: up to 32,000
Number of MAC address entries	Up to 512,000
Number of multicast routes	Up to 131,000
Number of Interior Gateway Management Protocol (IGMP) snooping groups	With VPC: 4000 to 32,000, Without VPC: 8000 to 32,000
Number of VLANs	Up to 4096
Number of VRF instances	Up to 16,000
Number of port channels	Up to 512
Number of port channel links	Up to 32
Number of Equal Cost Multipath (ECMP) paths	Up to 64
Number of active SPAN sessions	4 to 32
Number of Multiple Spanning Tree (MST) instances	Up to 64
Number of Rapid per-VLAN Spanning Tree (RPVST) instances	Up to 4000
Number of Hot Standby Router Protocol (HSRP) groups	Up to 490
Number of tunnel endpoints (VTEP) and VXLAN physical servers per VLAN	Up to 10,000
Weight	6.5 kg
Typical power	607 W
Maximum power	900 W
MTBF hours	420,050
Hot swappable	Yes
Airflow	Port-side intake

Apart from Linecard, integral part of Cisco Nexus 9504 chassis is a Fabric Module.

Figure 5. Cisco N9K-C9504-FM-G Fabric Module

Table 4. Cisco Nexus 9504 Switch supports fabric module N9K-C9504-FM-G with following characteristics

Parameter	N9K-C9504-FM-G
Total capacity	6.4 Tbps
Capacity (per slot)	1.6 Tbps
Maximum flow size	400 Gbps
Weight	3.4 kg
Typical power	455 W
Maximum power	504 W
MTBF hours	551,390
Hot swappable	Yes
Airflow	Port-side intake

A pair of redundant supervisor modules manages all switch operations using a state-synchronized, active-standby model. The supervisor accepts an external clock and supports management through multiple ports – two USB ports, a serial port, and a 10/100/1000-Mbps Ethernet port. All supervisors support Cisco ACI or NX-OS deployments. Redundant supervisors should be of the same type within a chassis.

Figure 6. Cisco N9K-SUP-A+ Supervisor Module

Table 5. Cisco N9K-SUP-A+ supervisor module specification

Parameter	N9K-SUP-A+
Processor	4 core, 8 thread 1.8GHz x86
DRAM	16GB
SSD	64GB
Weight	2.37 kg
Typical power	69 W
Maximum power	80 W
MTBF hours	414,240
Hot swappable	Yes
Airflow	Port-side intake

4.4. Leaves

Leaves are the nodes that are connected to the host or access devices. As a leaf switch sits on the edge of the network, it is also called as an edge or Network Virtualization Edge (NVE). When a host device on one leaf switch tries to communicate with a host device on another leaf switch, the traffic between the Leaves is sent through a spine switch. Leaves function as VTEPs in a VXLAN network and perform the encapsulation and decapsulation.

In BGP EVPN VXLAN fabric solution for CompanyA of Luxembourg multiple different Cisco Nexus 9300 Series switches models will be used.

- N9K-C9336C-FX2 - Leaf switch to provide 40G connectivity to hosts.
- N9K-C93180YC-FX3 - Leaf switch to provide 10G connectivity to hosts. Also, these switches will be used for Border Leaf and Border Gateway roles.
- N9K-C93180YC-FX - Currently used as ToR switch and will be converted to Leaf Switch.
- N9K-C93180YC-EX - Currently used as ToR switch and will be converted to Leaf Switch.
- N9K-C93108TC-EX - Currently used as ToR switch and will be converted to Leaf Switch.

4.4.1 N9K-C9336C-FX2

4.4.1.1 Overview

The Cisco Nexus 9300-FX2 Series switches belong to the fixed Cisco Nexus 9000 platform based on Cisco Cloud Scale technology. The platform supports cost-effective cloud-scale deployments, an increased number of endpoints, and cloud services with wire-rate security and telemetry. The platform is built on modern system architecture designed to provide high performance and meet the evolving needs of highly scalable datacentres and growing enterprises.

Cisco Nexus 9300-FX2 series is an extension of Nexus 9300-FX series switches with higher bandwidth capacity. The switches offer a variety of interface options to transparently migrate existing datacentres from 1-Gbps, and 10-Gbps speeds to 25-Gbps at the server, and from 10- and 40-Gbps speeds to 50- and 100-Gbps at the aggregation layer. The platforms provide investment protection for customers, delivering large buffers, highly flexible Layer 2 and Layer 3 scalability, and performance to meet the changing needs of virtualized datacentres and automated cloud environments.

In CompanyA of Luxembourg deployment N9K-C9336C-FX2 Switches will be used as Leaves to connect 40G endpoints.

Figure 7. Cisco N9K-C9336C-FX2 Leaf Switch

Table 6. Cisco Nexus N9K-C9336C-FX2 switch specification

Feature	N9K-C9336C-FX2
Ports	36 x 40/100-Gbps QSFP28 ports
Supported speeds	1/10/25/40/100-Gbps Ethernet; Breakout supported on all ports, 1-36: 100G, 2x50G NRZ, 40G native, 4x10/25G (10G w/QSA) 1G w/QSA except ports 1-6 and 33-36
CPU	4 cores
System memory	24 GB
SSD drive	128 GB
System buffer	40 MB
Management ports	2 ports: 1 RJ-45 and 1 SFP+
Power supplies (up to 2)	750W AC, 1100W AC, 1100W DC, 1100W HVAC/HVDC
Typical power (AC)	337W
Maximum power (AC)	719W
Input voltage (AC)	100 to 240V
Input voltage (High-Voltage AC [HVAC])	100 to 277V
Input voltage (DC)	–40 to –72V
Input voltage (High-Voltage DC [HVDC])	–240 to –380V
Frequency (AC)	50 to 60 Hz
Fans	3 dual fan trays
Airflow	Port-side intake and exhaust
MTBF	352,590 hours
Minimum NX-OS image	NXOS-703I7.3

Table 7. Cisco Nexus N9K-C9336C-FX2 switch Hardware performance and scalability specifications

Feature	N9K-C9336C-FX2
Maximum number of IPv4 Longest Prefix Match (LPM) routes	896,000
Maximum number of IPv4 host entries	896,000
Maximum number of IPv6 Longest Prefix Match (LPM) routes	498,000
Maximum number of IPv6 host entries	896,000
Maximum number of MAC address entries	256,000
Maximum number of multicast routes	128,000
Number of Internet Group Management Protocol (IGMP) snooping groups	Shipping: 8,000, Maximum: 32,000

Maximum number of Access Control List (ACL) entries	Per slice of the forwarding engine: 5000 ingress, 2000 egress
Maximum number of VLANs	4096
Number of Virtual Routing and Forwarding (VRF) instances	Shipping: 1,000, Maximum: 16,000
Maximum number of ECMP paths	64
Maximum number of port channels	512
Maximum number of links in a port channel	32
Number of active SPAN sessions	4
Maximum number of VLAN's in Rapid per-VLAN Spanning Tree (RPVST) instances	3,967
Maximum number of Hot-Standby Router Protocol (HSRP) groups	490
Number of Network Address Translation (NAT) entries	1,023
Maximum number of Multiple Spanning Tree (MST) instances	64
Flow-table size used for Cisco Tetration Analytics platform	64,000
Number of Queues	8

4.4.1.2 Connection to Spines

Last 2 interfaces on each Leaf switch will be used for connection to local Spines. In CompanyA of Luxembourg N9K-C9336C-FX2 Leaves are intended to connect 40G hosts, therefore they will be connected with 100G Uplinks towards each Spine at site.

Transceivers

Transceivers QSFP-100G-SR4-S will be used to connect N9K-C9336C-FX2 to the local Spine Switches.

The Cisco 100GBASE-SR4-S QSFP Module supports link lengths of up to 70m over OM3 and 100m over OM4 multi-mode fiber with MPO connectors. It primarily enables high-bandwidth 100G optical links over 12-fiber parallel fiber terminated with MPO multi-fiber connectors. QSFP-100G-SR4-S supports 100GBase Ethernet rate.

Figure 8. Cisco QSFP-100G-SR4-S 100G Transceiver

Existing Transceivers will be re-used to connect endpoints with SFP interfaces.

Transceivers QSFP-40G-SR4-S will be used on downlinks towards 40G endpoints.

The S-Class Cisco 40GBASE-SR4-S QSFP module supports link lengths of 100 and 150 meters, respectively, on laser-optimized OM3, and OM4/OM5 multi-mode fibers. QSFP-40G-SR4-S is aligned to IEEE 40GBASE-SR4 optical specifications which support high-bandwidth 40G optical links over 12-fiber parallel fiber terminated with MPO/MTP multi-fiber female connectors. The QSFP-40G-SR4-S does not support 4x10G breakout connectivity. QSFP-40G-SR4-S does not support FCoE.

Figure 9. Cisco QSFP-40G-SR4-S Transceiver

4.4.2 N9K-C93180YC-FX3

4.4.2.1 Overview

The Cisco Nexus 93180YC-FX3 Switch is a 1RU switch that supports 3.6 Tbps of bandwidth and 1.2 Bpps. The 48 downlink ports on the 93180YC-FX3 are capable of supporting 1-, 10-, or 25-Gbps Ethernet, offering deployment flexibility and

investment protection. The 6 uplink ports can be configured as 40 or 100-Gbps Ethernet, offering flexible migration options. The Cisco Nexus 93180YC-FX3 switch supports standard PTP telecom profiles with SyncE and PTP boundary clock functionality for telco datacentre edge environments.

In CompanyA of Luxembourg deployment N9K-C93180YC-FX3 Switches will be used as Border Leaf and Border Gateway Switches for external connectivity of datacentres and for interconnections between datacentres itself and campuses. This model of switches will also be used as a regular Leaf switch to connect endpoints.

Figure 10. Cisco N9K-C93180YC-FX3 Leaf Switch

Table 8. Cisco Nexus N9K-C93180YC-FX3 switch specification

Feature	N9K-C93180YC-FX3
Ports	Downlinks: 48 x 1/10/25G SFP28 ports; Uplinks: 6 x 40/100G QSFP28 ports
CPU	4 cores
System memory	Default: 16GB, Expandable: 16GB
SSD drive	128 GB
System buffer	40 MB
Management ports	1 port: 1 RJ-45
Power supplies (up to 2)	650W AC port-side intake and port-side exhaust, 930W DC port-side intake and port-side exhaust, 1200W HVAC/HVDC dual direction
Typical power (AC)	325W
Maximum power (AC)	600W
Input voltage (AC)	100 to 240V
Input voltage (High-Voltage AC [HVAC])	200 to 277V
Input voltage (DC)	–48 to –60V
Input voltage (High-Voltage DC [HVDC])	–240 to –380V
Frequency (AC)	50 to 60 Hz
Fans	4
Airflow	Port-side intake and exhaust
MTBF	288,760 hours
Minimum NX-OS image	NXOS-9.3.7

Table 9. Cisco Nexus N9K-C93180YC-FX3 switch Hardware performance and scalability specifications

Feature	N9K-C93180YC-FX3
Maximum number of IPv4 Longest Prefix Match (LPM) routes	1,792,000
Maximum number of IPv4 host entries	1,792,000
Maximum number of IPv6 Longest Prefix Match (LPM) routes	896,000
Maximum number of IPv6 host entries	1,792,000
Maximum number of MAC address entries4	512,000
Maximum number of multicast routes	128,000
Number of Internet Group Management Protocol (IGMP) snooping groups	32,000
Maximum number of Access Control List (ACL) entries	Per slice of the forwarding engine: 5000 ingress, 2000 egress
Maximum number of VLANs	4096
Number of Virtual Routing and Forwarding (VRF) instances	16,000
Maximum number of ECMP paths	128
Maximum number of port channels	512
Maximum number of links in a port channel	32
Number of active SPAN sessions	4
Maximum number of VLAN's in Rapid per-VLAN Spanning Tree (RPVST) instances	3,967
Maximum number of Hot-Standby Router Protocol (HSRP) groups	490
Number of Network Address Translation (NAT) entries	1,023
Maximum number of Multiple Spanning Tree (MST) instances	64
Flow-table size used for Cisco Tetration Analytics platform	64,000
Number of Queues	8

4.4.2.2 Connection to Spines, Core and External Networks

Last 2 interfaces on each Leaf switch will be used for connection to local Spines. In CompanyA of Luxembourg N9K-C93180YC-FX3 Leaves also play role of Border Gateway, therefore they will be connected with 100G Uplinks towards each Spine and Core switch at site with SFPs supported Multi-Mode cables and with 100G Uplinks towards remote Core router with SFPs supported Single-Mode cables. In addition to Border Gateway, same switches also act as Border Leaves providing connectivity to the external networks. Connection to external services is done via another set of 100G Uplinks towards Core Switches.

Transceivers

Transceivers QSFP-100G-SR4-S will be used to connect N9K-C93180YC-FX3 to the local Spine Switches.

The Cisco 100GBASE-SR4-S QSFP Module supports link lengths of up to 70m over OM3 and 100m over OM4 multi-mode fiber with MPO connectors. It primarily enables high-bandwidth 100G optical links over 12-fiber parallel fiber terminated with MPO multi-fiber connectors. QSFP-100G-SR4-S supports 100GBase Ethernet rate.

Figure 11. Cisco QSFP-100G-SR4-S 100G Transceiver

Some of devices of those models are located in S1R04 Rack and due to the high distance between this rack and Spine switches, SFPs QSFP-40G-LR4-S supported Single-Mode will be used for interconnection with spines.

The Cisco 40GBASE-LR4 QSFP module supports link lengths of up to 10 kilometer over a standard pair of G.652 single-mode fiber with duplex LC connectors. The QSFP-40G-LR4-S module supports 40GBASE Ethernet rate only. The 40 Gigabit Ethernet signal is carried over four wavelengths. Multiplexing and demultiplexing of the four wavelengths are managed in the device. QSFP-40G-LR4-S does not support FCoE.

Figure 12. Cisco QSFP-40G-LR4-S Transceiver

Transceivers QSFP-100G-LR4-S will be used to connect N9K-C93180YC-FX3 to the Core Switches.

The Cisco QSFP100 LR4 Module supports link lengths of up to 10km over a standard pair of G.652 single-mode fiber with duplex LC connectors. It complies with the IEEE 100GBASE-LR4 specification, which does not employ the use of FEC. QSFP-100G-LR4-S supports 100GBase Ethernet rate.

Figure 13. Cisco QSFP-100G-SR4-S 100G Transciever

4.4.3 N9K-C93180YC-FX

4.4.3.1 Overview

The Cisco Nexus 93180YC-FX Switch is a 1RU switch with latency of less than 1 microsecond that supports 3.6 Tbps of bandwidth and 1.2 bpps. The 48 downlink ports on the 93180YC-FX are capable of supporting 1-, 10-, or 25-Gbps Ethernet or as 16-, 32-Gbps Fibre Channel ports, creating a point of convergence for primary storage, compute servers, and back-end storage resources at the top of rack. The uplink can support up to six 40- and 100-Gbps ports, or a combination of 1-, 10-, 25-, 40, 50-, and 100-Gbps connectivity, offering flexible migration options. The

switch has IEEE compliant, FC-FEC and RS-FEC enabled for 25-Gbps support. All ports support wire-rate MACsec encryption.

In CompanyA of Luxembourg deployment N9K-C93180YC-FX Switches will be repurposed from current ToR role to a regular Leaf switch to connect endpoints.

Figure 14. Cisco N9K-C93180YC-FX Leaf Switch

Table 10. Cisco Nexus N9K-C93180YC-FX switch specification Table: Cisco Nexus N9K-C93180YC-FX switch specification

Feature	N9K-C93180YC-FX
Ports	48 x 1/10/25-Gbps fiber ports and 6 x 40/100-Gbps QSFP28 ports
CPU	6 cores
System memory	Up to 32 GB
SSD drive	128 GB
System buffer	40 MB
Management ports	1 port: 1 RJ-45
Power supplies (up to 2)	500W AC, 930W DC, or 1200W HVAC/HVDC
Typical power (AC)	260W
Maximum power (AC)	425W
Input voltage (AC)	100 to 240V
Input voltage (High-Voltage AC [HVAC])	200 to 277V
Input voltage (DC)	–48 to –60V
Input voltage (High-Voltage DC [HVDC])	–240 to –380V
Frequency (AC)	50 to 60 Hz
Fans	4
Airflow	Port-side intake and exhaust
MTBF	238,470 hours
Minimum NX-OS image	NXOS-703I7.1

Table 11. Cisco Nexus N9K-C93180YC-FX switch Hardware performance and scalability specifications

Feature	N9K-C93180YC-FX
Maximum number of IPv4 Longest Prefix Match (LPM) routes	1,792,000
Maximum number of IPv4 host entries	1,792,000

Maximum number of IPv6 Longest Prefix Match (LPM) routes	896,000
Maximum number of IPv6 host entries	1,792,000
Maximum number of MAC address entries4	512,000
Maximum number of multicast routes	128,000
Number of Internet Group Management Protocol (IGMP) snooping groups	Shipping: 8,000, Maximum: 32,000
Maximum number of Access Control List (ACL) entries	Per slice of the forwarding engine: 5000 ingress, 2000 egress
Maximum number of VLANs	4096
Number of Virtual Routing and Forwarding (VRF) instances	Shipping: 1,000, Maximum: 16,000
Maximum number of ECMP paths	64
Maximum number of port channels	512
Maximum number of links in a port channel	32
Number of active SPAN sessions	4
Maximum number of VLAN's in Rapid per-VLAN Spanning Tree (RPVST) instances	3,967
Maximum number of Hot-Standby Router Protocol (HSRP) groups	490
Number of Network Address Translation (NAT) entries	1,023
Maximum number of Multiple Spanning Tree (MST) instances	64
Number of Queues	8

4.4.3.2 Connection to Spines

Last 2 interfaces on each Leaf switch will be used for connection to local Spines. In CompanyA of Luxembourg N9K-C93180YC-FX switches play role of regular Leaves, therefore they will be connected with 40G Uplinks towards each Spine at site.

Transceivers

Transceivers QSFP-40G-SR4-S will be used on Uplinks towards Spines.

The S-Class Cisco 40GBASE-SR4-S QSFP module supports link lengths of 100 and 150 meters, respectively, on laser-optimized OM3, and OM4/OM5 multi-mode fibers. QSFP-40G-SR4-S is aligned to IEEE 40GBASE-SR4 optical specifications which support high-bandwidth 40G optical links over 12-fiber parallel fiber terminated with MPO/MTP multi-fiber female connectors. The QSFP-40G-SR4-S does not support 4x10G breakout connectivity. QSFP-40G-SR4-S does not support FCoE.

Figure 15. Cisco QSFP-40G-SR4-S Transceiver

Existing Transceivers will be re-used to connect endpoints with SFP interfaces. Transceivers SFP-10G-T-X will be used on downlinks to connect endpoints with RJ-45 interfaces.

The Cisco 10GBASE-T module offers connectivity options at the following data rates: 100M/1G/10Gbps. It has the SFP+ form factor and an RJ-45 interface so that CAT5e/CAT6A/CAT7 cables can be used to connect to end points with embedded 10GBASE-T ports. They are suitable for distances up to 30 meters and offers a cost-effective way to connect within racks and across adjacent racks.

Figure 16. Cisco SFP-10G-T-X Transceiver

4.4.4 N9K-C93180YC-EX and N9K-C93108TC-EX

4.4.4.1 Overview

The Cisco Nexus 93180YC-EX Switch is a 1-Rack-Unit (1RU) switch with latency of less than 1 microsecond that supports 3.6 Terabits per second (Tbps) of bandwidth and over 2.6 billion packets per second (bpps). The 48 downlink ports on the 93180YC-EX can be configured to work as 1-, 10-, or 25-Gbps ports, offering deployment flexibility and investment protection. The uplink can support up to six 40- and 100-Gbps ports, or a combination of 1-, 10-, 25-, 40-, 50, and 100-Gbps connectivity, offering flexible migration options. The switch has FC-FEC enabled for 25Gbps, and supports up to 3m in DAC connectivity.

In CompanyA of Luxembourg deployment N9K-C93180YC-EX Switches will be repurposed from current ToR role to a regular Leaf switch to connect endpoints with SFP ports.

Figure 17. Cisco N9K-C93180YC-EX Leaf Switch

The Cisco Nexus 93108TC-EX Switch is a 1RU switch that supports 2.16 Tbps of bandwidth and over 1.6bpps. The 48 10GBASE-T downlink ports on the 93108TC-EX can be configured to work as 100-Mbps, 1 Gbps, or 10-Gbps ports. The uplink can support up to six 40- and 100-Gbps ports, or a combination of 1-, 10-, 25-, 40, 50-, and 100-Gbps connectivity, offering flexible migration options.

In CompanyA of Luxembourg deployment N9K-C93108TC-EX Switches will be repurposed from current ToR role to a regular Leaf switch to connect endpoints with RJ-45 ports.

Figure 18. Cisco N9K-C93108TC-EX Leaf Switch

Table 12. Cisco Nexus N9K-C93180YC-EX and N9K-C93108TC-EX switches specifications

Feature	N9K-C93180YC-EX	N9K-C93108TC-EX
Ports	48 x 1/10/25-Gbps and 6 x 40/100-Gbps QSFP28 ports	48 x 100-Mbps/1/10GBASE-T and 6 x 40/100-Gbps QSFP28 ports
CPU	4 cores	4 cores
System memory	24 GB	24 GB
SSD drive	64 GB	64 GB
System buffer	40 MB	40 MB
Management ports	2 ports: 1 RJ-45 and 1 SFP	2 ports: 1 RJ-45 and 1 SFP

Power supplies (up to 2)	500W AC, 650W AC, 930W DC, or 1200W HVAC/HVDC	500W AC, 650W AC, 930W DC, or 1200W HVAC/HVDC
Typical power (AC)	210W	210W
Maximum power (AC)	470W	470W
Input voltage (AC)	100 to 240V	100 to 240V
Input voltage (High-Voltage AC [HVAC])	200 to 277V	200 to 277V
Input voltage (DC)	–48 to –60V	–48 to –60V
Input voltage (High-Voltage DC [HVDC])	–240 to –380V	–240 to –380V
Frequency (AC)	50 to 60 Hz	50 to 60 Hz
Fans	4	4
Airflow	Port-side intake and exhaust	Port-side intake and exhaust
MTBF	390,330 hours	366,130 hours
Minimum NX-OS image	NXOS-703I4.2	NXOS-703I4.3

Table 13. Cisco Nexus N9K-C93180YC-EX switch Hardware performance and scalability

Feature	**N9K-C93180YC-EX and N9K-C93108TC-EX**
Maximum number of Longest Prefix Match (LPM) routes	896,000
Maximum number of host entries	896,000
Maximum number of MAC address entries4	512,000
Maximum number of multicast routes	32,000
Number of Internet Group Management Protocol (IGMP) snooping groups	Shipping: 8,000, Maximum: 32,000
Maximum number of Access Control List (ACL) entries	Per slice of the forwarding engine: 4000 ingress, 2000 egress
Maximum number of VLANs	4096
Number of Virtual Routing and Forwarding (VRF) instances	Shipping: 1,000, Maximum: 16,000
Maximum number of ECMP paths	64
Maximum number of port channels	512
Maximum number of links in a port channel	32
Number of active SPAN sessions	4
Maximum number of VLAN’s in Rapid per-VLAN Spanning Tree (RPVST) instances	3,967

Maximum number of Hot-Standby Router Protocol (HSRP) groups	490
Number of Network Address Translation (NAT) entries	1,023
Maximum number of Multiple Spanning Tree (MST) instances	64
Number of Queues	8

4.4.4.2 Connection to Spines

Last 2 interfaces on each Leaf switch will be used for connection to local Spines. In CompanyA of Luxembourg N9K-C93180YC-EX and N9K-C93108TC-EX switches play role of regular Leaves, therefore they will be connected with 40G Uplinks towards each Spine at site.

Transceivers

Transceivers QSFP-40G-SR4-S will be used on Uplinks towards Spines.

The S-Class Cisco 40GBASE-SR4-S QSFP module supports link lengths of 100 and 150 meters, respectively, on laser-optimized OM3, and OM4/OM5 multi-mode fibers. QSFP-40G-SR4-S is aligned to IEEE 40GBASE-SR4 optical specifications which support high-bandwidth 40G optical links over 12-fiber parallel fiber terminated with MPO/MTP multi-fiber female connectors. The QSFP-40G-SR4-S does not support 4x10G breakout connectivity. QSFP-40G-SR4-S does not support FCoE.

Figure 19. Cisco QSFP-40G-SR4-S Transceiver

Some of devices of those models are located in S1R04 Rack and due to the high distance between this rack and Spine switches, SFPs QSFP-40G-LR4-S supported Single-Mode will be used for interconnection with spines.

The Cisco 40GBASE-LR4 QSFP module supports link lengths of up to 10 kilometer over a standard pair of G.652 single-mode fiber with duplex LC connectors. The QSFP-40G-LR4-S module supports 40GBASE Ethernet rate only. The 40 Gigabit Ethernet signal is carried over four wavelengths. Multiplexing and demultiplexing of the four wavelengths are managed in the device. QSFP-40G-LR4-S does not support FCoE.

Figure 20. Cisco QSFP-40G-LR4-S Transceiver

Existing Transceivers will be re-used to connect endpoints with SFP interfaces. Transceivers SFP-10G-T-X will be used on downlinks to connect endpoints with RJ-45 interfaces to N9K-C93180YC-EX if needed.

The Cisco 10GBASE-T module offers connectivity options at the following data rates: 100M/1G/10Gbps. It has the SFP+ form factor and an RJ-45 interface so that CAT5e/CAT6A/CAT7 cables can be used to connect to end points with embedded 10GBASE-T ports. They are suitable for distances up to 30 meters and offers a cost-effective way to connect within racks and across adjacent racks.

Figure 21. Cisco SFP-10G-T-X Transceiver

4.5. Nexus Dashboard

Cisco Nexus Dashboard provides multiple services that bring the one of the best of cloud-operational models to your networks, whether Cisco® Application Centric Infrastructure (Cisco ACI®), Cisco NX-OS (through the Cisco Nexus Dashboard

Fabric Controller service and/or running in standalone mode), or a Cisco Cloud Network Controller running in a public cloud provider environment. Cisco Nexus Dashboard Fabric Controller (NDFC) consolidates management for multiple NX-OS–based switches, bringing automation and monitoring for LAN, EVPN VXLAN, and SAN fabrics.

Nexus Dashboard will be used in CompanyA of Luxembourg for initial greenfield BGP EVPN VXLAN fabric deployment. CompanyA of Luxembourg has no intention to use Nexus Dashboard for operational support of the fabric after the deployment giving preference to already existing in the CompanyA environment open-source orchestration and automation tools such as Ansible.

Nexus Dashboard will be deployed in a form of single VM in the legacy network of CompanyA of Luxembourg.

When deploying in VMware ESX, Nexus Dashboard can be deployed in two types of nodes:

- Data Node — node profile with higher system requirements designed for specific services that require the additional resources.
- App Node — node profile with a smaller resource footprint that can be used for most services.

For deployment of fabric in CompanyA of Luxembourg, node of “App Node” type would be sufficient.

Below are resource requirements for Nexus Dashboard VM.

Table 14. Nexus Dashboard VM Requirements

Nexus Dashboard Version	Requirements
3.1	VMware ESXi 7.0, 7.0.1, 7.0.2, 7.0.3, 8.0 VMware vCenter 7.0.1, 7.0.2, 7.0.3, 8.0 if deploying using vCenter 16 vCPUs with physical reservation of at least 2.2GHz 64GB of RAM with physical reservation 550GB SSD/NVMe

4.5.1 Out-of-Band Management Network

In order to ease automation of BGP EVPN VXLAN fabric deployment, it is recommended to connect all new fabric devices via Management interface to the existing OOB Management network. Nexus Dashboard VM should also have one interface connected to this network. OOB Management network will be used to discover fabric devices by Nexus Dashboard and to push required fabric configuration.

In case if OOB Management network is not available for initial deployment, it is possible to use In-Band management network over legacy infrastructure of CompanyA of Luxembourg. In that case management interfaces of new fabric devices will be connected to existing In-Band management network. It is important to note that once fabric devices are discovered by Nexus Dashboard and BGP EVPN VXLAN fabric configuration is pushed to the devices, it becomes labor-intensive task to re-ip management interfaces. It can be done at the very final stage before decommissioning of the Nexus Dashboard.

Deployment of VM and provision of necessary connectivity from Nexus Dashboard to new devices is done by CompanyA of Luxembourg.

4.6. Connectivity Matrix

Full connectivity matrix can be found in the “Connectivity Matrix.xls” spreadsheet that comes along with this design document.

5. Naming Convention

The full list of hostnames will be provided in a separate document. We are discussing only convention here. There are a few examples of names provided below.

6. Logical EVPN VXLAN Fabric Design

6.1. Design Key Points

In this section, we provide an overview of the key design points that will be further elaborated below in the chapter.

6.1.1 Multi-Site Design

- Site-external design: Multi-Site with “BGW between spine and superspine” topology
- BGWs
  - Number of BGWs per site: 2
  - Redundancy mode: anycast BGW
- Inter-Site Network:
  - Number of superspines per site: 1
- Border Leaves (BL):
  - BGWs also play a role of BL
  - L3EXT: superspines play also a role of core devices providing L3 external connectivity

6.1.2 Control Plane

- I-E-I option for underlay and overlay control plane will be implemented that means the following:
  - Site-internal: IGP (IS-IS) as underlay CP, iBGP as overlay CP
  - Site-external: eBGP for both underlay and overlay CPs
- Site-internal Underlay
  - Routing: IS-IS with authentication, BFD
  - BUM traffic transmission: PIM-ASM multicasting with 2 RPs per site (on Spines)
  - Each compute leaf will use a common set of L2 and L3 VNIs. For these L2VNIs, a mapping to a single multicast group will be configured. However, for service leaves, the L2 VNIs responsible for connecting with L3-L7 devices will have dedicated multicast group(s) configured.
- Site-internal Overlay
  - Routing: EVPN iBGP with authentication, 2 RRs per site (on Spines), BFD, ECMP
- Site-external Underlay
  - Routing: IPv4 eBGP with authentication, BFD, ECMP
  - BUM traffic transmission: ingress replication
- Site-external Overlay
  - Routing: EVPN eBGP with authentication, 2 Route Servers (Core), BFD
- Overlay L2 Loops Prevention

6.1.3 Overlay Data Plane

- Multi-tenancy
  - 25 VRFs
- IPv6 Capability
- Multicast
  - Layer 3 Mode TRM
  - Internal Anycast RP approach
- DHCP relay (Option 82)

6.1.4 Management

- OOB management on deploying stage
- In-band management
  - Loopbacks in dedicated VRF SwitchMGMT

6.1.5 VXLAN BGP EVPN Enhancements

- ARP Suppression
- Anycast Gateway
- VPC Fabric Peering
- NGOAM Loop Detection and Mitigation

6.1.6 Cisco Best Practices

Deployment will be implemented with Nexus Dashboard Fabric Controller (NDFC)

6.2. VXLAN

Virtual Extensible LAN (VXLAN) is a central element in the data plane functionality of BGP EVPN VXLAN fabric. VXLAN, as outlined in RFC 7348, serves as an overlay technology designed to deliver Layer 2 and Layer 3 connectivity services over IP network. The mechanism involves encapsulating Layer 2 frames within IP packets. Notably, VXLAN relies on IP reachability between the VXLAN edge devices, which is established through an IP underlay routing protocol.

VXLAN is point-to-multipoint tunnelling mechanism to extend Layer 2 networks over an IP network.

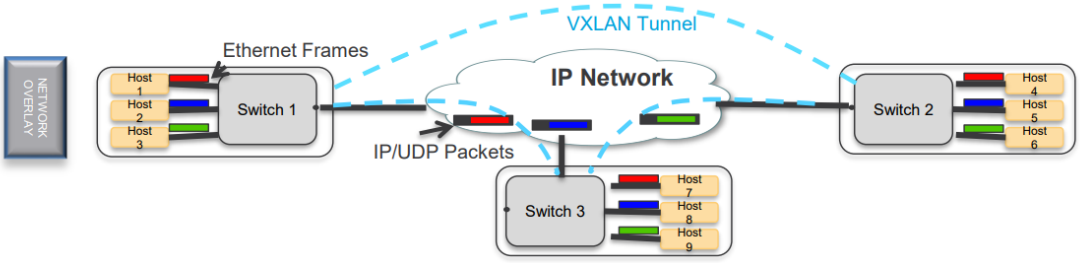

Figure 22. VXLAN Tunnel

VXLAN uses MAC in UDP encapsulation (UDP destination port 4789)

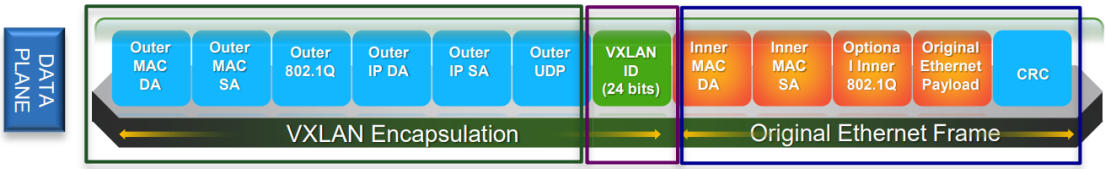

Figure 23. VXLAN Encapsulation

When discussing the essential components of a VXLAN fabric, the following terminology is commonly used:

- **VTEP** (Virtual Tunnel Endpoint): This hardware or software element resides at the network edge and is responsible for creating VXLAN tunnels, as well as handling VXLAN encapsulation and decapsulation.
- **VNI** (Virtual Network Instance): A logical network instance that provides either Layer 2 or Layer 3 services, defining a Layer 2 broadcast domain.
- **VNID** (Virtual Network Identifier): A 24-bit segment ID that enables addressing for up to 16 million logical networks within the same administrative domain.
- **Bridge Domain**: A collection of logical or physical ports participating in the flooding or broadcast packets exchange.

6.3. Underlay Network

6.3.1 Site-Internal

VXLAN relies on an underlying transport network for data plane forwarding. When we use the fabric as the underlay infrastructure, it follows a spine-leaf approach where each Leaf switch connects to all spines.

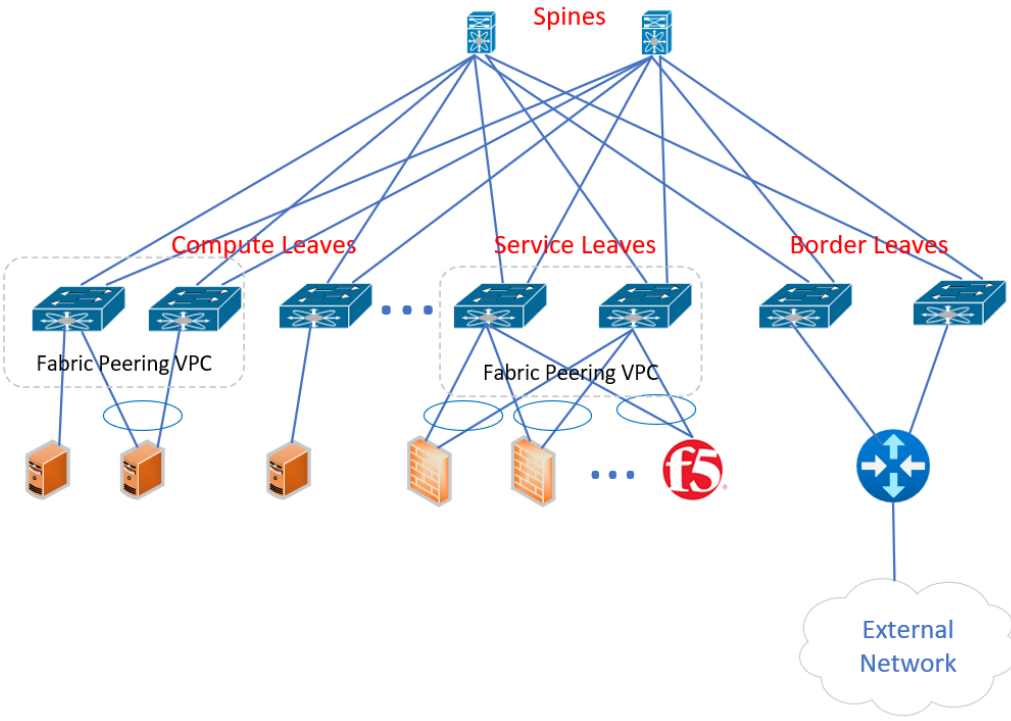

Figure 24. Site-Internal Fabric

In the fabric topology, end devices are typically connected to the Leaf switches. From a VXLAN perspective, these switches play the role of Virtual Tunnel Endpoints (VTEPs). Leaf switches can also be logically subdivided into roles such as service, compute, and border Leaf. While this division is considered best practice, it's not mandatory and depends on resource capacities and requirements.

Each VTEP is represented by a loopback IP address. In our implementation dedicated loopback interface will be configured to provide the VTEP IP address for each Leaf switches. These loopback IPs are used to establish VXLAN tunnels between VTEPs. The introduction of Virtual Port Channels (VPCs) adds complexity: the VPC domain presents an anycast VTEP, which manages dual-homed hosts. Individual VTEPs continue to serve orphan hosts.

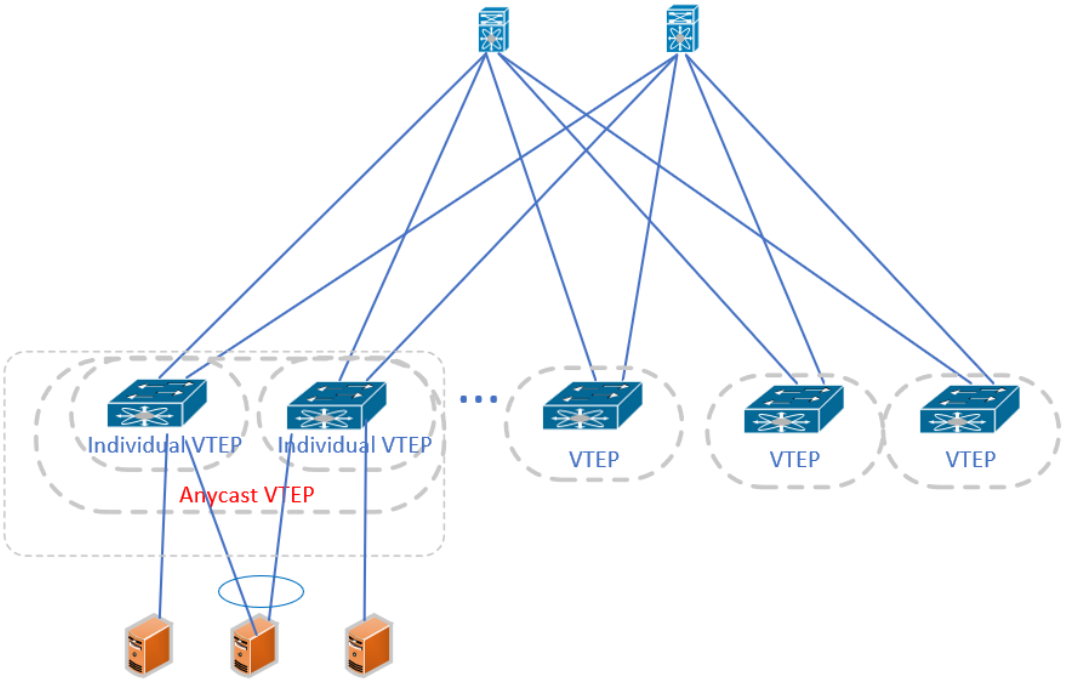

Figure 25. Site-Internal VTEPs

The anycast VTEP's IP address is configured as a secondary IP on this dedicated loopback interface (consistent across both VPC nodes in the domain).

6.3.2 Site-External

The central component of the EVPN Multi-Site architecture is the **border gateway** (BGW). BGWs serve to separate the site-internal fabric from the network that interconnects multiple sites. They effectively mask the site-internal VTEPs.

In a typical EVPN Multi-Site deployment, two or more sites are interconnected through a VXLAN BGP EVPN Layer 2 and Layer 3 overlay (as shown in Figure 4). Here's how it works:

- The BGW connects to the site-internal VTEPs (usually via spine nodes).
- It also connects to a site-external transport network, allowing traffic to reach BGWs at remote sites.
- The BGWs at remote sites have site-internal VTEPs behind them.

- Only the underlay IP addresses of the BGWs need to be visible within the transport network between BGWs. BGWs always mask the site-internal VTEPs.

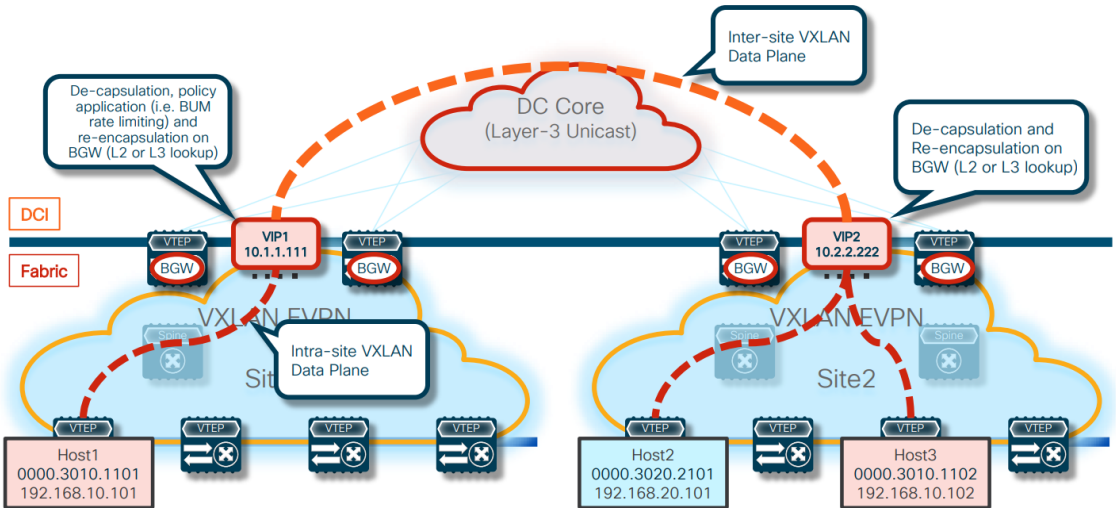

Figure 26. Site-External Example

BGWs can be deployed in various models:

- Border Gateway (Between Spine and External)
- Border Gateway Spine
- Border Gateway Super Spine

It can also be implemented as

- Single BGW
- VPC (Virtual Port Channel) pair
- Anycast BGWs (up to 6 devices per site)

Our implementation uses 2 Anycast Border Gateways (Leaf switches).

Within the fabric, BGWs establish VXLAN tunnels with other local VTEPs like ordinary Leaves with regular VTEPs.

To enable communication between sites, Border Gateway (BGW) devices establish additional VXLAN tunnels to remote BGWs. In this context, the BGW assumes the role of a Virtual Tunnel Endpoint (VTEP) within the Multi-Site VXLAN EVPN BGP fabric that interconnects remote sites. This ‘external’ VTEP is associated with a separate loopback interface. The IP address assigned to this BGW loopback interface is commonly referred to as the Virtual IP address (VIP) or anycast IP address. Conversely, the IP address used for regular internal VTEP communication is known as the Private IP Address (PIP) or individual IP address. The specific roles of these PIP and VIP addresses will be further discussed in the ‘Site-External Underlay’ section.

In our design, the switches configured as BGWs also serve as Border Leaves, providing connectivity to the Core.

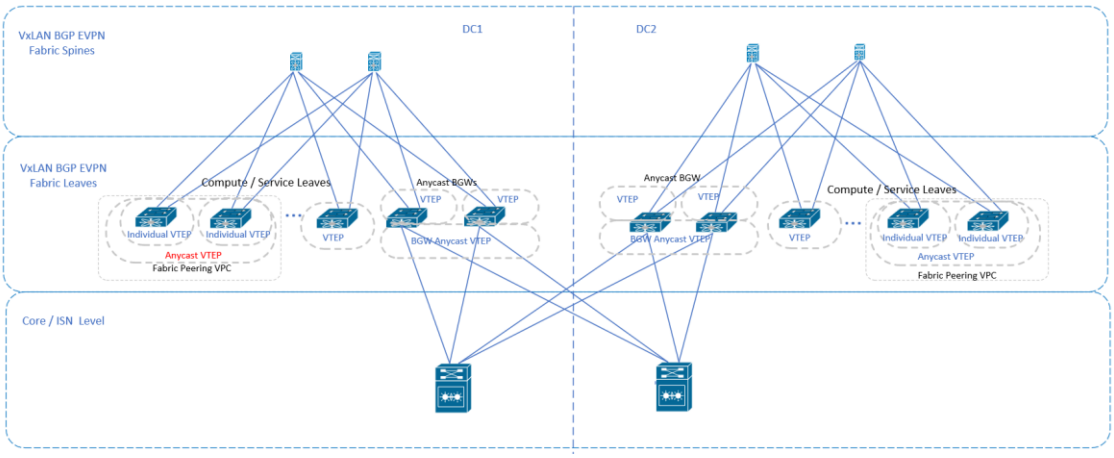

Figure 27. Multi-Site with Two Sites

6.4. Control Plane

6.4.1 Site-Internal Underlay CP

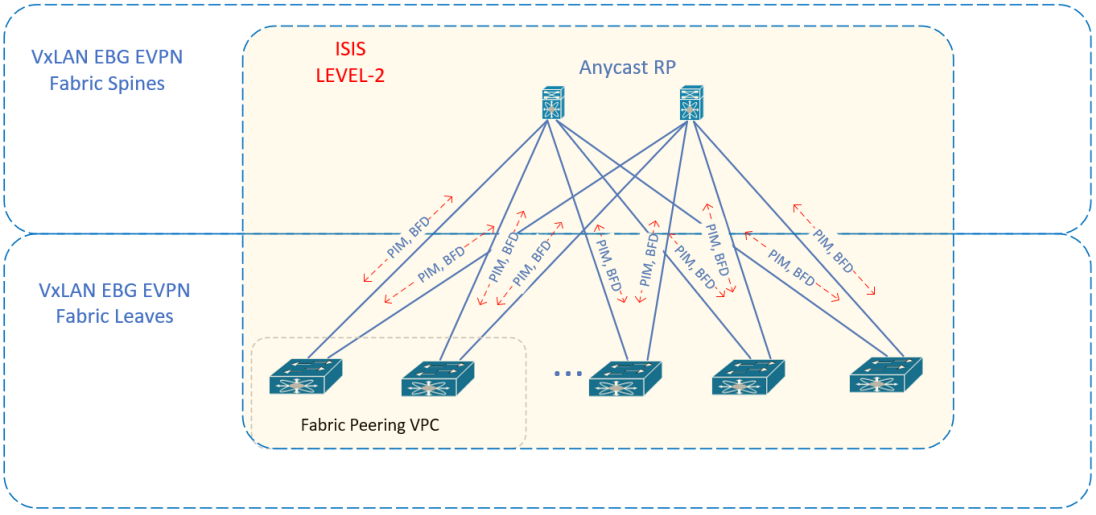

Figure 28. Site-Internal Underlay CP

BGP neighborship is established using loopback addresses. Underlay protocols (IGP like IS-IS, OSPF, or eBGP) provide reachability to these loopbacks. IGP is preferable for this purpose.

Multicasting is commonly used for handling Broadcast, Unknown Unicast, and Multicast (BUM) traffic. PIM-ASM or PIM BiDir are typical approaches.

Routed Interface Addressing

We will follow the recommended approach of using an unnumbered IP addresses to configure the interface IP addresses, requiring only one IP address per device, regardless of the number of internal fabric interfaces deployed.

Loopback Interface Addressing

- Each Leaf switch should have at least two loopback interfaces. The first loopback serves as the Router-ID (RID) and assigns an IP address to

unnumbered Layer 3 links. The second loopback represents the VTEP IP address used as the source and destination for VXLAN-encapsulated traffic.

- Spines also have a minimum of 2 loopback interfaces, with the second loopback providing multicast RP functionality.
- BGWs require a minimum of 3 loopbacks, with the third one related to the BGW role.

IGP Protocol Selection

IS-IS is chosen as underlay routing protocol. The reason this protocol is preferred over OSPF is because this protocol is already widely used in CompanyA network infrastructure. This link state routing protocol is gaining popularity with fast convergence in a large-scale environment. IS-IS uses Connectionless Network Protocol (CLNP) for communication between peers and doesn't depend on IP. There is no SPF calculation on link change and SPF calculation only happens when there is a topology change which helps with faster convergence and stability in the underlay. No significant tuning is required for IS-IS to achieve an efficient, fast converging underlay network.

BUM Traffic

When designing network functionality, it's essential to consider multi-destination traffic in addition to unicast traffic. This type of traffic is commonly referred to as **BUM**, which stands for Broadcast, Unknown Unicast, and Multicast traffic.

To enable the transmission of BUM traffic across the VXLAN fabric, two different approaches can be taken:

- **Leverage Multicast Technology**: In this approach, multicast technology is utilized in the underlay network. By doing so, the native replication capabilities of multicasting are harnessed to deliver traffic to all the edge VTEP devices efficiently.
- **Source Replication**: When multicast deployment is not feasible, an alternative method involves using source-replication capabilities within the VTEP nodes. These nodes create multiple unicast copies of the BUM frames, sending them to each remote VTEP device. However, it's important to note that this approach is not as efficient as using multicast for BUM traffic replication.

IP multicast is a recommended method of distribution of multi-destination traffic in the site-internal underlay. To deploy IP multicast in the underlay, a Protocol Independent Multicast (PIM) routing protocol needs to be enabled and must be consistent across all the devices in the underlay network. The two common PIM protocols are Sparse-Mode (PIM-ASM) and Bidirectional (PIM-Bidir). This implies the requirement to deploy rendezvous Points (RPs). PIM-ASM has been chosen for this design.

The ASM multicast solution improves convergence in an RP failure scenario. This is achieved by deploying Anycast RP, which consists of using a common IP address across devices to identify the RP. A simple static RP mapping configuration is then applied to each node in the fabric to associate multicast groups with the RP so that each source or destination can then use the local RP that is closest from a topological point of view. A common approach is to use Spine nodes as RPs.

The idea behind using multicast when delivering BUM packets is as follows.

In a VXLAN network, a single copy of BUM traffic originated from the ingress or source VTEP is directed toward the underlay transport network. The network then forwards this single copy along the multicast tree, so that it reaches all egress or destination VTEPs that participate in the given multicast group. As the single copy traverses the multicast tree, it is replicated only at appropriate branch points where receivers have joined the multicast group associated with the VNI. This approach ensures that only one copy per wire or link is maintained within the network, resulting in an efficient way to forward BUM traffic.

In this approach, each Layer 2 VNI is associated with a multicast group. This mapping should be consistently configured on all VTEPs where this VNI is present. Once the mapping is set up, the VTEP sends out a corresponding multicast join message expressing interest in the multicast tree associated with the specific multicast group.

When mapping a Layer 2 VNI to a multicast group, several options are available in the mapping provision for both the VNI and the multicast group.

On one end of the spectrum, each Layer 2 VNI can be mapped to a unique multicast group. On the other end, the simplest approach involves using a single multicast group and mapping all Layer 2 VNIs to that group. While this strategy reduces the multicast state in the underlying network, it may not be the most efficient way to handle BUM traffic. We will adopt the second approach for all Layer 2 VNIs that span across all Leaf nodes. In this scenario, it will be essential to replicate BUM packets to all those Leaf nodes. For L2 VNIs that are localized on only a subset of Leaves, a combination of both approaches appears reasonable. Further details will be covered in the NIP (Network Implementation Plan) and migration plan documentation.

6.4.2 Site-External Underlay CP

eBGP will be implemented for site-external Underlay CP.

For BUM replication between sites, EVPN Multi-Site architecture uses ingress replication to simplify the requirements of the site-external underlay network.

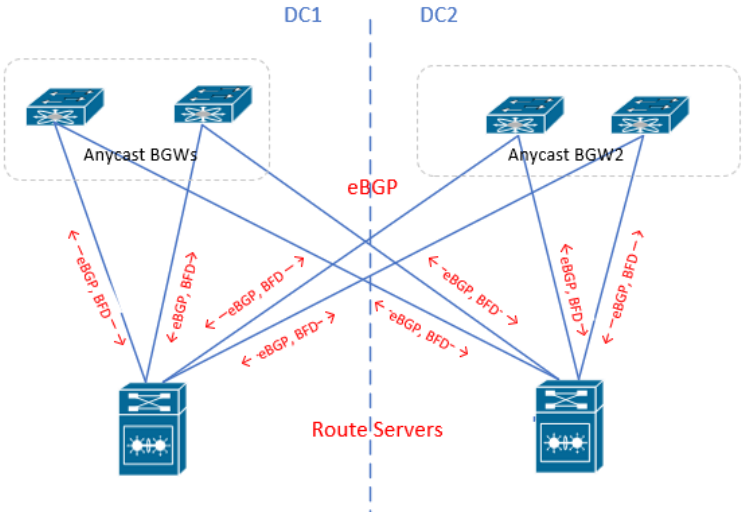

Figure 29. Site-External Underlay CP

PIP / VIP Roles

The Border Gateway (BGW) uses a virtual IP address for data plane communications, serving as a common address for traffic leaving a site and between sites when using the Multi-Site EVPN extension to connect to a remote site. This virtual IP address is also used within the local site to reach the egress point and facilitates communication between sites. The virtual IP address is associated with a dedicated loopback interface associated with the network virtualization endpoint (NVE) interface. With this approach, and with the existence of an Equal-Cost Multipath (ECMP) network, all BGWs are always equally reachable and active for data-traffic forwarding.

In addition to the virtual IP address or anycast IP address, every BGW has its own individual personality represented by the primary VTEP IP (PIP) address. The PIP address is responsible in the BGW for handling BUM traffic. Every BGW uses its PIP address to perform BUM replication, either in the multicast underlay or when advertising BGP EVPN Route Type 3 (inclusive multicast), used for ingress replication. Another case of using PIP is when BGWs play also a role of Border Leaves providing external connectivity with VRF-lite next to the EVPN Multi-Site deployment. In this case, routing prefixes that are learned from the external Layer 3 devices are advertised inside the VXLAN fabric with the PIP address as the next-hop address. A closely related scenario is the case in which the BGW advertises an IP prefix with its own PIP address through local connectivity. An endpoint can be directly connected to a BGW, but its IP address can be learned only through routing on a physical interface or sub-interface.

The use of VLANs and Switch Virtual Interfaces (SVIs) local to one BGW or across multiple BGWs is not currently supported.

The underlay CP for both fabrics in a Multi-Site deployment may be represented by the following diagram:

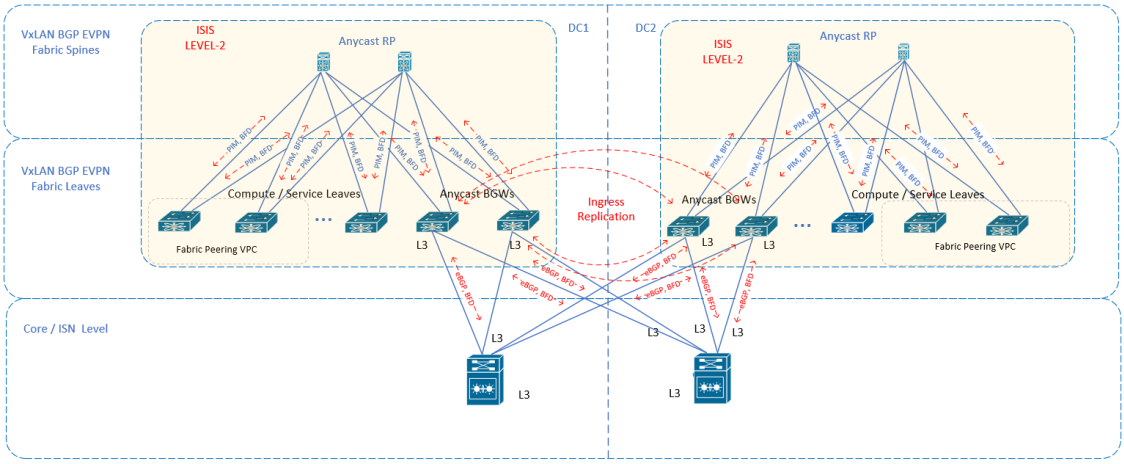

Figure 30. Multi-Site Underlay CP

6.4.3 Site-Internal Overlay CP

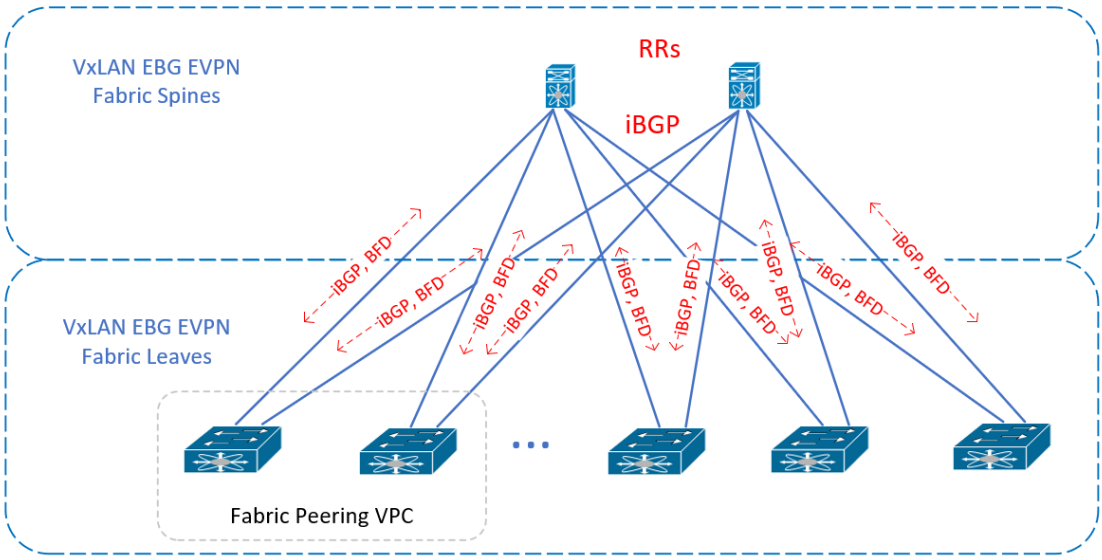

Figure 31. Site-Internal Overlay CP

The VXLAN protocol has primarily focused on the data plane, ensuring connectivity across hosts within a VXLAN domain. However, its original control plane, based on flood-and-learn behavior, is not scalable.

To address this, BGP EVPN is used as the standard control plane for VXLAN.

BGP EVPN has been proposed as the IETF standard control plane for VXLAN. It provides information about Layer 2 MAC routes, Layer 3 Host IP routes, and Layer 3 subnet IP routes. EVPN also introduces multi-tenancy support, VTEP peer discovery, security, and authentication mechanisms.

EVPN leverages constructs similar to the VPNv4 address family used in MPLS VPN architectures. These include Virtual Routing and Forwarding (VRFs), Route Distinguishers (RDs), and Route Targets (RTs). Notably, EVPN allows the exchange of both IP and MAC address information.

Virtual Routing and Forwarding (VRF)

Virtual Routing and Forwarding (VRF) defines the Layer 3 routing domain for each tenant supported in the VXLAN fabric. In VXLAN EVPN networks, each tenant VRF has a Layer 3 VNI used as a virtual backbone for routing within the VRF.

Route Distinguisher (RD)

Route Distinguisher (RD) is the identifier of a VRF since each VRF has its own unique RD in the network. When an EVPN advertises routes to the peers, the RD of the VRF to which this route belongs is prepended to the original route itself to render it unique within the network. This allows different VRFs to use overlapping IP addresses so that different tenants can have true autonomy for IP address management. The RD can be automatically defined to simplify configuration.

Route Target (RT)

Route Target (RT) is an extended attribute in EVPN route updates used to control route distribution in a multi-tenant network. EVPN VTEPs have an import RT setting

and an export RT setting for each VRF and each L2VNI. When a VTEP advertises EVPN routes, it affixes its export RT in the route update. The routes will be received by other VTEPs in the network. These devices will compare the RT value carried with the route against their own local import RT setting. If the two values match, the route will be accepted. Otherwise, the route will not be imported. The RT can be automatically defined to simplify configuration.

BFD

Bidirectional Forwarding Detection (BFD) is a detection protocol designed to provide fast forwarding path failure detection times for all media types, encapsulations, topologies, and routing protocols. The main benefit of implementing BFD for BGP is a marked decrease in convergence time.

EVPN Route Types

The EVPN control plane advertises different types of routing information:

- Type-2 - Endpoint reachability information, including MAC and IP addresses of the endpoints.
- Type-3 - Multicast route advertisement-announcing capability and intention to use Ingress Replication for specific VNIs.
- Type-5 - IP prefix route used to advertise internal IP subnet and externally learned routes onto the VXLAN fabric.

The EVPN route update also includes the following information:

- VNID for the L2VNI and VNID for the L3VNI for the tenant VRF.
- BGP next-hop IP address identifying the originating VTEP device.
- Router MAC address of the originating VTEP device.

Route Reflector

iBGP is the most common routing protocol deployed for the EVPN control plane in VXLAN fabrics (site-internal), and we will use this protocol for internal overlay CP. With iBGP, there is a requirement to have a full mesh between all iBGP speakers. To help scale and simplify the iBGP configuration, it is recommended to implement iBGP Route Reflectors (RR). The placement of the iBGP route reflectors is recommended to be implemented on the spines as they are central to all Leaf switches. The route reflector will reflect EVPN routes for the VTEP Leaf switches.

6.4.4 Site-External Overlay CP

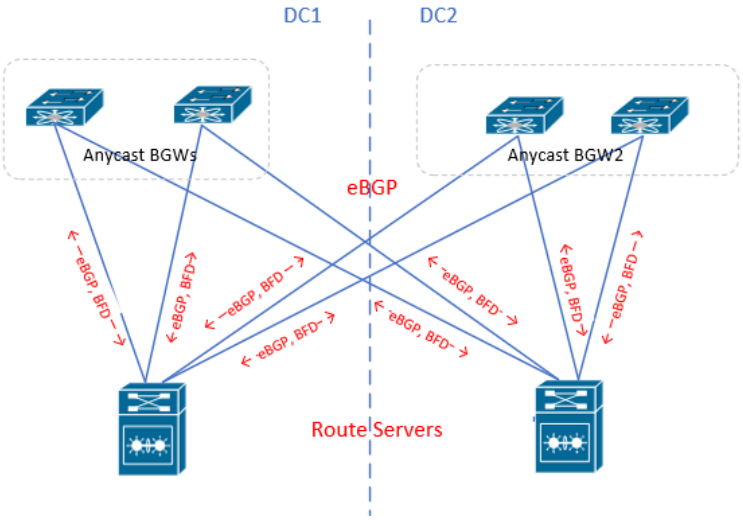

Figure 32. Site-External Overlay CP

Only eBGP is allowed as for underlay as for overlay control plane for site-external.

Route Server

Full mesh of MP-eBGP EVPN adjacencies across sites is mandatory. It is recommended to deploy a couple of Route-Servers with 3 or more sites. RS functions: EVPN routes reflection, next-hop-unchanged, route-target rewrite.

6.4.5 Overlay Both Sites

Combining the above, we have the following scheme for two sites:

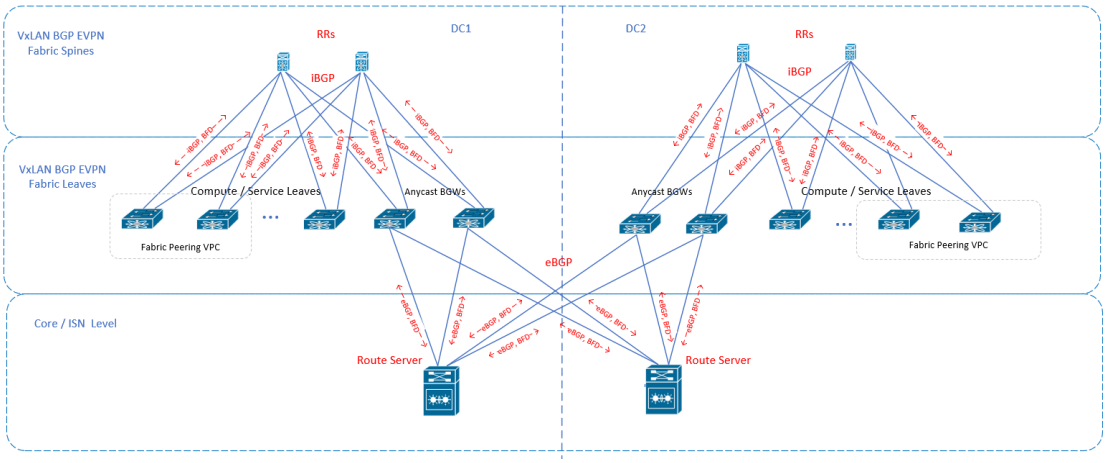

Figure 33. Multi-Site Overlay CP

6.5. VXLAN BGP EVPN Enhancements

6.5.1 Distributed Anycast Gateway

The use of the MP-BGP EVPN control plane introduces Distributed Anycast Gateway functionality. In this model, the default gateway function is fully distributed across all Leaf nodes within the Multi-Site VXLAN fabric (if corresponding L2 and

L3 VNIs are attached to the Leaf). Leveraging the Distributed Anycast Gateway provides several benefits:

- **No need in FHRP**: By distributing the gateway function, the network achieves better efficiency and increased bandwidth utilization. This approach eliminates the need to run a First Hop Redundancy Protocol (FHRP), which simplifies the network design.
- **Local Forwarding of Routed Traffic**: Workloads connected to the same Leaf node can communicate directly without involving the spine layer. This local forwarding reduces latency and improves overall performance.
- **Reduced L2 Hop Count**: Since the gateway function is distributed across all Leaf nodes, the number of hops between endpoints is minimized. This reduction in hop count contributes to lower network latency.

Symmetric IRB

With introduction of Anycast Gateway feature the question of symmetric / asymmetric IRB (Integrated Routing and Bridging) arrives.

Cisco supports only symmetric approach on NX-OS platforms with assigning of L3 VNI for each VRF.

Whereas asymmetric IRB follows the bridge–route–bridge mode of operation, symmetric IRB follows the bridge–route–route–bridge mode of operation.

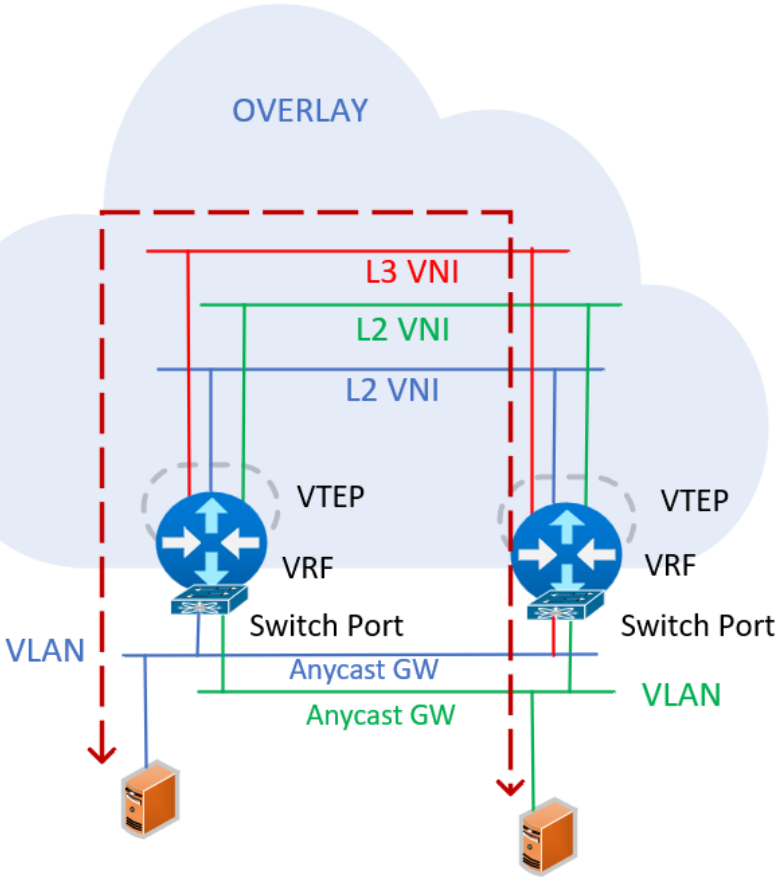

Figure 34. IRB Symmetric

On case of symmetric IRB, only the VNIs of locally-attached endpoints need to be defined in a VTEP (plus the transit L3 VNI), which in turns simplifies configuration and reduces scale requirements through optimized use of ARP and the MAC address table.

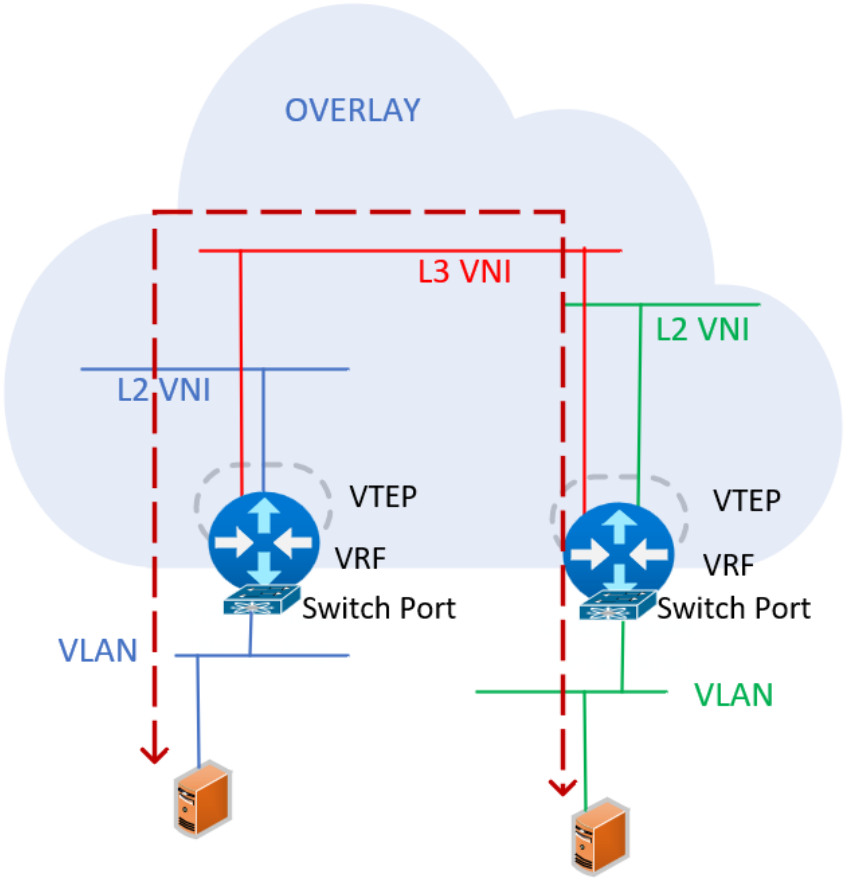

Figure 35. IRB-Symmetric Intra-VRF Routing

6.5.2 VPC

6.5.2.1 Fabric Peering

vPC Fabric Peering provides an enhanced dual-homing access solution without the overhead of wasting physical ports for vPC Peer Link. This feature preserves all the characteristics of a traditional vPC.

This is achieved by using of Virtual Peer Link over fabric. The vPC Fabric Peering peer-link is established over the transport network (the spine layer of the fabric). As communication between vPC peers occurs in this manner, control plane information CFS messages used to synchronize port state information, VLAN information, VLAN-to-VNI mapping, host MAC addresses are transmitted over the fabric. CFS messages are marked with the appropriate DSCP (we will use DSCP 56) value, which should be protected in the transport network.

The vPC Fabric Peering domain is not supported in the role of a Multi-Site vPC BGW

6.5.2.2 Peer keepalives

For peer keepalive Out-of-Band (mgmt0 or dedicated link) or In-Band (dedicated loopback interface) may be used. In our case In-Band loopback will be used.

6.5.2.3 PIP / VIP

VIP is used to present VTEP of VPC domain. Orphan Type-2 host as well as Type-5 routes are advertised using PIP. In this aspect, Virtual and Individual VTEPs can be considered.

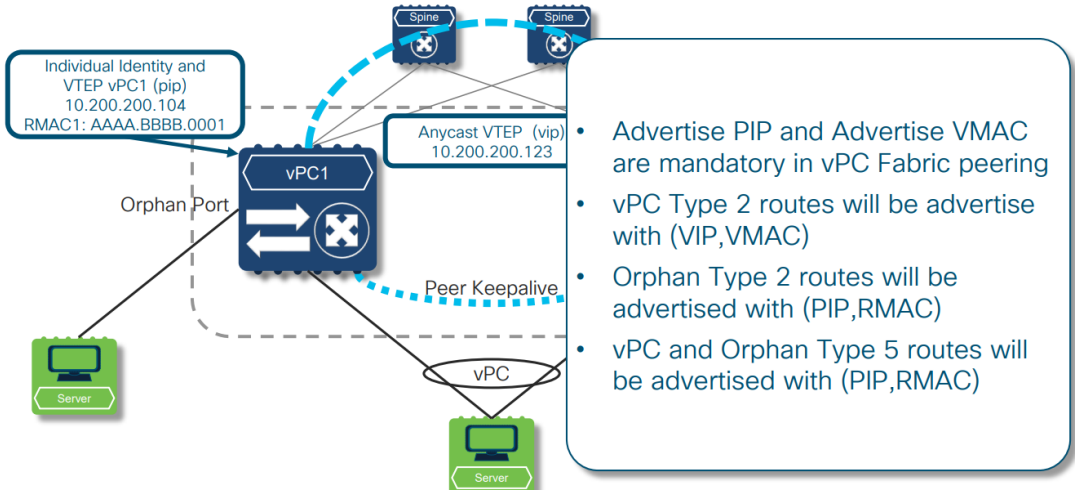

Figure 36. PIP vs VIP for VPC

6.5.2.4 Peer-gateway

VPC peer-gateway feature allows a vPC switch to act as the active gateway for packets addressed to the peer router MAC, it keeps forwarding of traffic local to the vPC node and avoids use of the peer link.

6.5.3 ARP Suppression

ARP communication allows creating IP/MAC mappings of the locally attached endpoints on the Leaf. In addition to the MAC-to-IP table being populated on the edge device, all the MAC information is added to the BGP EVPN control plane protocol.

Typically, when an endpoint wants to talk to another endpoint in the same subnet, it sends out an ARP request for determining the IP-to-MAC binding of the destination endpoint. The ARP request is flooded to all the endpoints that are part of that Layer 2 VNI. ARP snooping coupled with the BGP EVPN control plane information can help avoid flooding for known endpoints. By using ARP snooping, all ARP requests from an endpoint are redirected to the locally attached edge device. The edge device then extracts the destination IP address in the ARP payload and determines whether it is a known endpoint. Specifically, a query is done against the known endpoint information from the BGP EVPN control plane. If the destination is known, the IP-to-MAC binding information is returned. The local edge device then performs an ARP proxy on behalf of the destination endpoint. In other words, it sends out an unicast ARP response toward the requestor with the resolved MAC address of the known destination endpoint. In this way, all ARP requests to known endpoints are

terminated at the earliest possible point, which is the locally attached edge device or VTEP or Leaf. This is known as ARP suppression.

When the destination endpoint is not known to the BGP EVPN control plane (that is, a silent or undiscovered endpoint), the ARP broadcast needs to be sent across the VXLAN network. When the queried endpoint responds to the ARP request, the endpoint generates an ARP response that is learned by BGP EVPN and propagated to all Leaves where the L2 VNI is attached.

6.6. Overlay Data Forwarding

6.6.1 Multi-Tenancy in EVPN VXLAN

In an EVPN VXLAN overlay network, VXLAN network identifiers (VNIs) serve as the foundation for defining Layer-2 domains and enforcing Layer-2 segmentation. These VNIs prevent Layer-2 traffic from crossing VNI boundaries, ensuring isolation. This concept is often referred to as **L2 multi-tenancy**.

Similarly, Layer-3 segmentation among VXLAN tenants is achieved through Layer-3 VRF. Each tenant is assigned a separate Layer-3 VNI, which maps to an individual VRF instance. Within this setup:

- **Each tenant has its own VRF**: Each tenant (or customer) in the VXLAN fabric has its dedicated VRF routing instance. This separation ensures that the Layer-3 routing domain remains isolated from other tenants.
- **IP subnets within VNIs**: The IP subnets associated with a given tenant's VNIs reside within the same Layer-3 VRF instance. This arrangement further enhances the isolation between tenants.

This approach is commonly known as **L3 multi-tenancy** or simply **multi-tenancy**.

In the current setup, the INT VDC (Virtual Device Context) is configured with 25 different VRFs. Here's an example of some L3 functions these VRFs perform:

- **Intra-VRF routing**: These VRFs facilitate routing between campus networks and firewalls. Essentially, they handle traffic within the same VRF stretched between campuses and datacentre.
- **SVI interfaces as Default Gateways**: Each VRF also has one or more SVI interfaces that act as default gateways for datacentre endpoints.
- **Transit VRF for Inter-Segment Traffic**: The VRF "Transit" is specifically designed for inter-firewall traffic. It provides communication between network segments separated by firewalls.

Firewalls as Default Gateway

In the current setup, most subnets in the datacentre are terminated on firewalls. The legacy network infrastructure provides L1/L2 connectivity between the end hosts and the FW. We must retain this behavior and the new fabrics will also act as a pure Layer 2 (L2) infrastructure for these VLANs (L2 VNIs), forwarding traffic based on MAC addresses.

6.6.2 Example of Migration. VRF Academics

6.6.3 Tenant Routed Multicast

Layer 3 Mode Tenant Routed Multicast (TRM) will be implemented.

Another TRM (Layer 2/3) is necessary for a mixed platform fabric with VTEPs that can support VXLAN BGP EVPN unicast routing but do not support TRM (Non-TRM). All our switches support TRM.

VNI VRF L3 with TRM enabled will have an associated multicast group. The multicast group associated with each VRF is called the default **Multicast Distribution Tree** (MDT) group. There are separate multicast groups for L2 VNI and L3 VNI. L2 VNI will only carry traffic for Unknown Unicast and broadcast once TRM is enabled for VRF for IP multicast. Non-IP multicast traffic will still be treated as BUM traffic in the L2 VNI. VTEPs with VRF configured become the source and destination for the default MDT group. As the default recipient of the MDT group, each VTEP initiated a connection (*, G) to the RP existing on the spine nodes. As a source for the default MDT group, each VTEP registers as a source with the spine RP and initiates the (S, G) state in its MRIB table. The source address in the (S, G) entry will be the NVE VTEP loopback interface, and the multicast group will be the default MDT.

As result, a VXLAN BGP EVPN fabric with TRM will have two multicast domains: the underlay in the default VRF and the overlay in a custom-created tenant VRF. PIM Sparse mode is a requirement in the underlay to enable TRM, and PIM Sparse Mode can also be configured in the tenant VRF as part of the overlay. An RP is required in the underlay and overlay multicast domains in such scenarios.

The RP in the underlay will be configured on the spine switches. The spine switches are the RP for the same underlay multicast groups already used for L2 VNI BUM traffic.

The RP in the overlay supports three deployment models:

- Internal RP (RP-less) or Anycast RP. In the Anycast RP deployment model, every VTEP, including the border nodes, is configured as an RP inside the VRF. The RP function is enabled on every VTEP where the VRF is provisioned by configuring a PIM-enabled loopback interface with a common IP address.
- External RP. This mode means that the external RP is a PIM router attached to the border Leaf, or it can be placed anywhere in the external network with its redundancy model.
- RP Everywhere This mode is combination of two above options with using of PIM Anycast RP.

The mode should be chosen depending on current multicast configuration and future requirements. For now, multicasting is used for only 2 VRFs. RPs are configured for them. For these VRFs Anycast RP mode will be implemented.

6.6.4 Southbound Loop Detection and Mitigation

In Cisco NX-OS, native southbound loop detection and mitigation are provided for VXLAN EVPN fabrics. This functionality is also available as for a single fabric as for VXLAN EVPN Multi-Site deployments. For these Multi-Site scenarios, loop

detection and mitigation focus on identifying backdoor links—the most common cause of multi-site outages during extension or migrations.

Unlike many loop protection solutions that detect loops in the overall topology and shut down offending ports, VXLAN EVPN Loop Detection and Mitigation operates at the “VLAN-level.” Unlike Spanning Tree, which proactively calculates forwarding trees, VXLAN EVPN Loop Detection and Mitigation takes precautions to prevent loop existence and their introduction into the Overlay.

Beginning with Cisco NX-OS Release 10.1(1) VXLAN EVPN loop detection and mitigation is supported on Cisco Nexus 9300-FX3 and -GX platform switches.

To configure this, NGOAM feature should be enabled and space for TCAM ing-sup region should be created. After that **ngoam loop-detection** may be enabled.

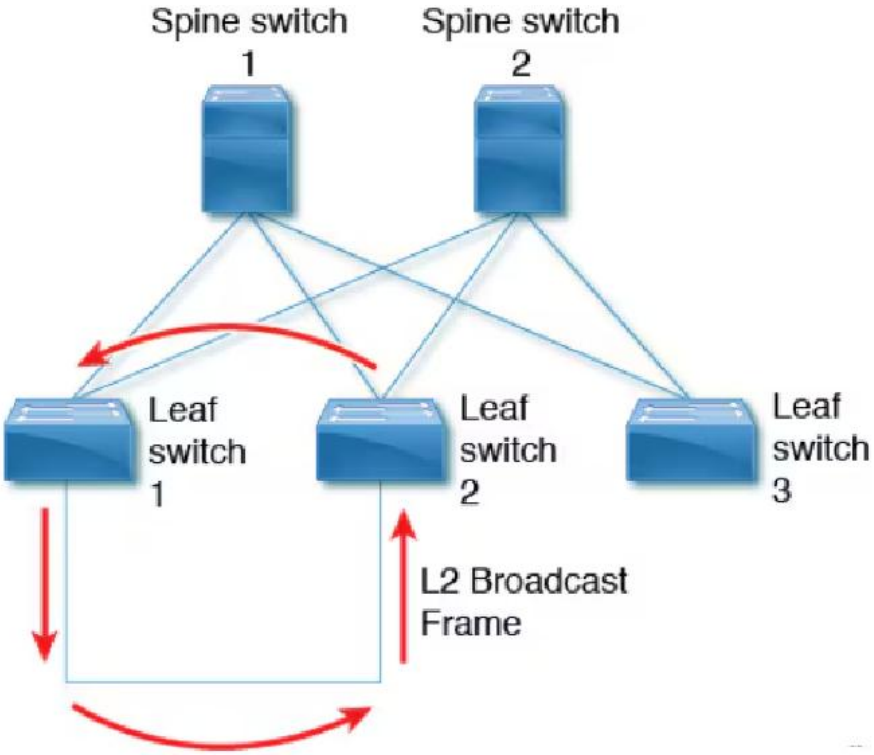

Figure 37. Loop Detection and Mitigation

The feature operates in three phases:

- **Loop Detection**: Sends a loop detection probe under the following circumstances:
  - When requested by a client, as part of a periodic probe task.
  - As soon as any port comes up.
- Loop Mitigation:
  - Blocks the VLANs on a port once a loop has been discovered.
  - Displays a syslog message.
  - Because loops can lead to incorrect local MAC address learning, this phase also flushes the local and remote MAC addresses.

- **Loop Recovery**: Once a loop is detected on a particular port or VLAN and the recovery interval has passed, recovery probes are sent to determine if the loop still exists. When NGOAM recovers from the loop, a syslog message appears.

6.7. Management

6.7.1 OOB Management

Nexus Dashboard Fabric Controller (NDFC) streamlines VXLAN BGP EVPN fabric configuration, accelerates deployment, and offers customization options while adhering to best practices. The efficacy of this approach has been tested within the COMPANYB Lab environment, yielding favorable results. NDFC will manage fabric infrastructure via OOB, so we have to provide connectivity form NDFC to switches OOB.

The CompanyA has decided not to incorporate the NDFC into its operational processes, and out-of-band (OOB) deployment is not within the scope of this project. Consequently, CompanyA will deploy a temporary OOB network, which will be deployed as a dedicated VLAN and corresponding subnet within the existing legacy data infrastructure.

6.7.2 In-Band Management

– Not Completed –

We will use existing approach to provide in-band access. Access will be provided via loopback interfaces in VRF SwitchMGMT.

6.8. Guidelines and Limitations for VXLAN EVPN Multi-Site (version 10.4(x))

Only the limitations relevant to our deployment are provided below.

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

- The **EVPN multi-site dci-tracking** is mandatory for anycast BGWs and vPC BGW DCI links. The EVPN multi-site fabric-tracking is mandatory only for anycast BGWs. For vPC based BGWs, this command is not mandatory. The NVE Interface will be brought up with just the dci tracked link in the up state.

We will rely on NDFC automation. What we can see in our LAB environment this requirement has been met.

- VXLAN EVPN Multi-Site and Tenant Routed Multicast (TRM) are supported between sources and receivers deployed across different sites.

This feature will be utilized in our deployment.

- The Multi-Site BGW allows the coexistence of Multi-Site extensions (Layer 2 unicast/multicast and Layer 3 unicast) as well as Layer 3 unicast and multicast external connectivity.

It means that we can use BGW in the BL role as for overlay unicast as for multicast forwarding.

- In TRM with Multi-Site deployments, all BGWs receive traffic from fabric. However, only the designated forwarder (DF) BGW forwards the traffic. All other

BGWs drop the traffic through a default drop ACL. This ACL is programmed in all DCI tracking ports. Don't remove the EVPN multi-site dci-tracking configuration from the DCI uplink ports. If you do, you remove the ACL, which creates a nondeterministic traffic flow in which packets can be dropped or duplicated instead of deterministically forwarded by only one BGW, the DF.

We will rely on NDFC automation. What we can see in our LAB environment this requirement has been met.

- Anycast mode can support up to six BGWs per site.

Only 2 BGWs per site will be deployed.

- BGWs in a vPC topology are supported.

Anycast mode will be used.

- iBGP EVPN Peering between BGWs of different fabrics/sites isn't supported.

eBGP EVPN will be deployed.

- The peer-type fabric-external command configuration is required only for VXLAN Multi-Site BGWs (this command must not be used when peering with non-Cisco equipment).

We will rely on NDFC automation.

- Anycast mode can support only Layer 3 services that are attached to local interfaces.

It is not planned to connect endhosts to BGWs

- In Anycast mode, BUM is replicated to each border Leaf. DF election between the border Leaves for a particular site determines which border Leaf forwards the inter-site traffic (fabric to DCI and conversely) for that site.
- In Anycast mode, all 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 BGWs.

BGWs will be configured in anycast mode.

- vPC mode can support both Layer 2 hosts and Layer 3 services on local interfaces.

BGWs will be configured in anycast mode.

- In vPC mode, BUM is replicated to either of the BGWs for traffic coming from the external site. Hence, both BGWs are forwarders for site external to site internal (DCI to fabric) direction.

BGWs will be configured in anycast mode.

- In vPC mode, BUM is replicated to either of the BGWs for traffic coming from the local site Leaf for a VLAN using Ingress Replication (IR) underlay. Both BGWs are forwarders for site internal to site external (fabric to DCI) direction for VLANs using the IR underlay.

BGWs will be configured in anycast mode.

- In vPC mode, BUM is replicated to both BGWs for traffic coming from the local site Leaf for a VLAN using the multicast underlay. Therefore, an election

happens, and the decapsulation/forwarder winner only forwards the site-local traffic to external site BGWs for VLANs using the multicast underlay.

BGWs will be configured in anycast mode.

- Prior to NX-OS 10.2(2)F only ingress replication was supported between DCI peers across the core. Beginning with Cisco NX-OS Release 10.2(2)F both ingress replication and multicast are supported between DCI peers across the core.

We will use ingress replication.

- In vPC mode, all 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.

BGWs will be configured in anycast mode.

- If different Anycast Gateway MAC addresses are configured across sites, enable ARP suppression for all VLANs that have been extended.

A single MAC address will be configured.

- 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).

We will rely on NDFC automation. What we can see in our LAB environment this requirement has been met.

- PIM BiDir is not supported for fabric underlay multicast replication with VXLAN Multi-Site.

PIM-ASM Anycast will be used.

- PIM is not supported on Multi-Site VXLAN DCI links.

Ingress replication will be used.

- Beginning with Cisco NX-OS Release 9.3(5), VTEPs support VXLAN-encapsulated traffic over parent interfaces if subinterfaces are configured. This feature is supported for VXLAN EVPN Multi-Site and DCI. DCI tracking can be enabled only on the parent interface.

We are considering setting up a parent interface facing the ISNs (cores) as a possible option to handle BGW related functions, while the subinterfaces will be used for external L3 connectivity.

- To improve the convergence in case of fabric link failure and avoid issues in case of fabric link flapping, ensure to configure multi-hop BFD between loopbacks of spines and BGWs. In the specific scenario where a BGW node becomes completely isolated from the fabric due to all its fabric links failing, the use of multi-hop BFD ensures that the BGP sessions between the spines and the isolated BGW can be immediately brought down, without relying on the configured BGP hold-time value.
- In a VXLAN Multi-Site environment, a border gateway device that uses ECMP for routing through both a VXLAN overlay and an L3 prefix to access remote site subnets might encounter adjacency resolution failure for one of these routes. If

the switch attempts to use this unresolved prefix, it will result in traffic being dropped.