DC fabric design & implementation

Paresh Khatri
CTO IP Networks APAC

NOKIA

SANOG

Agenda

1. Data center architectures
2. EVPN basics

Agenda

1. Data center architectures
2. EVPN basics

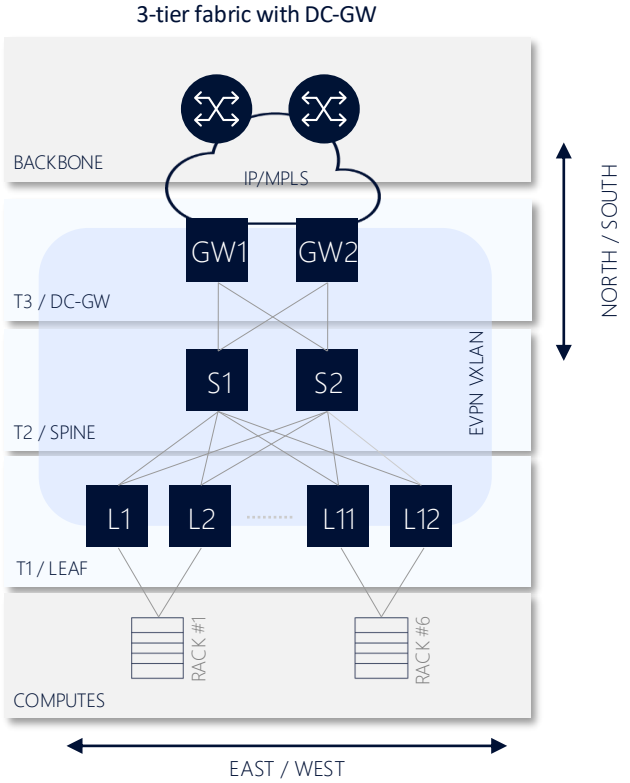
Agenda

1. Data center architectures
   a. In a nutshell
   b. Physical
   c. Logical
2. EVPN basics

Data center key foundations

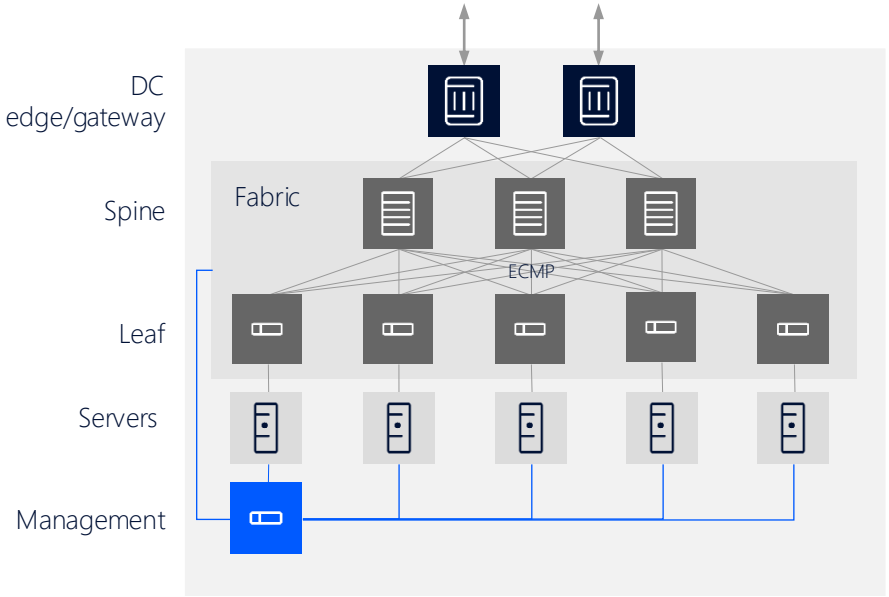
Goals and topology

- Purpose of a data center fabric
  - Interconnect compute, storage and external networks at scale
  - Ensure low-latency and high-throughput across the fabric
  - Provide high availability through redundant design
  - Enable secure communication between workloads
- In practice, this is done through a Clos architecture.
- Key factors like oversubscription ratio, fault tolerance, redundancy and scale will determine physical characteristics of the data center.



Modern data center network architectures

The industry has converged



Non-blocking fabrics

- IP and EVPN fabrics
- DC gateway or border leaf derivatives
- Collapsed core for edge DC
- Scale via super spines / pods

ASICs tailored per use case

- Range of different ASICs on the market
- Key properties: latency, programmability, port speed & density, feature set

OOB management

- Merchant silicon
- 1G/10G port speeds

Modern data center network architectures

Key elements

Physical architecture

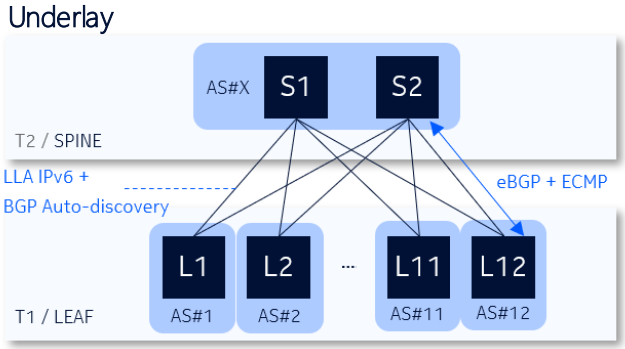
- (Folded) Clos leaf-spine topologies
- Multi-stage
- Multi-tier for scale

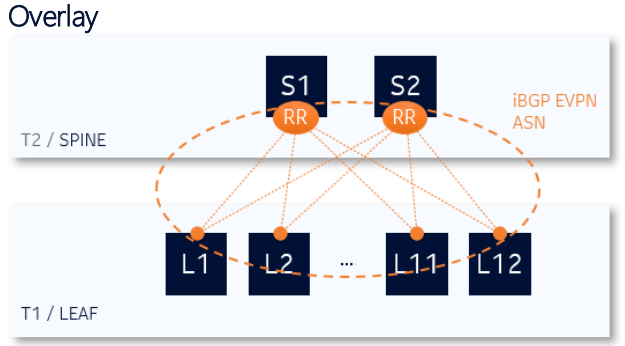
Underlay

- IP underlay
- Single BGP instance with no additional IGPs

Overlay

- BGP-EVPN control plane
- VXLAN data plane
- Layer 2 and Layer 3 multi-tenant services

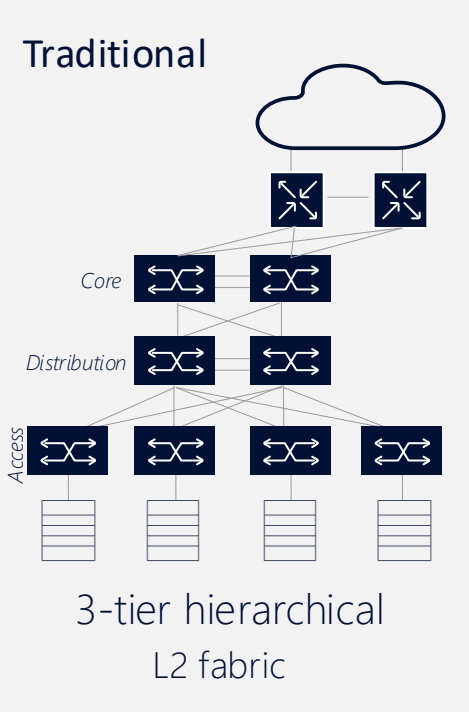


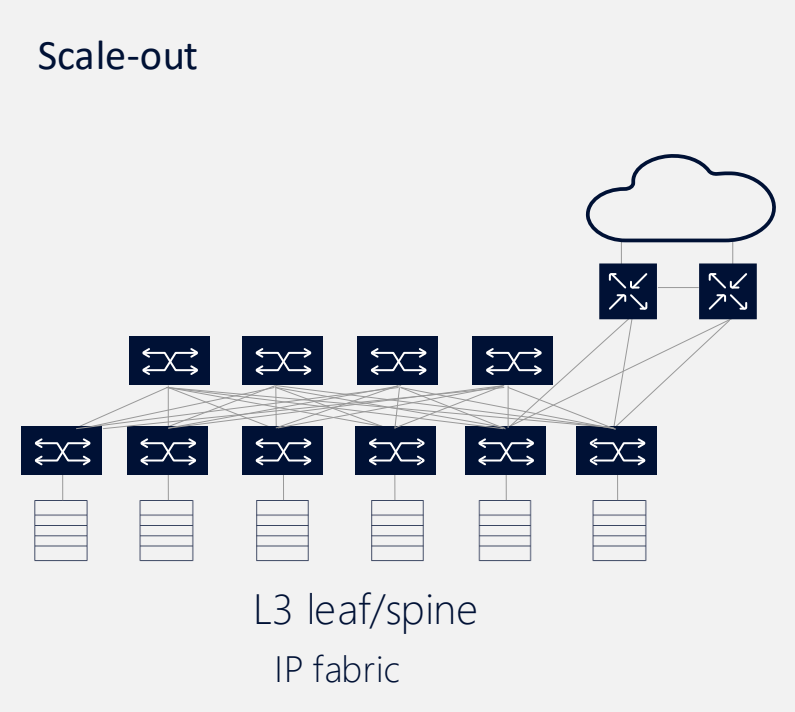


Agenda

1. Data center architectures
   a. In a nutshell
   b. **Physical**
   c. Logical
2. EVPN basics

Evolution of data center architecture





IP fabric architecture

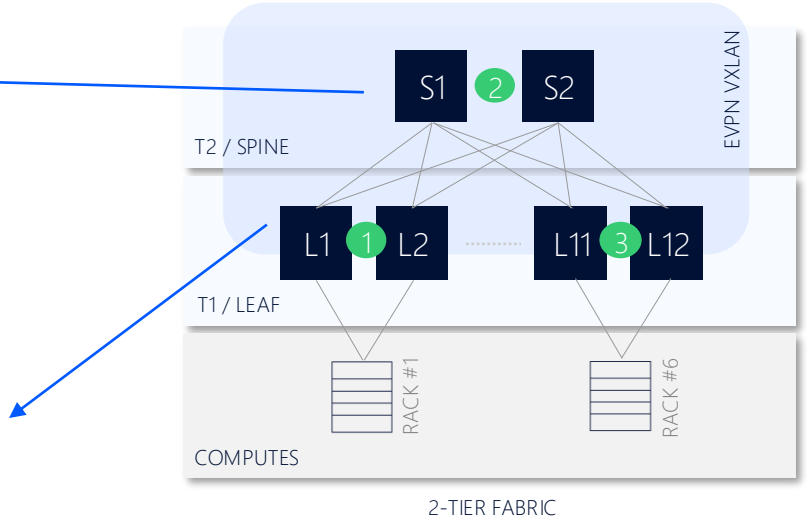
3-stage Clos

Spine:

- Provides connectivity between leaf switches. The major role of the spine is to participate in the control-plane and data plane operations for traffic forwarding between leafs
- Spines are not interconnected

Leaf:

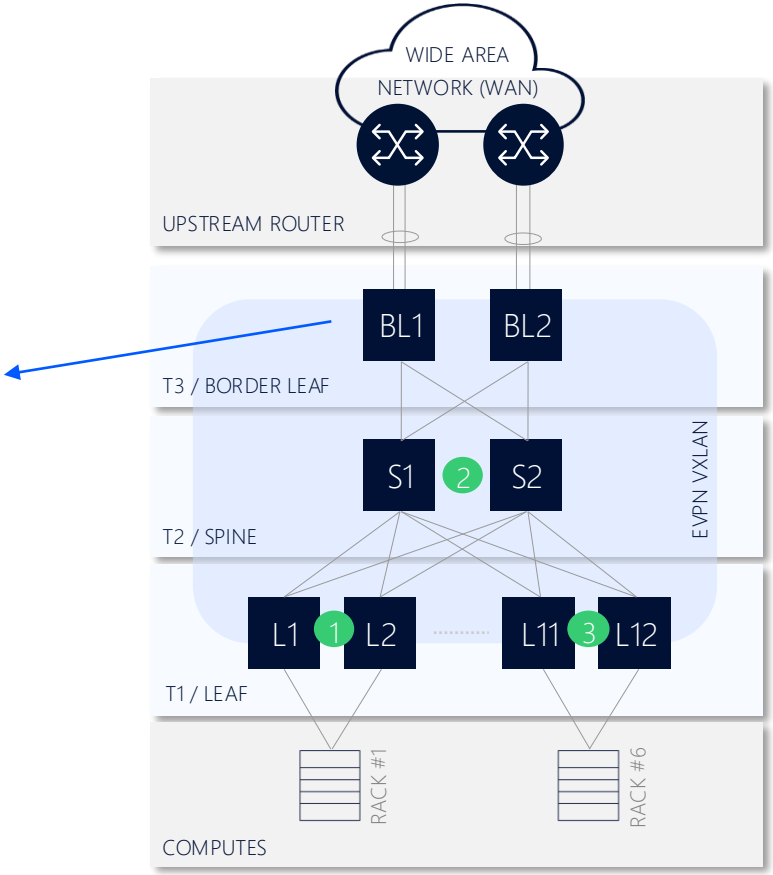
- Provides connectivity to the endpoints in the data center network which include compute servers and storage devices. The leaf provides VXLAN endpoint functionality and is also referred to as a Top of Rack (ToR) and can be deployed as a single switch or a dual switch to allow for server/storage redundant connections
- Each Leaf connects to each Spine



IP fabric architecture

External connectivity

Border Leaf: Provides connectivity in and out of the IP Fabric to the WAN edge or other networks within the Data Center. Endpoints attached here normally include networking devices like routers, switches, load balancers, firewalls, and any other physical or virtual networking endpoints. The Border Leaf also provides VXLAN endpoint functionality

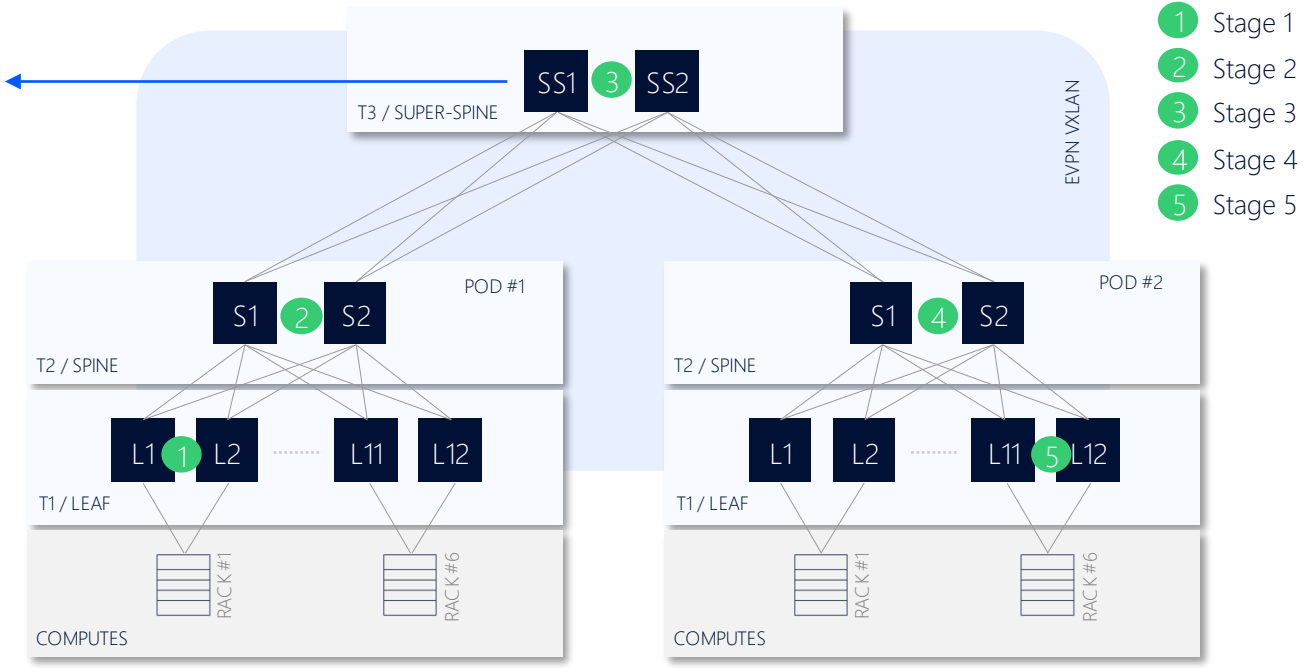


1 Stage 1
2 Stage 2
3 Stage 3

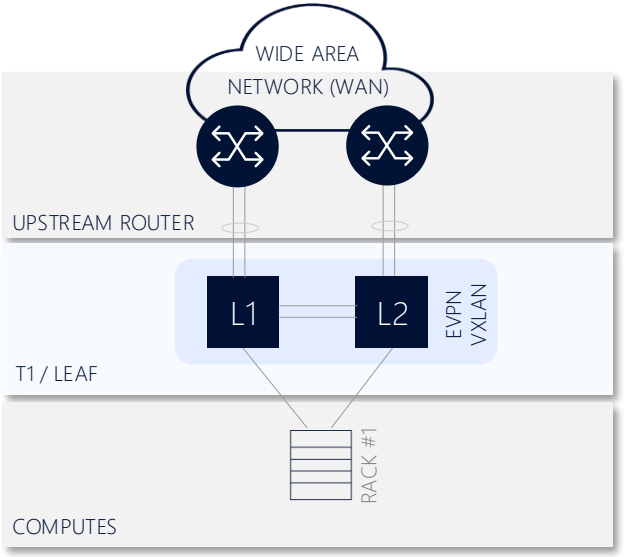
IP fabric architecture

5-stage Clos

Super Spine: Used in a 5-stage Clos providing connectivity between PoDs within the Data Center. Each PoD would consist of a 3-stage Clos IP Fabric. The major role of the spine is to participate in the control-plane and data plane operations for traffic forwarding between PoDs



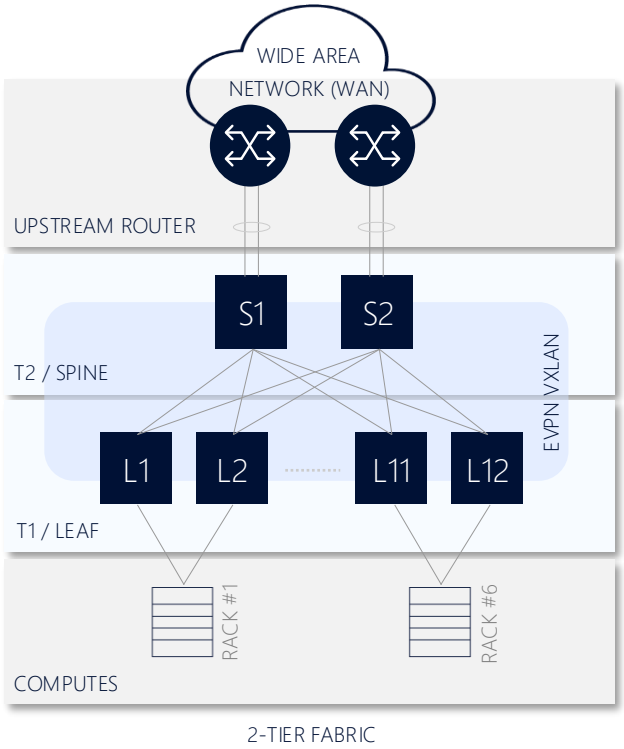
1-tier topology (single-rack)



1-TIER FABRIC

- Single-rack deployment with 2 leafs
- 2x back-to-back links are required (in this topology only)
- eBGP/iBGP between leafs (EVPN/VXLAN)
- The external bandwidth must not exceed the capacity of the back-to-back links.

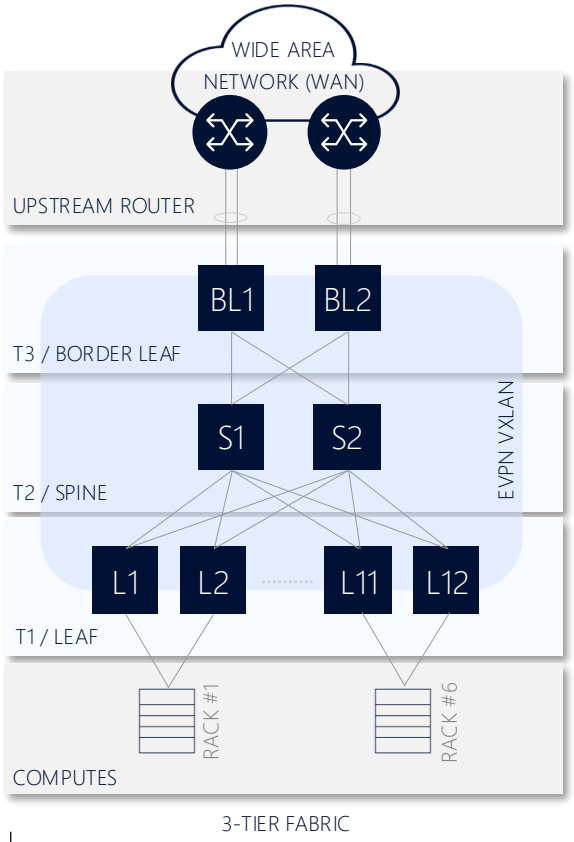
2-tier topology (leaf/spine)



2-TIER FABRIC

- 2 tiers:
  - Leaf
  - Spine
- VXLAN/VLAN handoff to upstream routers on spines.
- Well-suited for high external bandwidth (N/S)

3-tier topology (leaf/spine/border leaf)



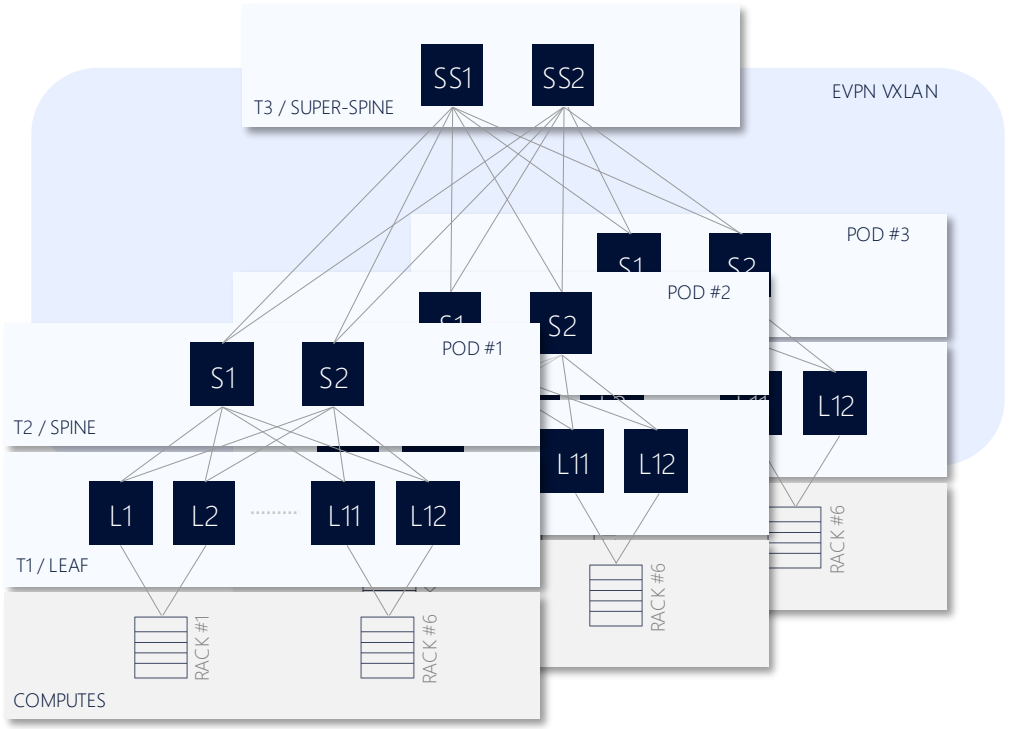
3-TIER FABRIC

- 3 tiers:
  - Leaf
  - Spine
  - Border leaf
- VXLAN/VLAN handoff to upstream routers on dedicated border leafs.
- Well-suited for high external bandwidth (N/S)

3-tier topology (multi-pod)

Design considerations (1/2)

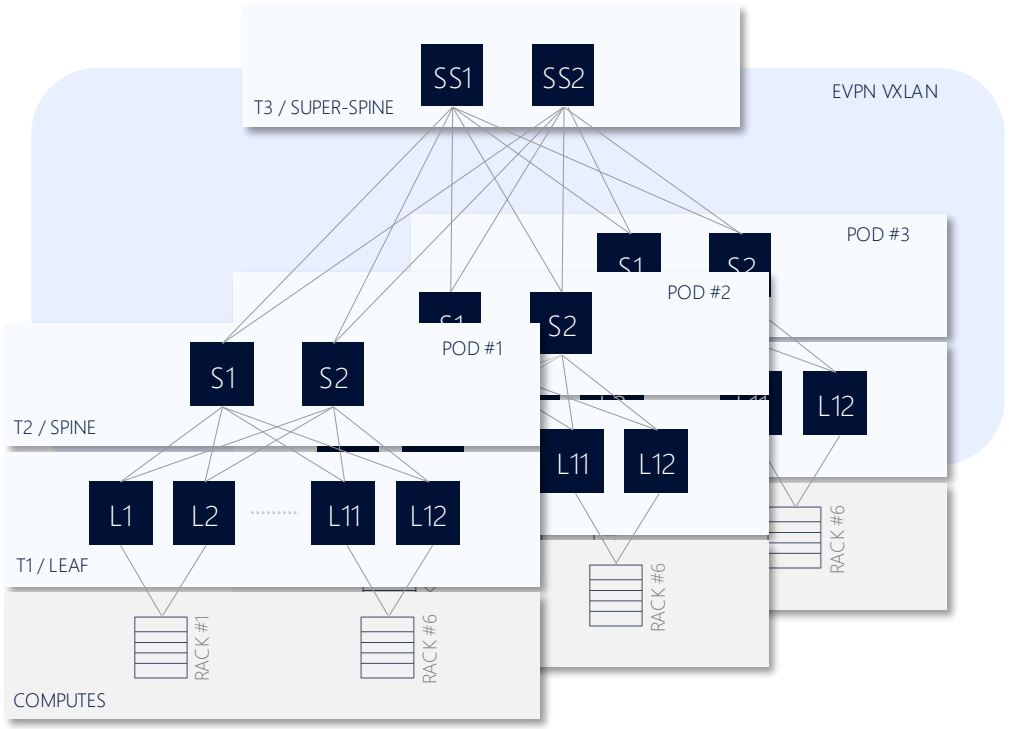
- A POD is composed of a pair of spines and its associated leafs.
- A multi-POD design is only possible with a 3-Tier topology.
- PODs are inter-connected at the super-spine layer.
- The EVPN VXLAN domain is extended accross the PODs.



3-tier topology (multi-pod)

Design considerations (2/2)

- The multi-POD design allows an incremental build up of the fabric. The fabric can grow without impacting the existing nodes & cabling.
- East-west communications between PODs require additional hops when traversing the super-spine layer – this impacts latency.
- It is recommended to keep applications having strict latency requirements within a single POD.



Agenda

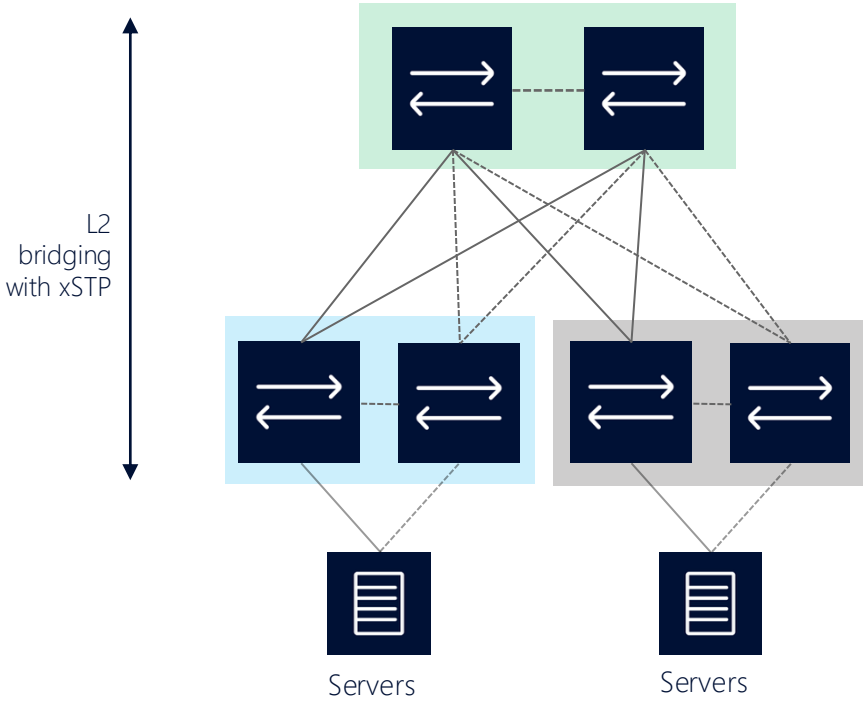
1. Data center architectures
   a. In a nutshell
   b. Physical
   c. Logical
2. EVPN basics

Agenda

1. Data center architectures
    a. In a nutshell
    b. Physical
    c. Logical
        a. Overview
        b. Underlay
        c. Overlay
2. EVPN basics

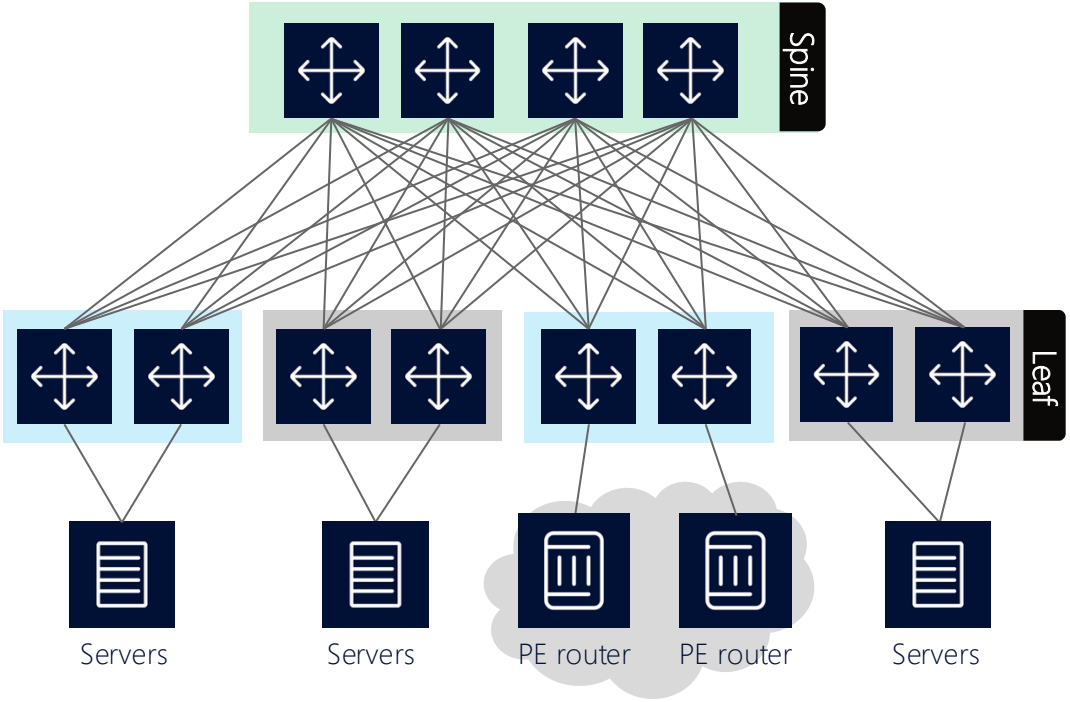
L2 fabric

Legacy design



- Scalability challenges
- Broadcast storms → STP and RSTP and MSTP
- Temporary L2 loops while converging
- Sub-optimal bandwidth usage
- Inefficient load distribution

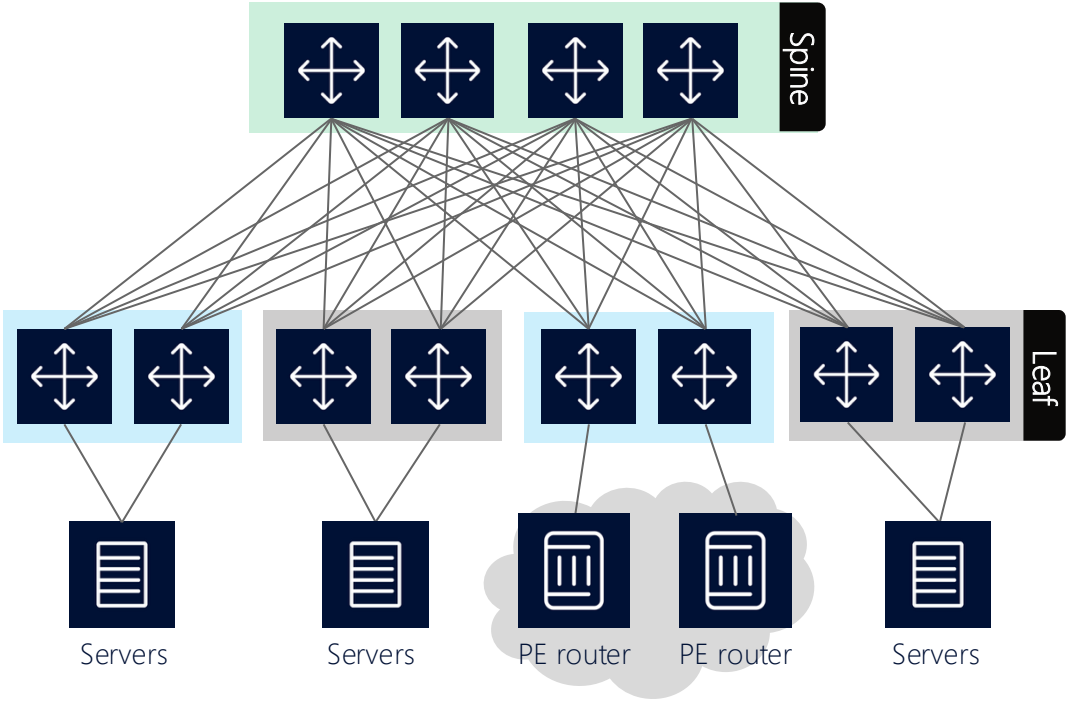
L3 Clos fabric



- Easy to scale up
- L2 loop-free
- Load balancing: ECMP and hash labels for extra entropy
- Design inputs
  - Workload BW – per rack BW
  - Traffic forecast and oversubscription ratio
  - Fault tolerance
  - Protocols: BGP, ISIS, OSPF

L3 fabric

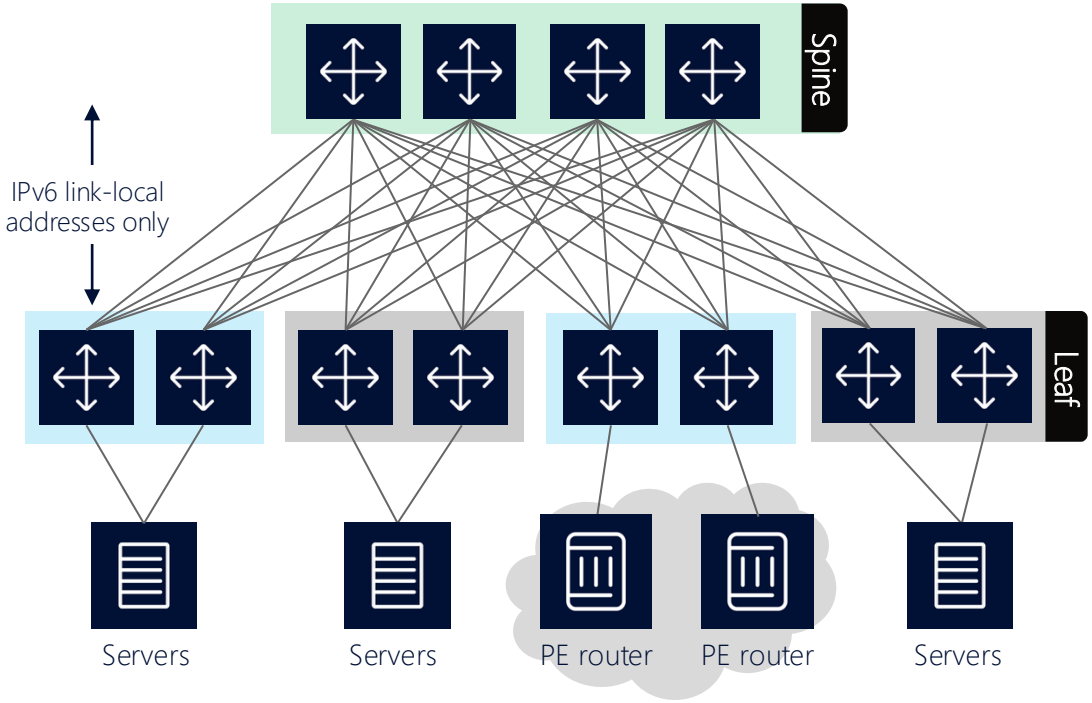
Underlay with eBGP/RFC 7938



- Simple design
- Scalable
- Lower HW resource consumption
- Flexible policy engine
- Smaller failure impact radius

L3 fabric

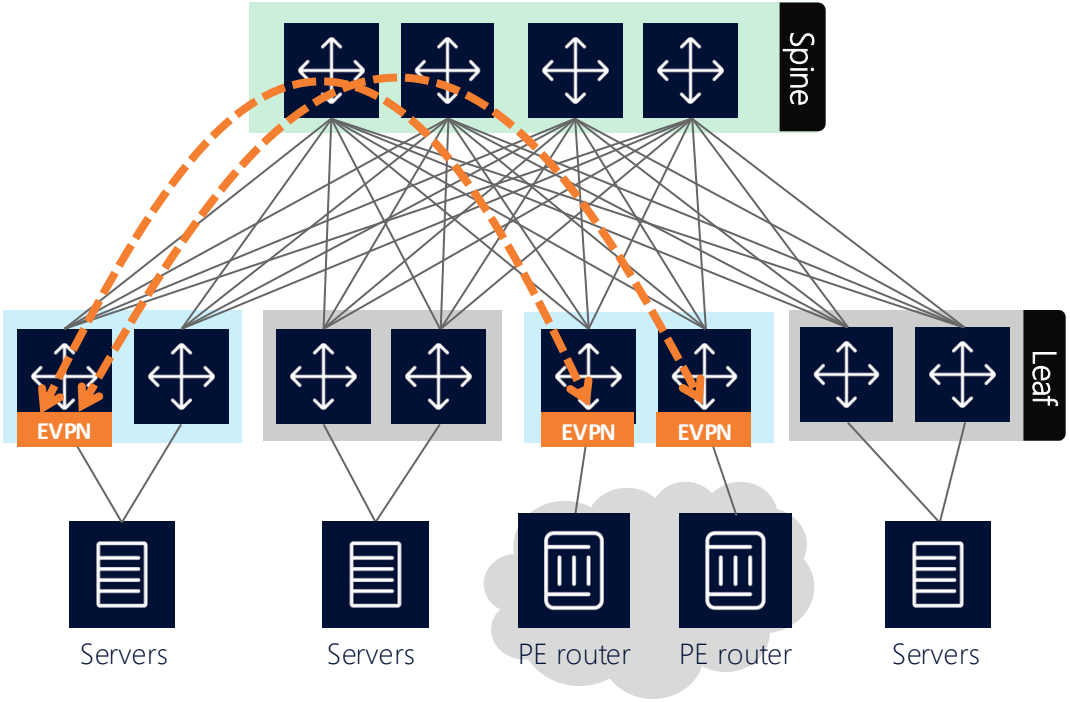
IPv6 unnumbered underlay



- Interfaces establish BGP sessions using IPv6 link-local addresses without needing a unique global IP address
- IPv6 ND Router Advertisements (RAs) are used to announce and learn peers' link-local addresses
- BGP unnumbered feature establishes the peerings based on the IPs exchanged

EVPN-VXLAN

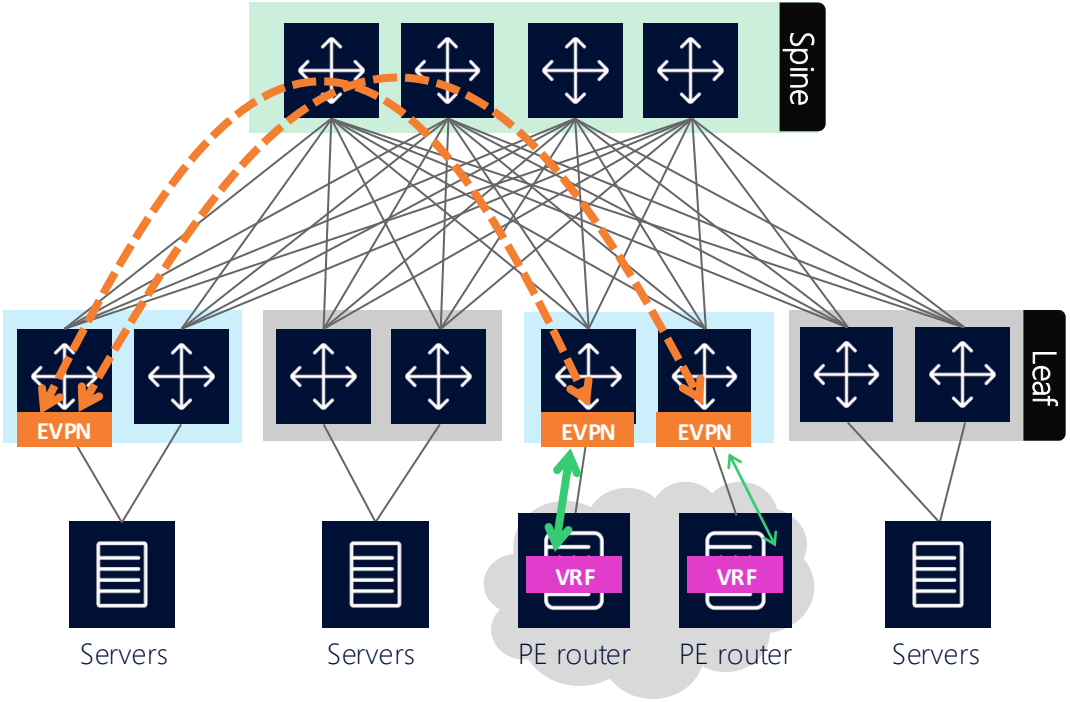
Overlay with EVPN



- Network segmentation
- Loop prevention
- L2 multi-homing
- Anycast gateway
- Host mobility
- ARP suppression

Connecting to WAN

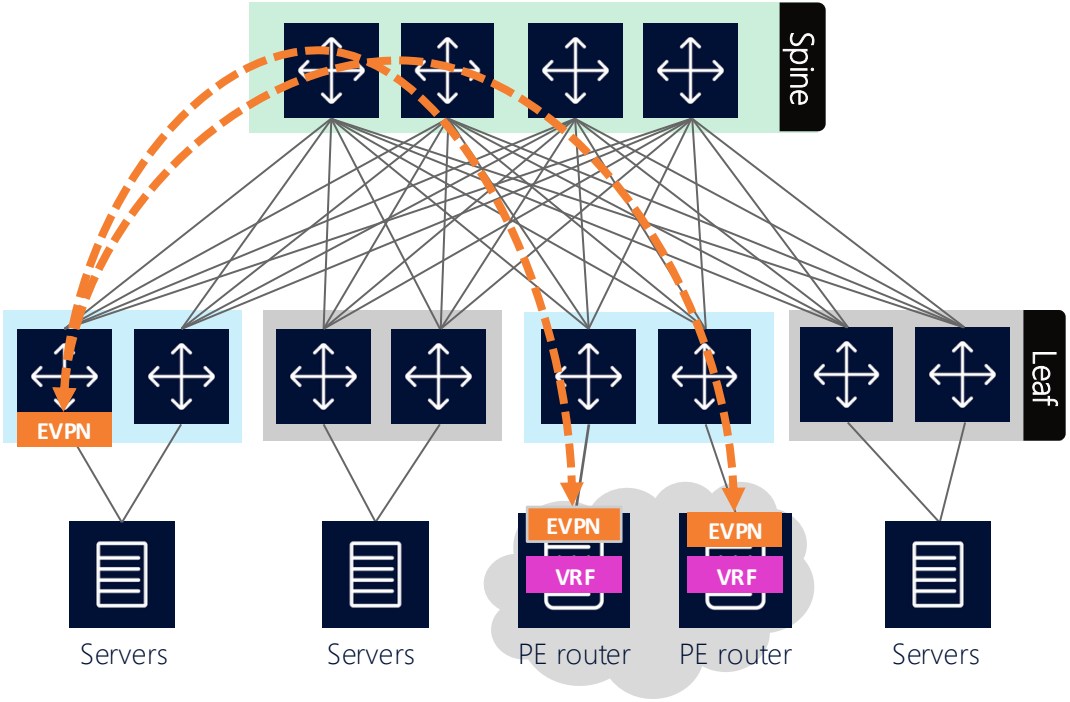
Vlan hand-off



- Clear demarcation between DC and WAN
- No need for high end PE
- For L2 networks, Ethernet Segments
- For L3 networks Inter-AS Option A

Connecting to WAN

Integrated model



- Faster convergence for L2
- Need for routing policies to segregate WAN and DC
- Need for a higher-end PE platform
- No need for Inter AS Option-A

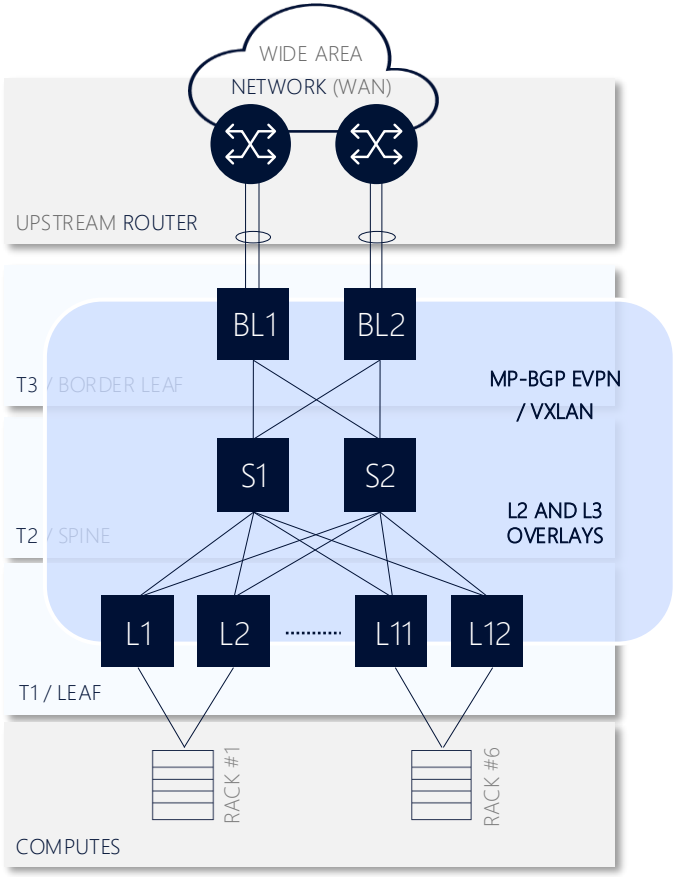
Agenda

1. Data center architectures
    a. In a nutshell
    b. Physical
    c. Logical
        a. Overview
        b. Underlay
        c. Overlay
2. EVPN basics

Fabric underlay

L3 Fabric with EVPN/VXLAN

- The Layer 3 spine and leaf architecture addresses limitations of the classic Layer 2 fabric such as spanning-tree protocol looping and multi-pathing uplink constraints by providing consistent and normalized any-to-any latency, throughput and performance for all racks while allowing simple scaling and a fully non-blocking architecture, if required.
- The L3 fabric routing is based on RFC7938 (Use of BGP for Routing in Large-Scale Data Centers) which describes a practical routing design that can be used in large-scale data centers.
- When building L3 Fabric, VXLAN and MP-BGP EVPN technologies are leveraged to provide control-plane and data-plane separation for both Layer 2 and Layer 3 forwarding in a VXLAN overlay network.

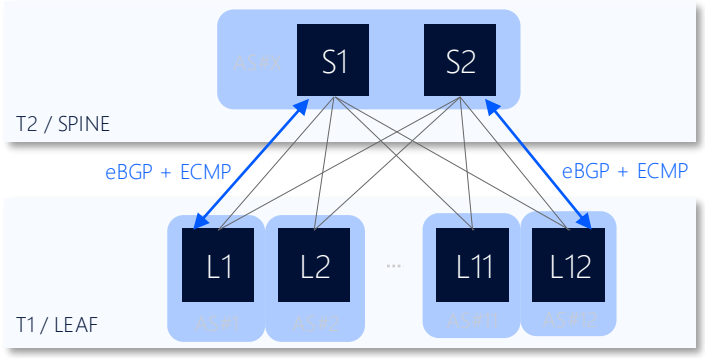


L3 FABRIC (MP-BGP EVPN/VXLAN)

Fabric underlay

BGP as underlay routing protocol (RFC7938)

- eBGP single-hop sessions are established over direct point-to-point links interconnecting the network nodes
- Private ASN (Autonomous System Number)
  - Each leaf with unique ASN
  - All spines/super-spines with same ASN per pod
- IPv6 link-local addresses can be used on point-to-point links, along with BGP peer auto-discovery to ease the planning & configuration process (implies that only a single IP address needs to be configured on each switch)
- No multi-hop or loopback sessions are used, even in the case of multiple links between the same pair of nodes.
- The point-to-point links are not advertised into BGP. Since the eBGP-based design changes the next-hop address at every device, distant networks will automatically be reachable via the advertising eBGP peer.
- There is no requirement for an IGP!
- Only the system loopback interfaces (VTEP) are advertised across the fabric.



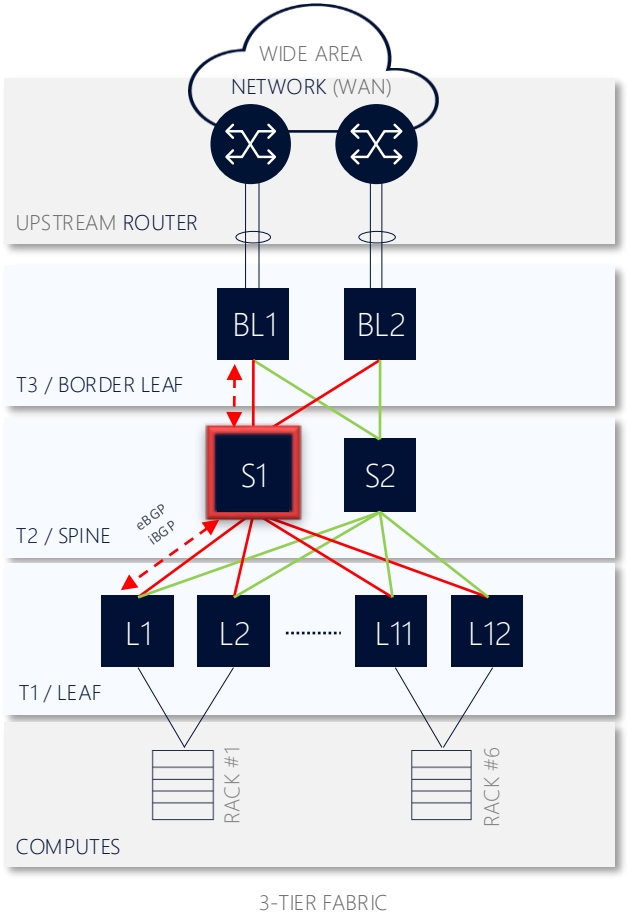
Recommendations for underlay

- Private ASNs (64512 – 65534), reused in different DCs
- IPv6 LLAs & BGP peer auto-discovery (*)

Fabric underlay

Resiliency

- Failure detection in the underlay relies on:
  - The **physical state of the links (L1)**. Any loss of signal (LOS) is immediately detected on both sides.
  - Or **Bidirectional Failure Detection (BFD)**, L3 protocol applied to BGP. A 100ms transmit interval with a multiplier of 3 is applied, offering a maximum failure detection time of 300ms.
- The BGP timers and route advertisements are optimized to speed up the convergence for both underlay and overlay EVPN (Rapid update / Rapid withdrawal)
- Once the fabric has converged, the flows are redirected to the other paths (BGP Multipath / ECMP)



3-TIER FABRIC

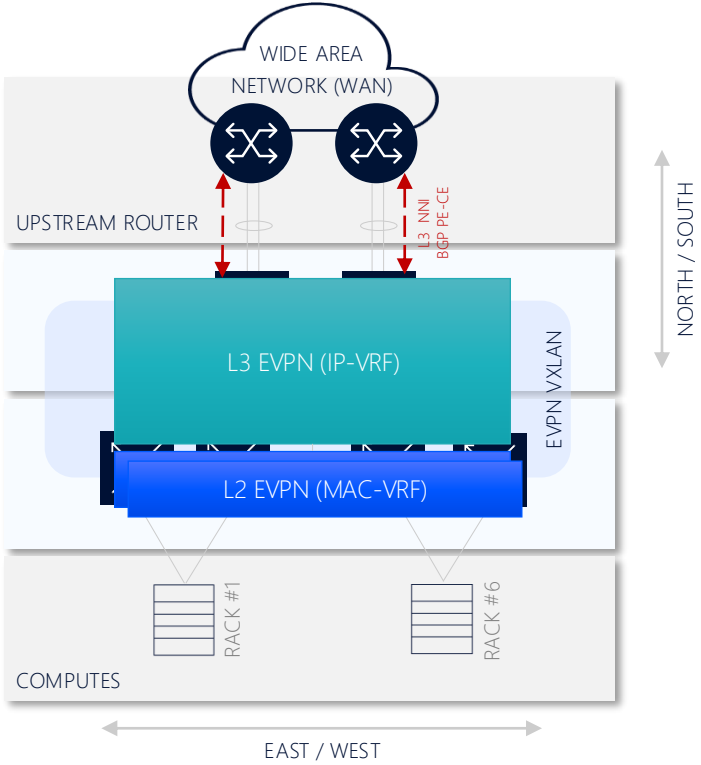
Agenda

1. Data center architectures
   a. In a nutshell
   b. Physical
   c. Underlay
   d. Overlay
2. EVPN basics

Overlay networks

EVPN/VXLAN Overlays

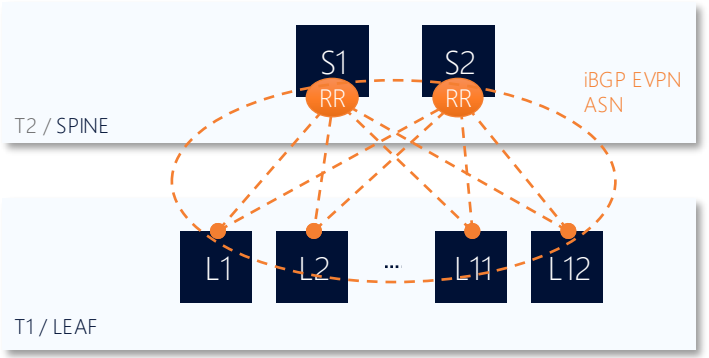
- Workloads (or tenants) networks are implemented as EVPN/VXLAN overlay networks.
- VXLAN and MP-BGP EVPN technologies are leveraged to provide control-plane and data-plane separation for both Layer 2 and Layer 3 forwarding.



Overlay networks

MP-BGP EVPN / VXLAN control plane

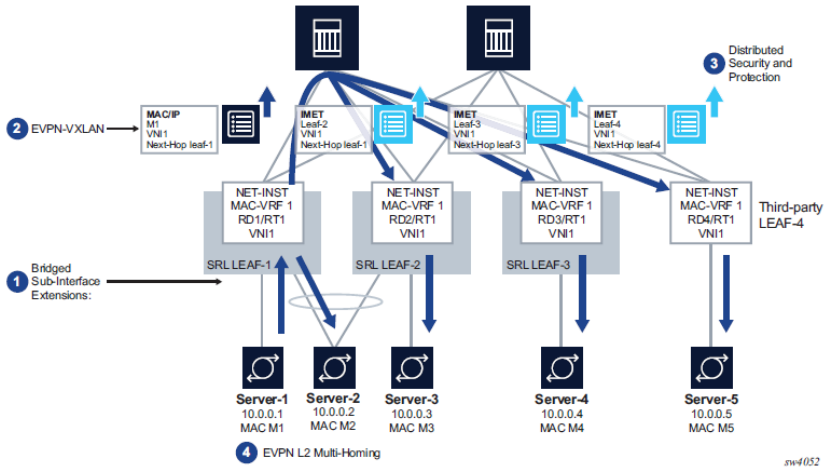
- The EVPN standard (RFC 7432) defines the functionality for delivering multi-tenant Layer 2/3 VPN services using VXLAN encapsulation across a common physical IP infrastructure.
- The EVPN control plane is supported by dedicated iBGP multi-hop sessions between VTEPs (leafs, border leafs) and route reflectors.
- The iBGP EVPN sessions are established between system loopbacks.
- In the 3-Tier & 2-Tier topologies the EVPN route reflector function is typically enabled on the spines. Each RR runs independently.



Overlay networks

L2 EVPN / MAC-VRF

- The EVPN-VXLAN solution supports using **Layer 2 Broadcast Domains (BDs)** in multi-tenant data centers. The primary usage for EVPN VXLAN tunnels (layer 2) is the extension of a BD in overlay multi-tenant DCs.
- The network-instance type mac-vrf functions as a broadcast domain. Each mac-vrf network-instance builds a bridge table composed of MAC addresses that can be learned via the data path on interfaces or via static configuration.
- **MAC duplication is the mechanism used for loop prevention.** MAC duplication monitors MAC addresses that move between sub-interfaces.
- Detection of duplicate MAC addresses is necessary when extending broadcast domains to multiple leaf nodes. Upon detecting a duplicate MAC, the MAC address will not be relearned anymore on this or any sub-interface ("stop-learning" action)



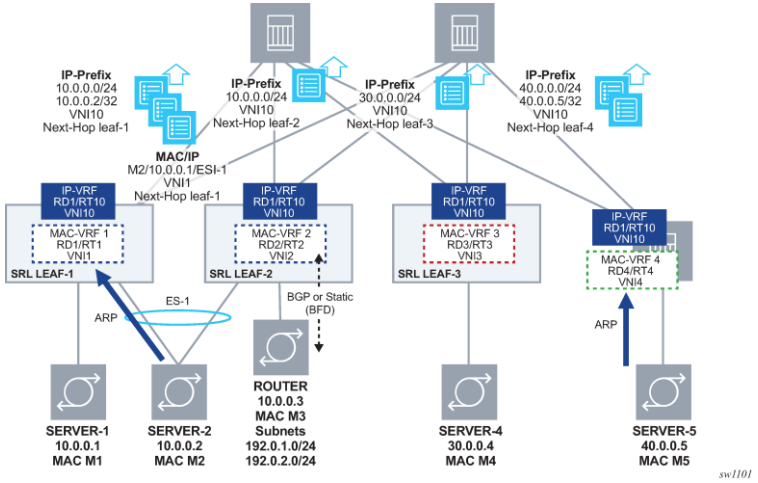
L2 EVPN (MAC-VRF)

Overlay networks

L3 EVPN / IP-VRF

- The EVPN-VXLAN solution supports **L3 EVPN** routing, which provides **connectivity between subnets across multiple Broadcast Domains (BDs)** of the same tenant.
- The L3 EVPN solution implements the **EVPN Interface-less (IFL) model**, based on the advertisement and processing of IP prefixes using **EVPN type 5 routes (RT-5)**.
- All interfaces and local routes in an IP-VRF are automatically advertised in RT5s without the need for any export policy.

.



L3 EVPN (IP-VRF)

Overlay networks

L3 EVPN / PE-CE routing

- Routing mechanism is needed when server workloads need to advertise/receive non connected prefixes to/from the DC leaves or to/from a PE router connected. Two routing mechanisms exist in L3 IP-VRF to route to external workloads:
  - Static Routing: this requires both sides of the routing to know the prefixes behind the next hop - BFD is necessary to avoid blackholing traffic to far end that may not be operational.
  - eBGP (PE-CE): using dynamic routing to allows dynamic prefix exchange and control of advertisements - usually toward PE workloads.

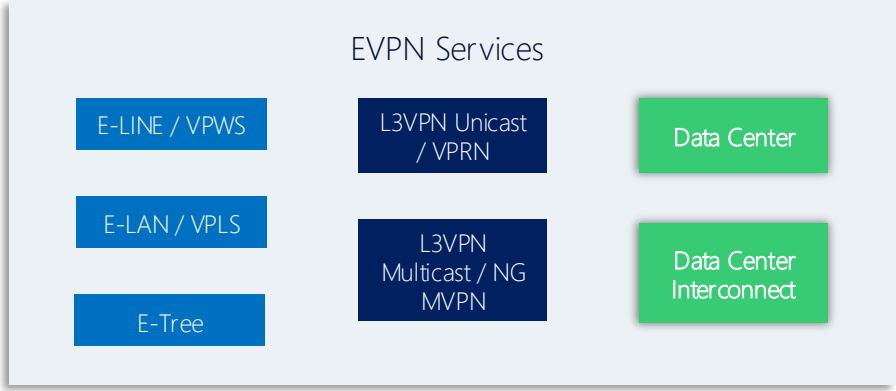
Agenda

1. Data center architectures
2. EVPN basics

What is Ethernet VPN ?

- **VPN services** were traditionally delivered using **a different technology per service type** : for instance, BGP/LDP to deliver VPLS and VPWS, MP-BGP/MPLS to deliver IP VPNs, and BGP/PIM to deliver multicast VPNs. Combined together, these technologies **add complexity and increase the operational costs** for service providers.

- Ethernet virtual private network (**EVPN**) introduces a **unified model** for VPNs and cloud-based services, by providing a **control plane framework** that can deliver **any type of VPN services**.



➢ Specifications for **overlays in data centers** are defined based on RFCs. Notably, they describe how to use EVPN across VXLAN or MPLS tunnels.

For further reference,
RFC 7432 - BGP MPLS-Based Ethernet VPN (ietf.org)
RFC 8365 - A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN) (ietf.org)
RFC 9014 - Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks (ietf.org)

Ethernet Virtual Private Networks RFC7432

EVPN baseline spec as a VPLS replacement

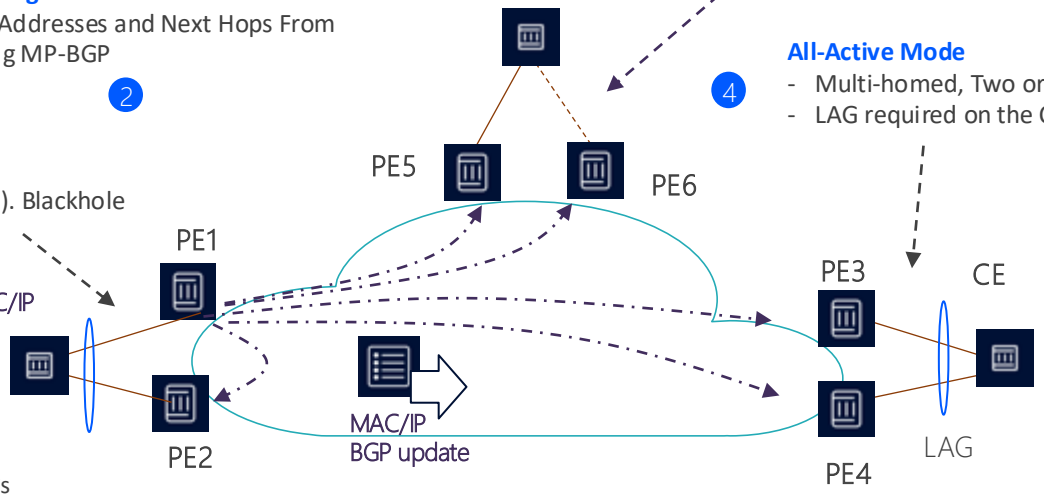
Control Plane Learning
PEs Advertise MAC Addresses and Next Hops From Connected CEs Using MP-BGP

2

Data Plane Learning
Dynamic or Static (Provisioned). Blackhole available.

1

Customer Edge (CE)
Host, VM, Router or Switch



MAC-VRF / BD (Broadcast Domain)
Virtual Routing Forwarding table for MACs
MAC-VRF==BD (if vlan-based/vlan-bundle)

EVPN Instance (EVI)
Group of MAC-VRFs (BDs) forming a VPN

3

Data Plane Encapsulation
Agnostic
MPLS, SR-MPLS, VXLAN, SRv6

Single-Active Mode
- Multi-homed
- One Active PE

4

All-Active Mode
- Multi-homed, Two or More Active PEs
- LAG required on the CE

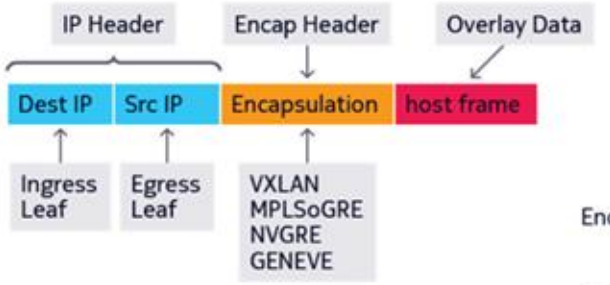
Ethernet Segment (ES) and virtual ES
- Link(s) that Connect the CE to PEs (ESIs are Unique Across the Network)
- An ES on a PE can contain LAG, port, MPLS tunnel, PW, vlan, QinQ tag combinations, VXLAN instance

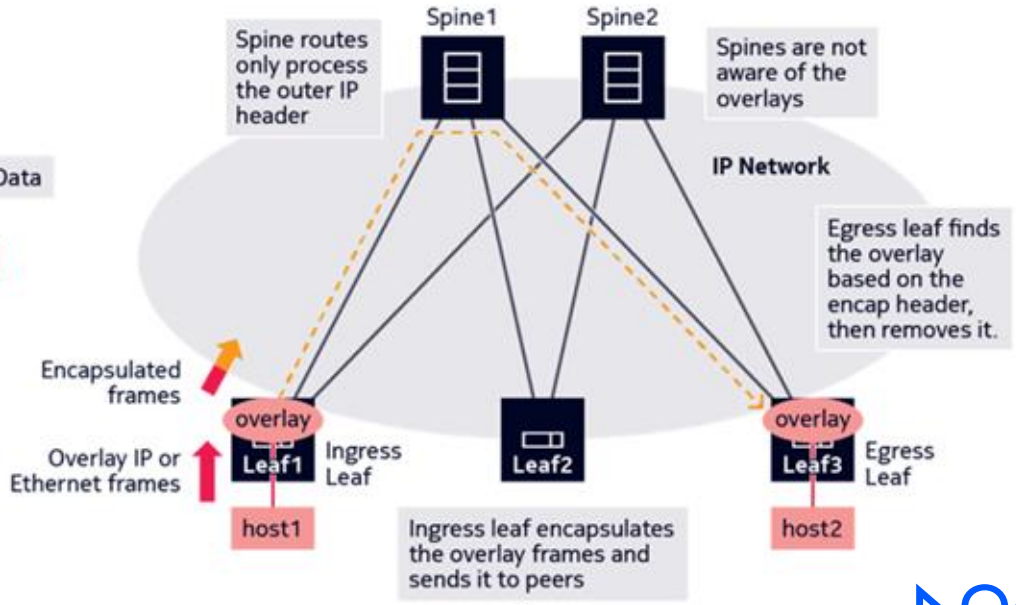
EVPN

Supported data planes

EVPN supports multiple data plane technologies for tunneling overlay networks

- Service Providers prefer MPLS (and PBB for very large-scale networks)
- IP-based encapsulation in the data center:
  - VXLAN is the most popular
  - Also MPLSoGRE, NVGRE, GENEVE





EVPN

VXLAN in data centers

VXLAN - Virtual eXtensible Local Area Network

- VXLAN is a Layer 2 overlay using an existing Layer 3 network infrastructure (underlay)
- VXLAN is defined in RFC7348
- Was in use well before EVPN

VXLAN became a de-facto standard in data center networking

- Allows the creation of L2 overlay networks that span the whole data center:
  - Scale up to 16M tenants (vs 4K with VLANs), isolating each overlay from each other
  - Seamless VM mobility within the data center
- Leverages the underlay IP networks:
  - To avoid loops (no more Spanning Tree Protocol)
  - To provide efficient multi-path load-balancing with ECMP
- However, RFC 7348 does not specify a control plane:
  - VXLAN uses Flood-and-Learn in the data plane
  - Requires static configuration to learn about other VXLAN endpoints

EVPN

VXLAN terminology

VXLAN tunnels:

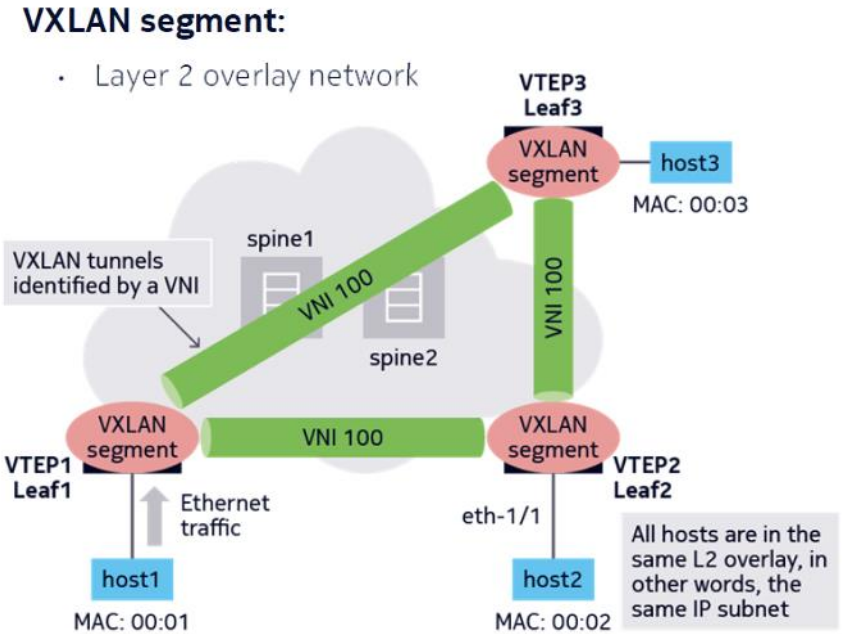
- To implement an overlay network, traffic is encapsulated with an extra header identifying the overlay. VXLAN is one of such encapsulation techniques.
- Encapsulated traffic flows between two end-points over the underlay network - hence the name 'tunnel'.

VXLAN Tunnel End Point (VTEP):

- Egress and Ingress point of the encapsulated traffic.
- Typically, the leaf routers
- The spine routers are not aware of the VLAN tunnels

VXLAN Network Identifier (VNI):

- 24-bit integer uniquely identifying a VLAN segment network-wide



EVPN

VXLAN-encapsulated frame

VLAN encapsulates Ethernet in IP:

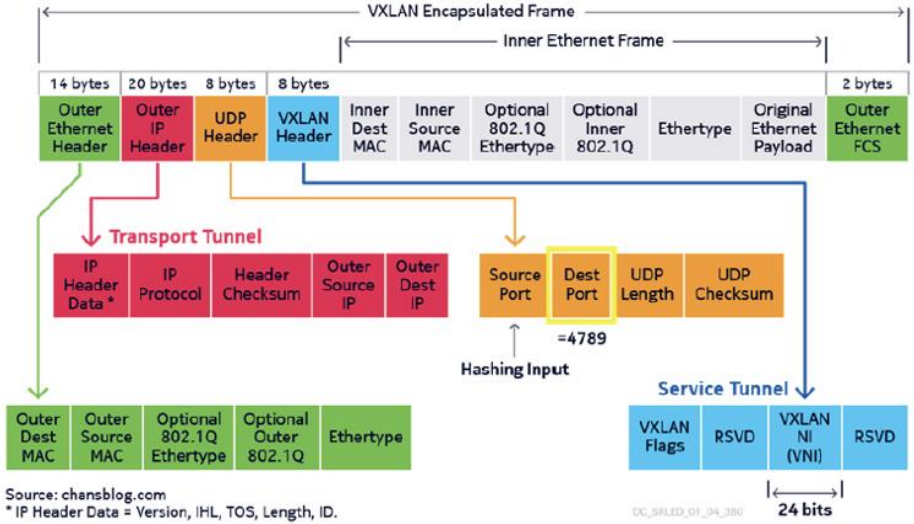
- UDP-based (port 4789)
- UDP source port is a hash of the original payload MAC, IPs, and ports to provide flow-based load balancing entropy

8-byte VXLAN header:

- VXLAN Network Identifier (VNI) 24-bits
- Allows for 16M overlays

Since VLAN is IP-based, it can be routed over any IP network:

- Enable ECMP for load-balancing
- Network must support the 50-byte encapsulation overhead



Why is EVPN used in modern DCs?

RFC8365

Modern DCs are based on:

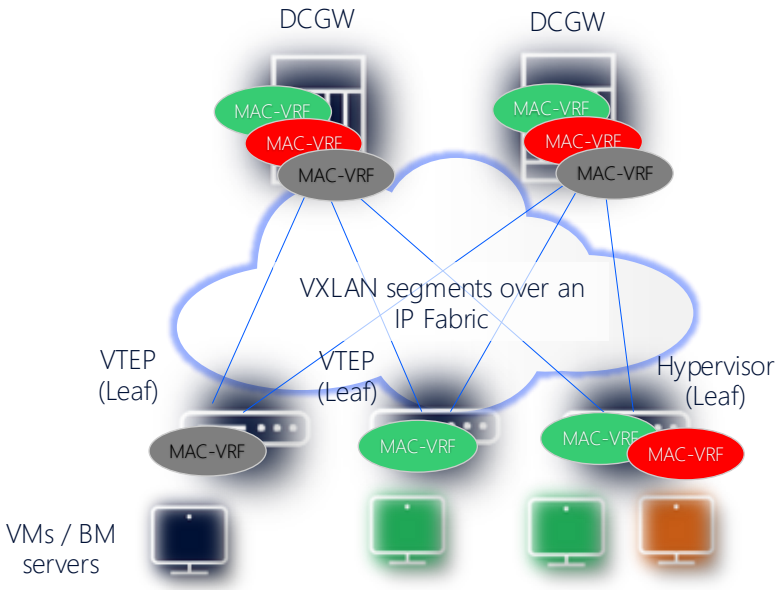
- CLOS architecture and IP Fabrics
  - No loops, no flooding, fast convergence
  - ECMP
- Multi-tenancy with intra (L2) and inter-subnet (L3) connectivity
- IP overlay tunnels are therefore needed (VXLAN is the most popular option)

Why do I need a control plane?

- Auto-discovery of the remote VTEPs
- Distribution of MAC/IP information in order to reduce/suppress flooding

What are my options?

- PIM, but…
  - Requires tenant states in the core (SG, *G)
  - Does not reduce/suppress broadcast/multicast
  - Does not do inter-subnet-forwarding
- EVPN
  - YES!



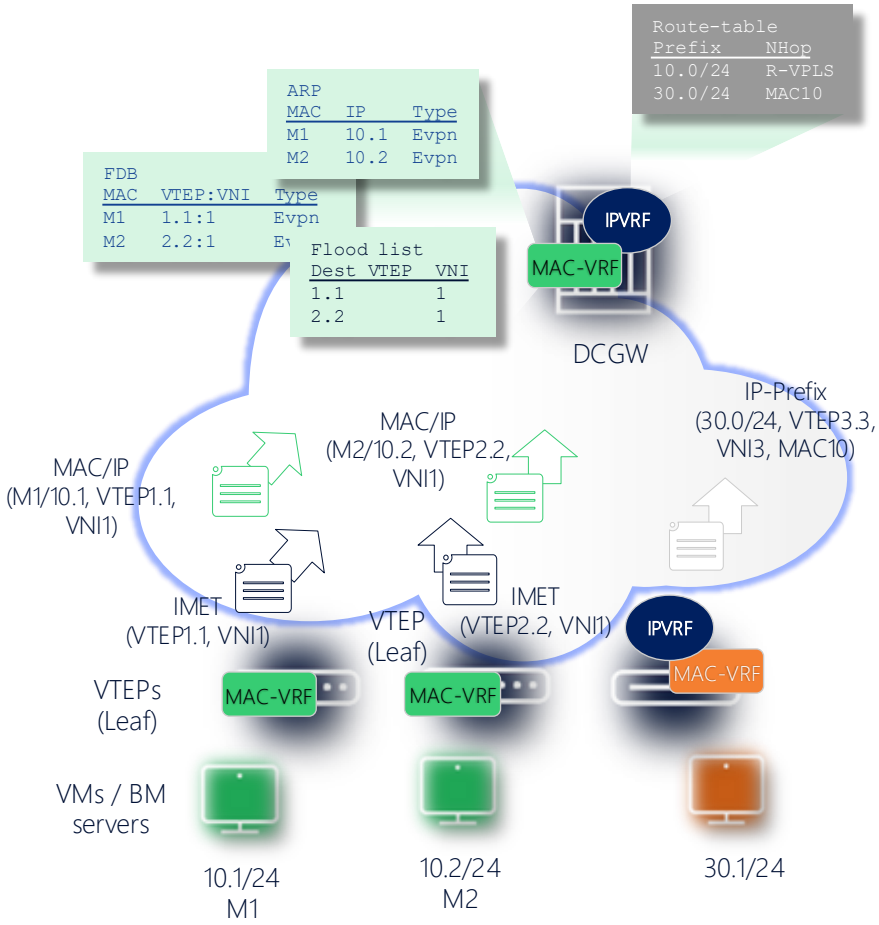
EVPN for VXLAN tunnels

EVPN provides the basic Control Plane needs…

- Auto-discovery of remote VTEPs through Inclusive Multicast Routes (IMET)
- Distribution of MAC/IP information in order to reduce/suppress flooding through MAC/IP routes

But also advanced options

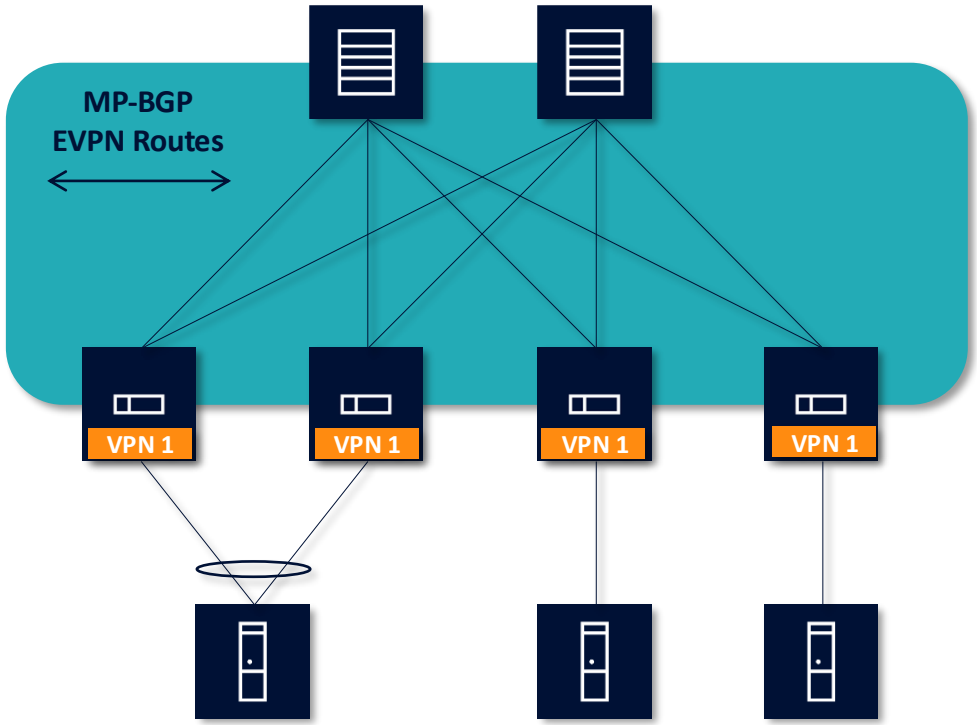
- Distribution of IP-Prefixes so that not all the subnets must be defined in all the NVEs (through IP-Prefix type 5 routes)
- MAC/IP mobility
- MAC protection, duplication detection, loop protection
- Proxy-ARP/ND
- Assisted-Replication
- Service-chaining



Control plane

EVPN in data centers

- EVPN relies on **MP-BGP** in the service provider network, i.e. the data center fabric.
- Several EVPN route types are defined:
  - EVPN routes are **exchanged between leafs**, via a route reflector or not.
  - Route types advertised are based on the **use case** and the **type of service** being delivered.



Control plane – EVPN route types

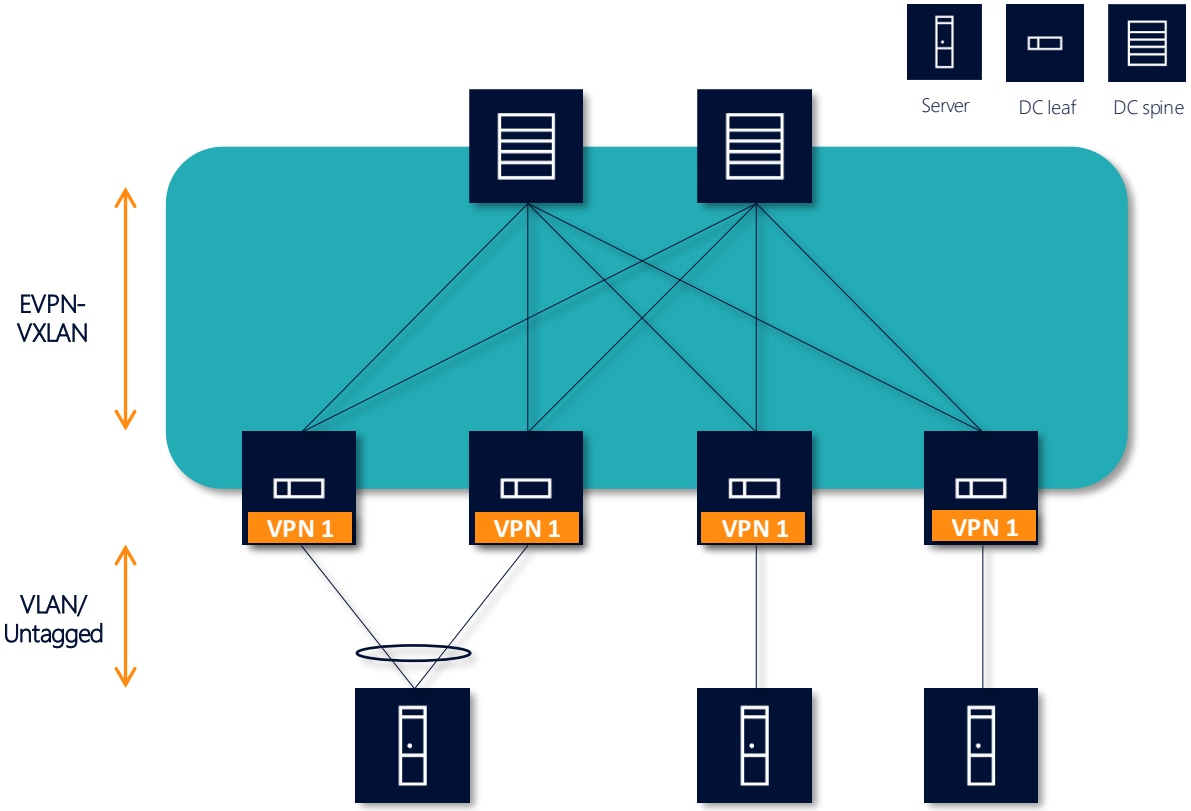
EVPN in data centers

Route type	Route name	Purpose
1	Ethernet Auto-Discover (A-D)	Used in multi-homing scenarios to support aliasing and fast convergence
2	MAC/IP Advertisement	Used to advertise host MAC address or host MAC/IP addresses
3	Inclusive Multicast Ethernet Tag (IMET)	Used to discover member PEs and to setup the flooding tree for BUM traffic
4	Ethernet Segment (ES)	Used in multi-homing scenarios to suppport Ethernet segment discovery and DF election
5	IP-Prefix	Used to advertise IP prefixes for inter-subnet connectivity in L3VPN services

Data plane

EVPN in data centers

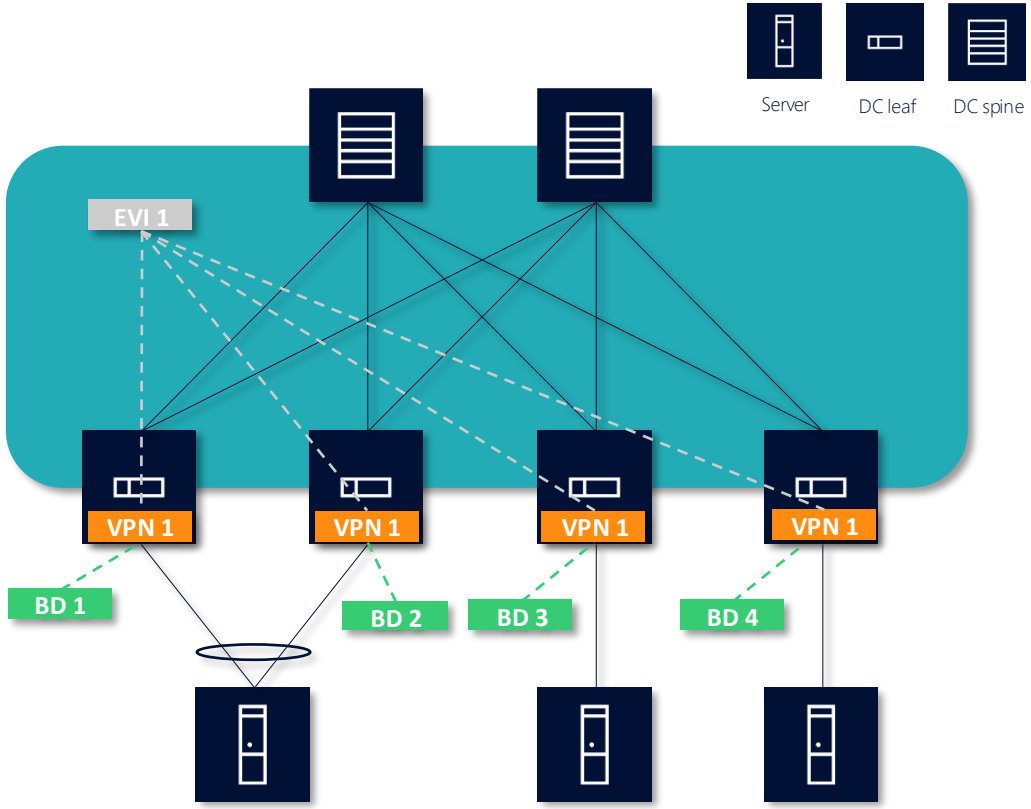
- EVPN allows the delivery of multiple services over a single core network
- Leafs encapsulates customer data with a **label that uniquely identifies each service**
- Encapsulated data is **tunnelled between leafs** (or border leafs, or DCGWs).
- **VXLAN** tunnels are used within the fabric.
- **MPLS** tunnels are used within IP/MPLS networks.



Layer-2 services - terminology

EVPN in data centers

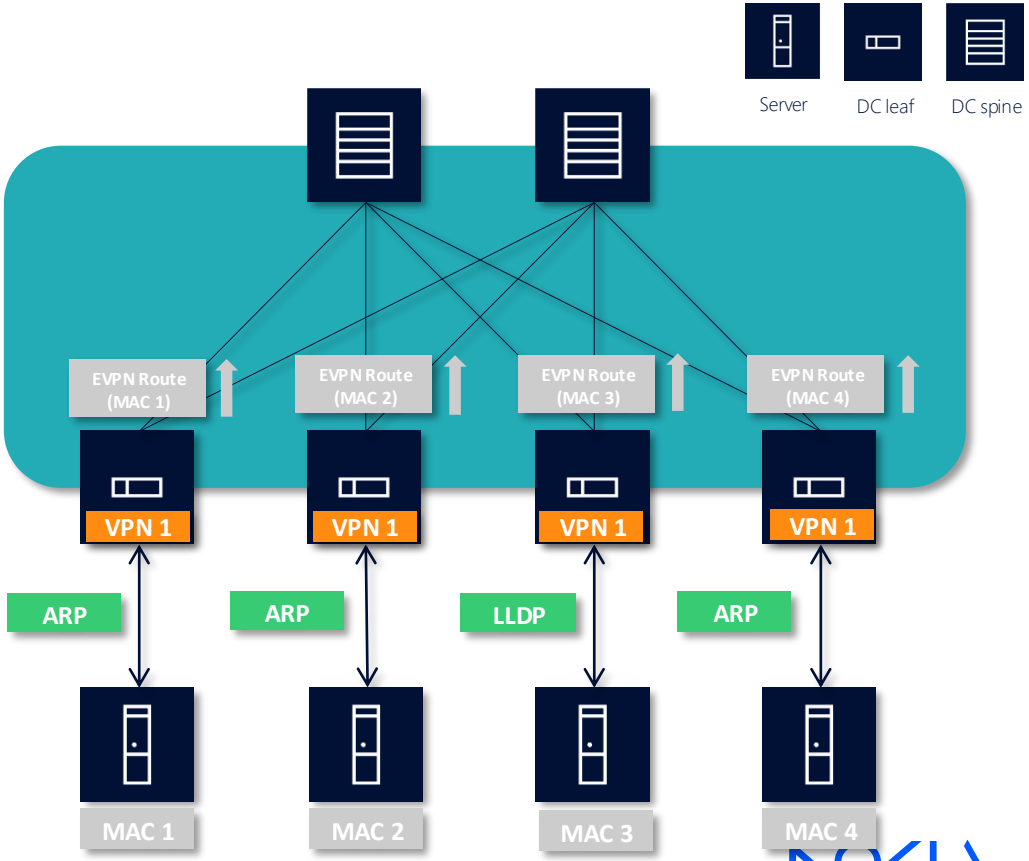
- Broadcast domain (BD)
  - An instantiation of an EVPN service on a given leaf
- MAC-VRF
  - The virtual routing and forwarding (VRF) table that contains the MAC addresses for an EVPN service
- EVPN instance (EVI)
  - The group of BDs that are part of the same EVPN service



Layer-2 services – EVPN operation

EVPN in data centers

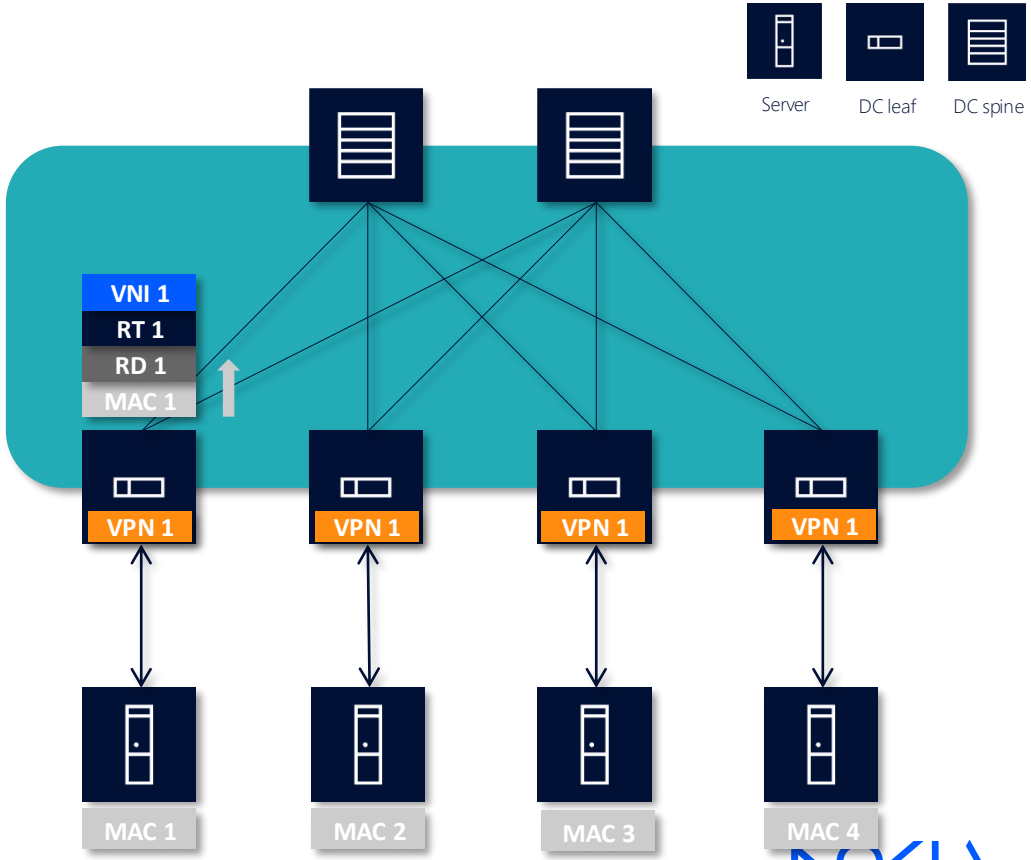
- EVPN enables control-plane MAC learning in the core
- Leafs exchange EVPN routes over **MP-BGP** to **advertise local MAC addresses** across the fabric
- Attached clients can be physical or virtual
- Leafs learn MAC addresses of locally-connected clients using data-plane learning, static provisioning or control-plane protocols.



Layer-2 services – leaf-to-leaf MAC address advertisement

EVPN in data centers

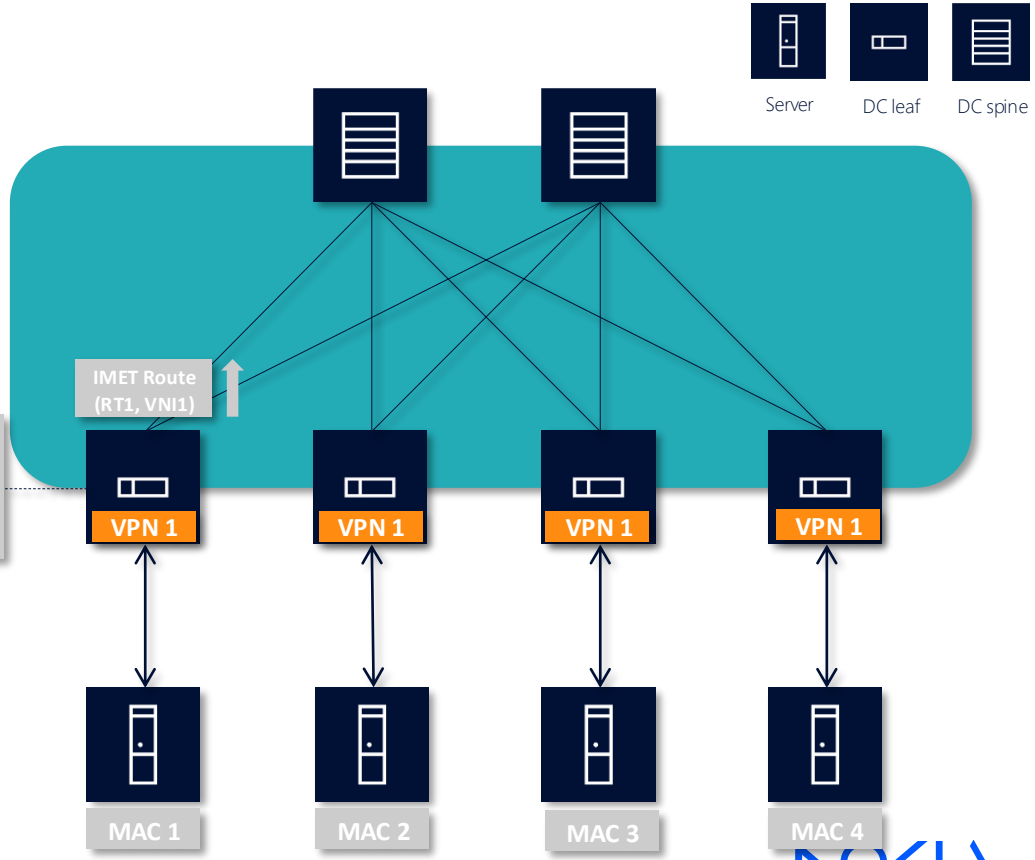
- Leafs advertise **locally-learned MAC addresses** using **MAC/IP routes** (EVPN route type 2)
- A single MP-BGP instance handles the exchange of routes for all EVIs on the leaf
- **Route distinguisher** is used to distinguish routes between EVIs in case of overlaps. **Route targets** identify which EVPN routes are to be installed in the local MAC-VRF
- Provided label (i.e. **VxLAN network identifier**) is used by remote leafs when encapsulating frames destined to the advertised MAC address



Layer-2 services – flooding lists & BUM traffic handling

EVPN in data centers

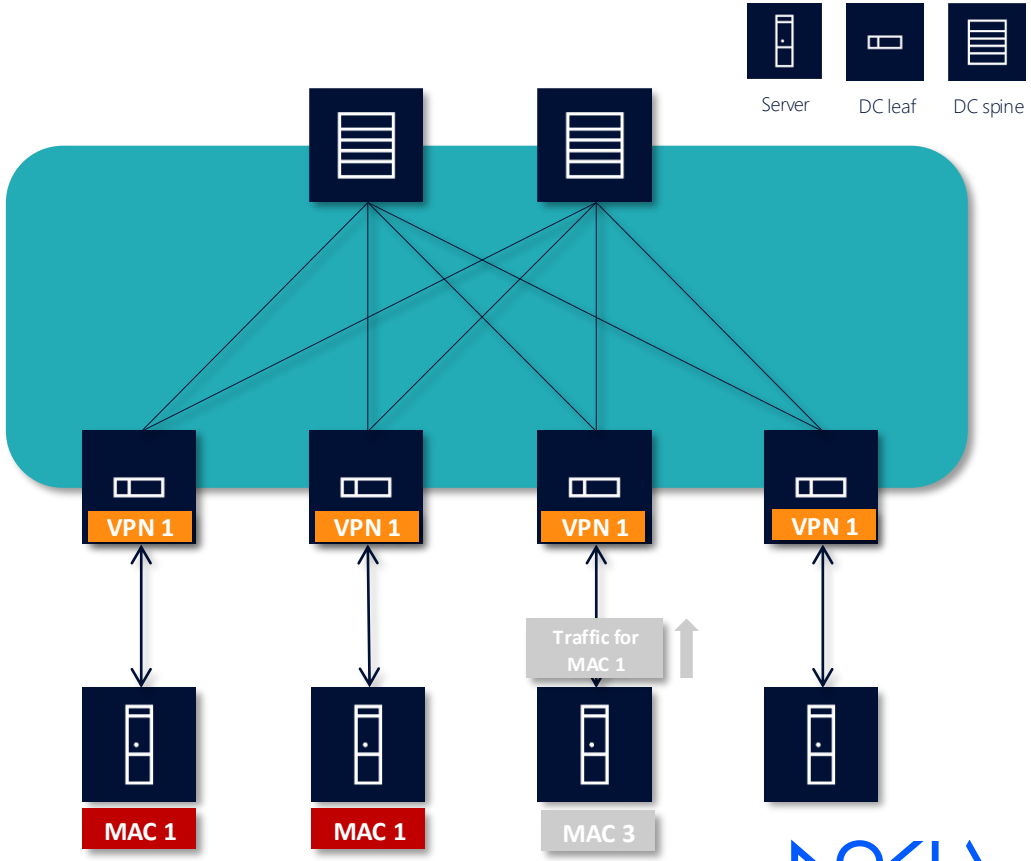
- When an EVPN service is enabled, leafs participating in this service exchange **inclusive multicast Ethernet tag routes** (**IMET** or type 3)
  - Discover all leafs attached to the same EVI
  - Build a flooding list in each leaf
- A leaf receiving a **BUM frame** from a client, **consults the flooding list** of the EVPN service to determine the leafs to which it needs to flood the frame
- A leaf receives this encapsulated BUM frame from the fabric, **decapsulates** the packet and then **floods it to its local interfaces**
  - Due to **split-horizon**, the BUM frames are never flooded back to the fabric



Layer-2 services – dealing with layer 2 loops

EVPN in data centers

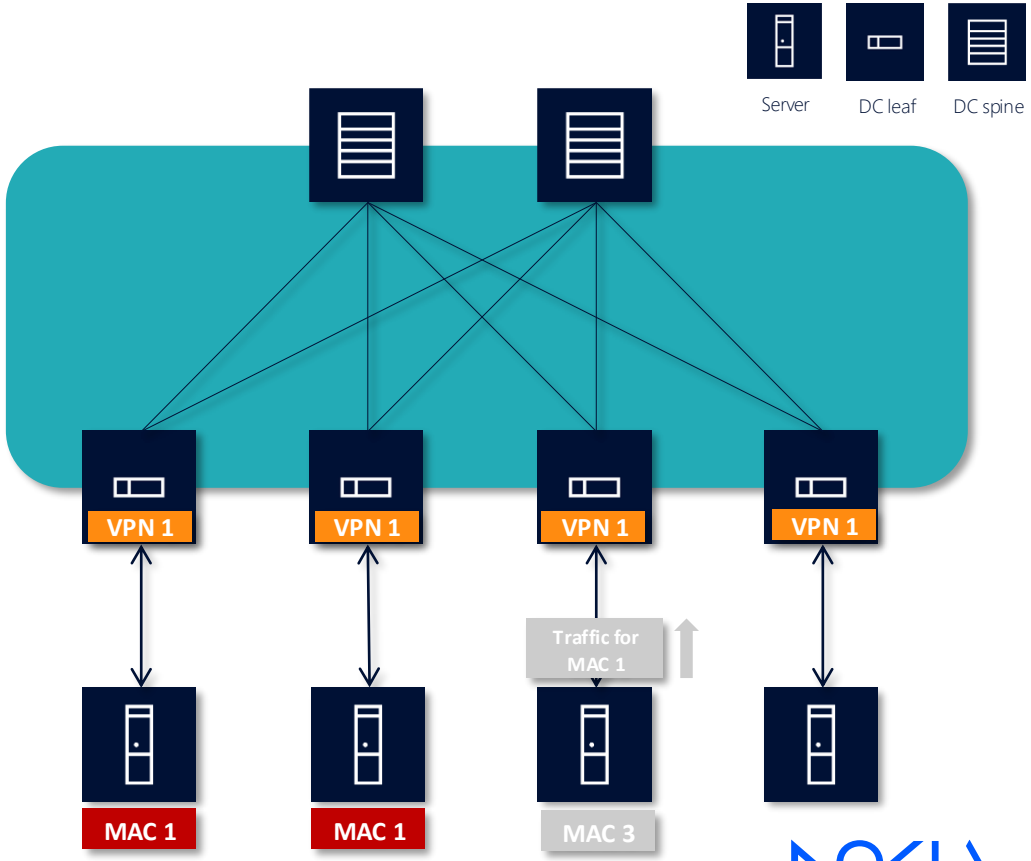
- The detection of duplicate MAC addresses and loops is a fundamental feature when extending a broadcast domain
- L2 loops or duplicate MACs are typically due to a configuration mistake or an intended spoofing attack.
- **MAC duplication** is the mechanism used by for **loop prevention**. MAC duplication **monitors MAC addresses that move between subinterfaces**. It consists of detection, actions, and process restart.



Layer-2 services – MAC-duplication as loop protection mechanism

EVPN in data centers

- A MAC is declared **duplicate** if it is learnt on **different interfaces** and **the number of moves is higher than a certain value** (num-moves) within a certain interval (monitoring-window).
- A **configurable action** can be performed on the subinterface when a duplicate MAC is detected. One of three options can be selected (stop-learning, blackhole, oper-down).
- The **MAC remains "duplicate"** for the duration of the **hold-down-time** parameter. At the end of that interval, it is flushed from the bridge table and the action on the subinterface is cleared.




NOKIA