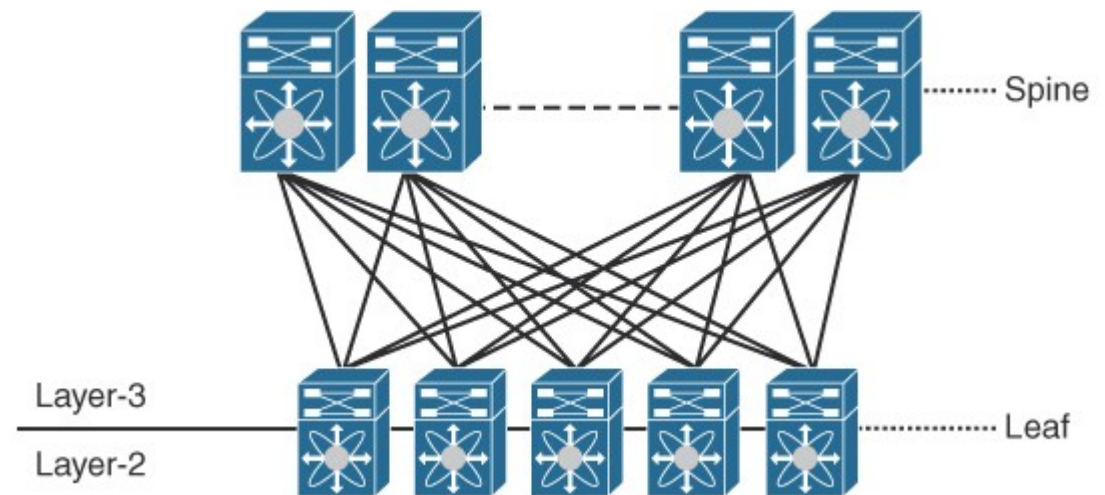
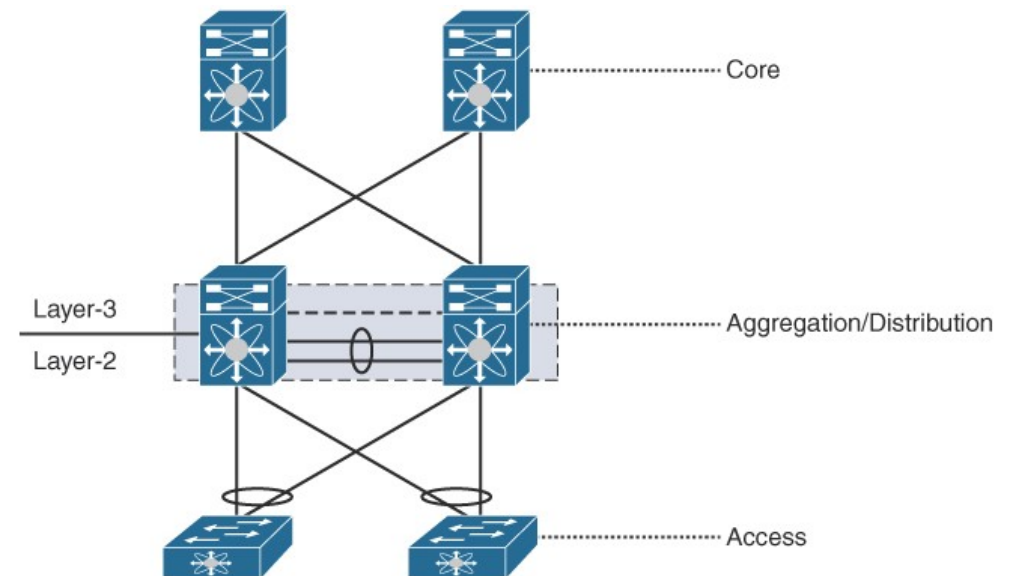


Layer 2 VPN

VXLAN and BGP EVPN

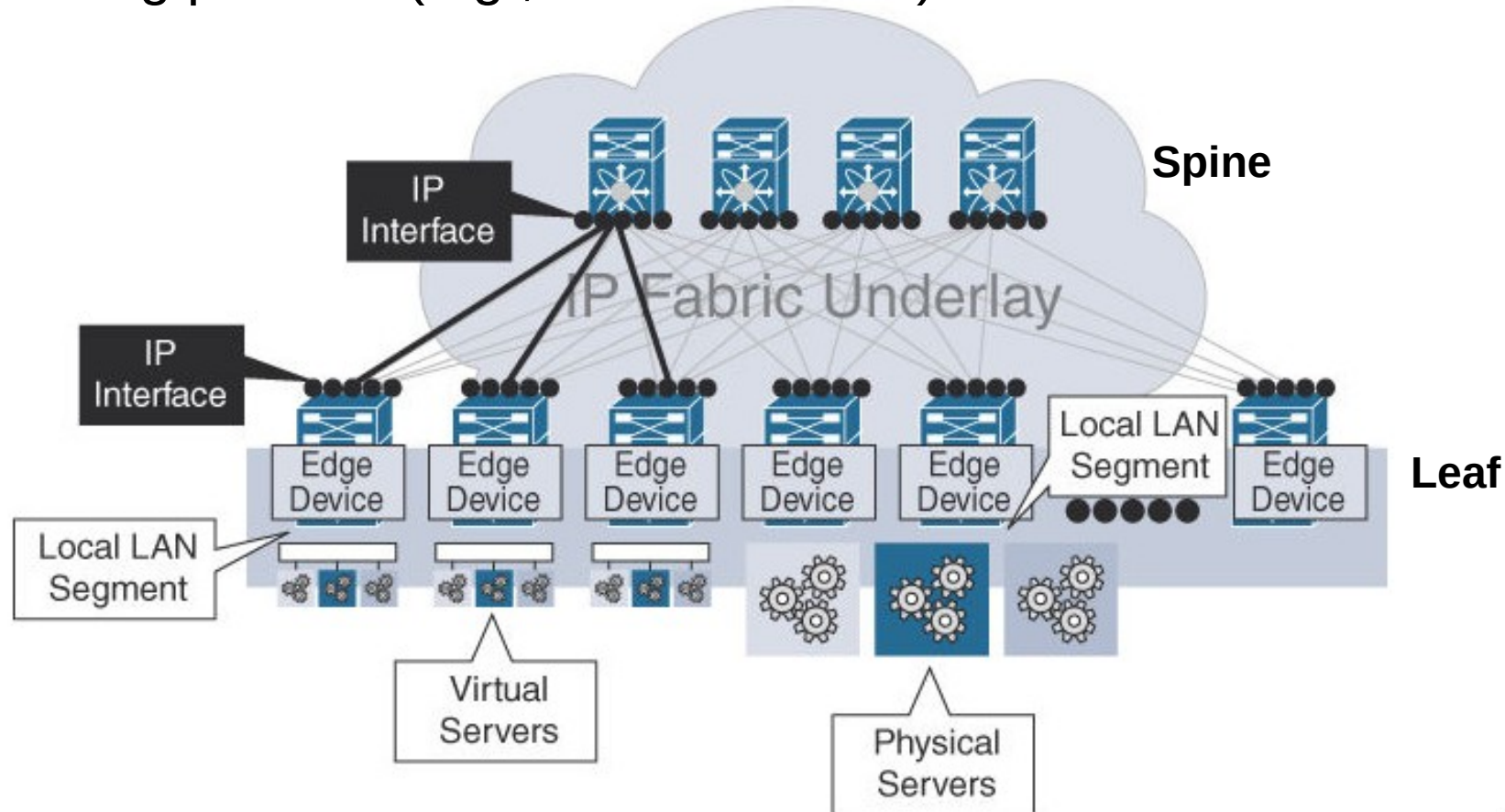
Datacenter CLOS Topology

- With large-scale data center deployments, three-tier topologies have become scale bottlenecks.
- The classic three-tier topology evolved to a CLOS topology.
 - Original designed by Charles Clos in 1950 to find a more efficient way to handle telephonic call transfers.
- Eliminating the need for STP the network evolved to greater stability and scalability.
- Layer 3 moves to the Access Layer.
- Usually called Spine-and-Leaf Architecture.



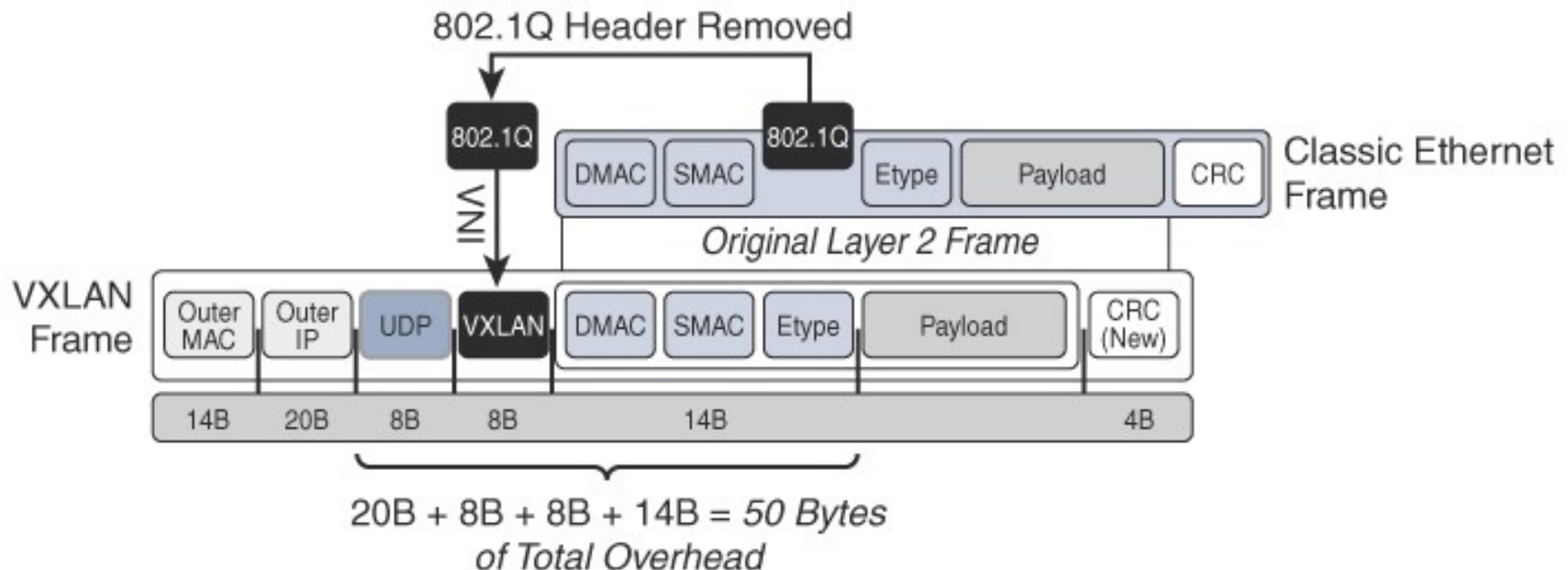
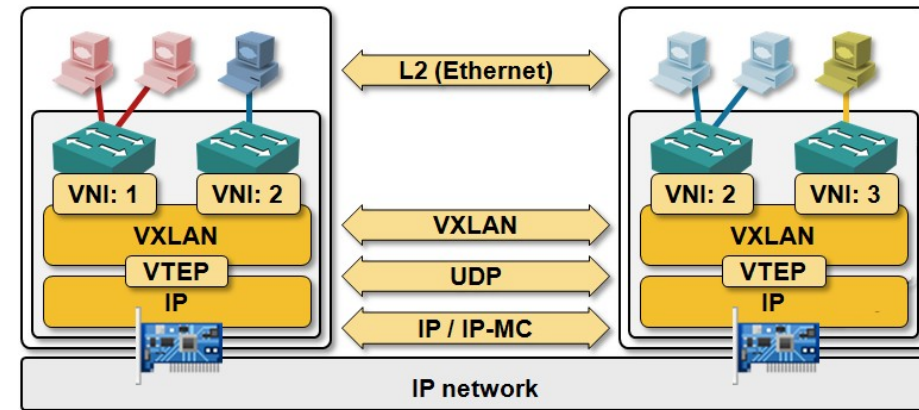
Spine-and-Leaf Architecture

- The access layer with Layer 3 support is typically called the Leaf layer.
- The aggregation layer that provides the interconnection between the various leafs is called the Spine layer.
- The IP underlay transport between Spines and Leaves requires an IGP routing protocol (e.g., OSPF or IS-IS).

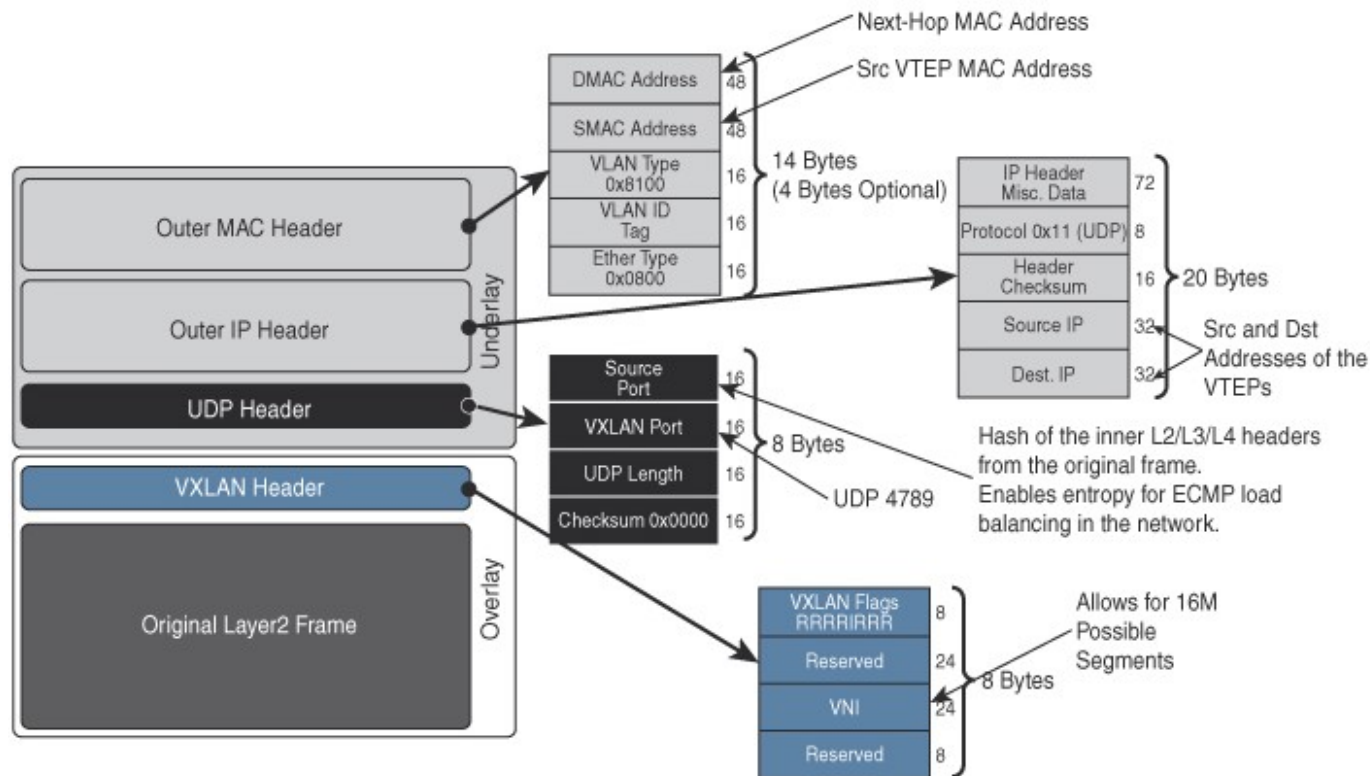


Virtual Extensible LAN (VXLAN)

- Encapsulates OSI Layer 2 Ethernet frames within Layer 4 UDP/IP datagrams .
 - ◆ Default port 4789.
- VLAN may be additionally identified by a **VNI field** with 24 bits.
 - ◆ 802.1Q tag only has 12 bits.
- The original inner 802.1Q header of the Layer 2 Ethernet frame is removed and mapped to a VNI to complete the VXLAN header.



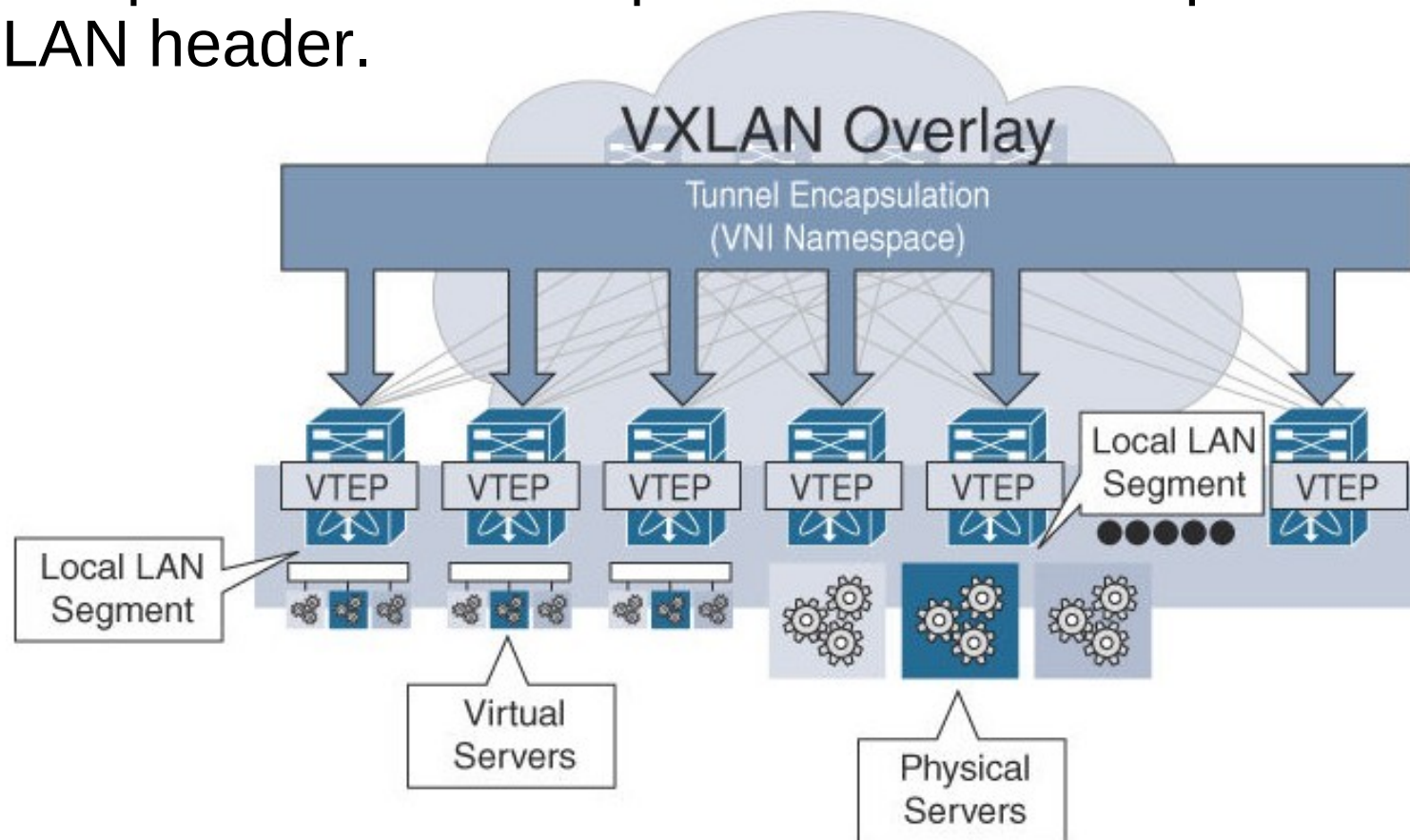
VXLAN Header/Packet



- › Ethernet II, Src: ca:01:25:92:00:08 (ca:01:25:92:00:08), Dst: 0c:32:45:6d:00:00 (0c:32:45:6d:00:00)
- › Internet Protocol Version 4, Src: 192.0.0.3, Dst: 192.0.0.1
- › User Datagram Protocol, Src Port: 56255, Dst Port: 8472
- › Virtual eXtensible Local Area Network
 - › Flags: 0x0800, VXLAN Network ID (VNI)
 - Group Policy ID: 0
 - VXLAN Network Identifier (VNI): 101
 - Reserved: 0
- › Ethernet II, Src: 0c:88:63:63:00:01 (0c:88:63:63:00:01), Dst: Private_66:68:00 (00:50:79:66:68:00)
- › Internet Protocol Version 4, Src: 10.1.3.100, Dst: 10.1.1.100
- › Internet Control Message Protocol

VTEP (VXLAN Tunnel Endpoint)

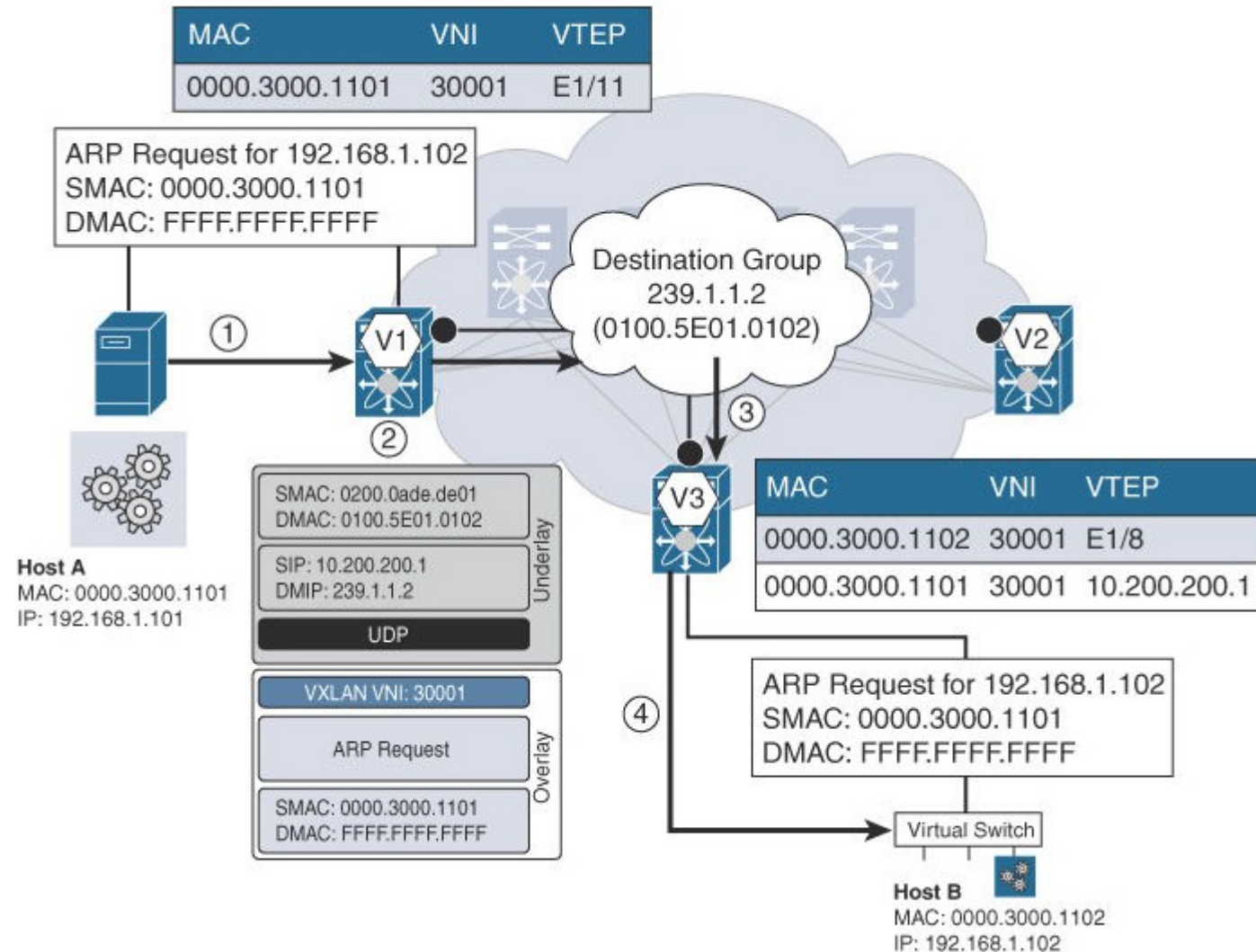
- The edge devices in a VXLAN network have the VXLAN Tunnel Endpoints (VTEP).
- Are responsible for encapsulation and decapsulation of the VXLAN header.




VTEP: VXLAN Tunnel Endpoint
VNI/VNID: VXLAN Network Identifier


VXLAN Flood and Learn


- The multidestination traffic is flooded over the VXLAN between VTEPs.
 - To learn about the host MACs located behind the VTEPs so that subsequent traffic can be unicast.
 - This is referred to as an F&L mechanism.
- A native F&L based approach is far from optimal since the broadcast domain for a VXLAN now spans Layer 3 boundaries.



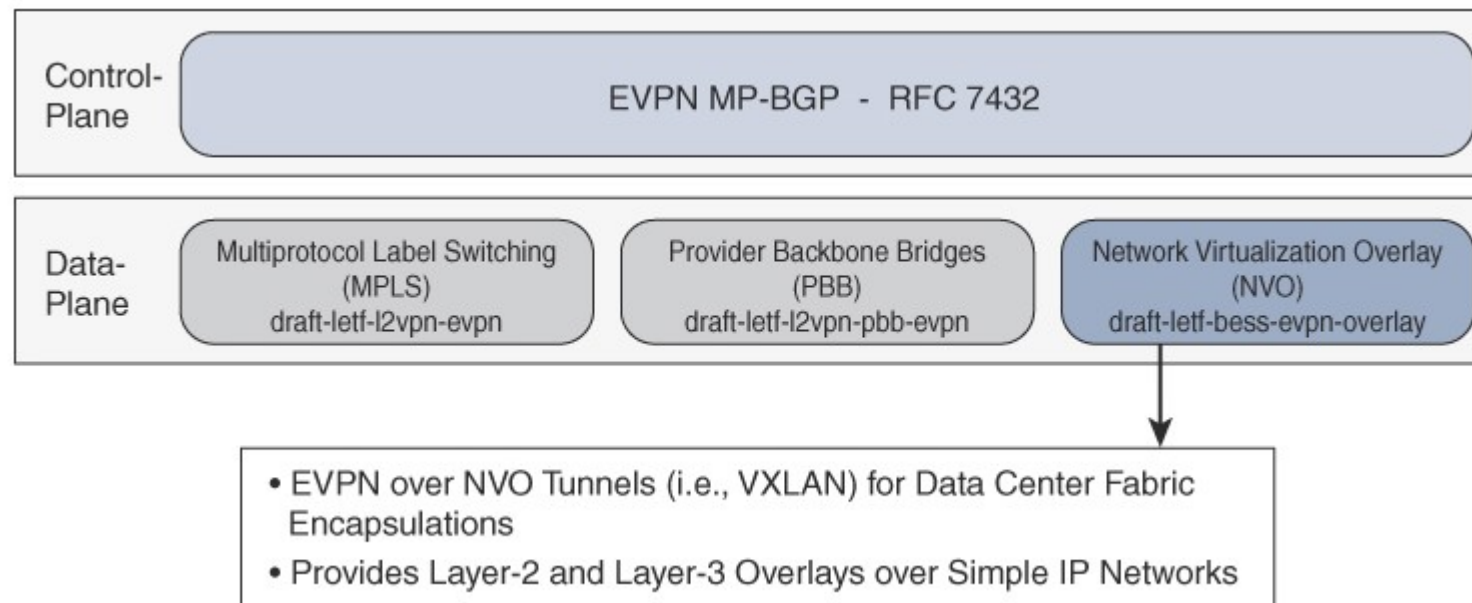
EVPN MP-BGP

 To mitigate the VXLAN Flood and Learn problem, it was introduced the concept of Ethernet VPN with MP-BGP, provided by the Address family L2VPN EVPN.

 The Address family L2VPN EVPN provides a method to transport VPN-aware Layer 2 (MAC) and Layer 3 (IP) information across a single MP-BGP peering session.

 The EVPN MP-BGP RFC allow for multiple data plane transport: MPLS, PBB and NVO.

- ◆ A possible (and more common) EVPN over NVO solution for datacenters is VXLAN.



BGP EVPN wih VXLAN



BGP is used to announce and learn remote VTEP addresses.



VXLAN is used to transport to the specific remote VTEP where the destination device is.

- BGP relations can be:

- Only internal BGP.

- To avoid a full BGP mesh, Route Reflectors should be used (usually all or some of the spines).



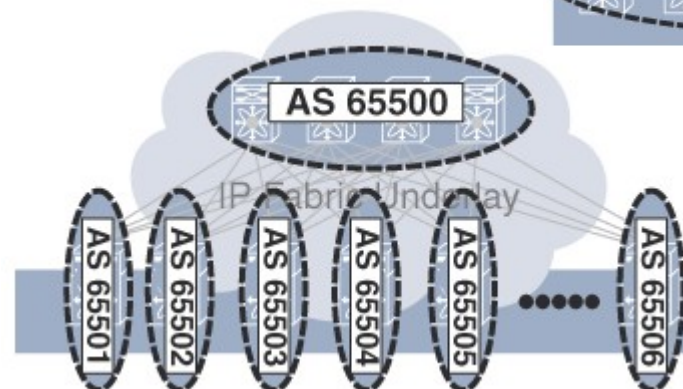
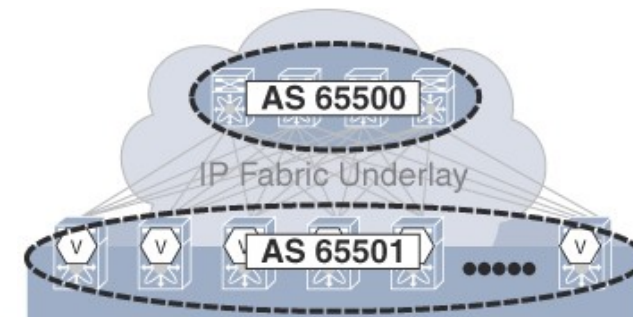
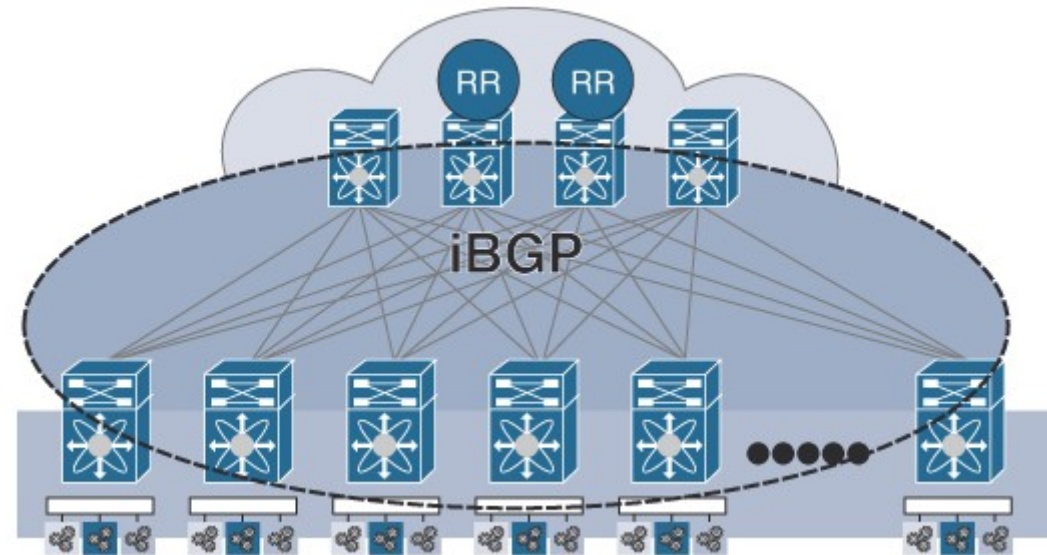
- External BGP relations between private AS.



- Leaf in a private AS.



- Each Leaf is a private AS.



EVPN Route types

- Route Type-2



Defines the MAC/IP advertisement route.

Responsible for the distribution of MAC and IP address reachability information.

- Route Type-3



- Called “inclusive multicast Ethernet tag route”.

- Used to create the a distribution list for unknown unicast, multicast and broadcast packets (ingress replication).

- Provides a way to replicate multideestination traffic in a unicast.

| |
|---------------------------------|
| RD (8 Octets) |
| ESI (10 Octets) |
| Ethernet Tag ID (4 Octets) |
| MAC Address Length (1 Octet) |
| MAC Address (6 Octets) |
| IP Address Length (1 Octet) |
| IP Address (0, 4, or 16 Octets) |
| MPLS Label1 (3 Octets) |
| MPLS Label2 (0 or 3 Octets) |

| |
|--|
| RD (8 Octets) |
| ESI (10 Octets) |
| Ethernet Tag ID (4 Octets) |
| IP Address Length (1 Octet) |
| Originating Router's IP Address (4 or 16 Octets) |



EVPN Route Type-2

```
Border Gateway Protocol - UPDATE Message
Marker: ffffffffffffffffffffffffffffffff
Length: 144
Type: UPDATE Message (2)
Withdrawn Routes Length: 0
Total Path Attribute Length: 121
Path attributes
  Path Attribute - MP_REACH_NLRI
    Flags: 0x90, Optional, Extended-Length, Non-transitive, Complete
    Type Code: MP_REACH_NLRI (14)
    Length: 83
    Address family identifier (AFI): Layer-2 VPN (25)
    Subsequent address family identifier (SAFI): EVPN (70)
    Next hop: 192.0.0.3
    Number of Subnetwork points of attachment (SNPA): 0
  Network Layer Reachability Information (NLRI)
    EVPN NLRI: MAC Advertisement Route
      Route Type: MAC Advertisement Route (2)
      Length: 33
      Route Distinguisher: 0001c00000030003 (192.0.0.3:3)
      ESI: 00:00:00:00:00:00:00:00
      Ethernet Tag ID: 0
      MAC Address Length: 48
      MAC Address: Private_66:68:02 (00:50:79:66:68:02)
      IP Address Length: 0
      IP Address: NOT INCLUDED
      VNI: 101
    EVPN NLRI: MAC Advertisement Route
      Route Type: MAC Advertisement Route (2)
      Length: 37
      Route Distinguisher: 0001c00000030003 (192.0.0.3:3)
      ESI: 00:00:00:00:00:00:00:00
      Ethernet Tag ID: 0
      MAC Address Length: 48
      MAC Address: Private_66:68:02 (00:50:79:66:68:02)
      IP Address Length: 32
      IPv4 address: 10.1.3.100
      VNI: 101
  Path Attribute - ORIGIN: IGP
  Path Attribute - AS_PATH: empty
  Path Attribute - LOCAL_PREF: 100
  Path Attribute - EXTENDED_COMMUNITIES
    Flags: 0xc0, Optional, Transitive, Complete
    Type Code: EXTENDED_COMMUNITIES (16)
    Length: 16
    Carried extended communities: (2 communities)
      Encapsulation: VXLAN Encapsulation [Transitive Opaque]
      Route Target: 100:101 [Transitive 2-Octet AS-Specific]
```

- Announces a MAC address and respective IP address of a remote device.



- And respective next-hop.

- EXTENDED_COMMUNITY attribute is used to announce the type of encapsulation and the route target.



- Sent when Leaf device learns a new MAC address.



EVPN Route Type-3

```

- Border Gateway Protocol - UPDATE Message
  Marker: ffffffffffffffffffffffffffffffffff
  Length: 122
  Type: UPDATE Message (2)
  Withdrawn Routes Length: 0
  Total Path Attribute Length: 99
- Path attributes
  - Path Attribute - MP_REACH_NLRI
    - Flags: 0x90, Optional, Extended-Length, Non-transitive, Complete
    Type Code: MP_REACH_NLRI (14)
    Length: 28
    Address family identifier (AFI): Layer-2 VPN (25)
    Subsequent address family identifier (SAFI): EVPN (70)
  - Next hop: 192.0.0.2
    Number of Subnetwork points of attachment (SNPA): 0
  - Network Layer Reachability Information (NLRI)
    - EVPN NLRI: Inclusive Multicast Route
      Route Type: Inclusive Multicast Route (3)
      Length: 17
      Route Distinguisher: 0001c00000020002 (192.0.0.2:2)
      Ethernet Tag ID: 0
      IP Address Length: 32
      IPv4 address: 192.0.0.2
  - Path Attribute - ORIGIN: IGP
  - Path Attribute - AS_PATH: empty
  - Path Attribute - MULTI_EXIT_DISC: 0
  - Path Attribute - LOCAL_PREF: 100
  - Path Attribute - ORIGINATOR_ID: 192.0.0.2
  - Path Attribute - CLUSTER_LIST: 192.0.0.1
  - Path Attribute - EXTENDED_COMMUNITIES
  - Path Attribute - PMSI_TUNNEL_ATTRIBUTE
    - Flags: 0xc0, Optional, Transitive, Complete
    Type Code: PMSI_TUNNEL_ATTRIBUTE (22)
    Length: 9
    Flags: 0
    Tunnel Type: Ingress Replication (6)
    VNI: 102
  - Tunnel ID: tunnel end point -> 192.0.0.2

```

- Defines the next hop for unknown unicast, multicast and broadcast.



- Must also carry a Provider Multicast Service Interface (PMSI) Tunnel attribute.



- Defines tunnel type.
- For EVPN with VXLAN the tunnel type is “Ingress Replication”.

- Sent when a new Leaf (BGP peer) is added.



BGP Route Table – L2VPN EVPN

```
# show bgp l2vpn evpn
```

```
BGP table version is 1, local router ID is 192.0.0.1
```

```
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
```

```
Origin codes: i - IGP, e - EGP, ? - incomplete
```

```
EVPN type-1 prefix: [1]:[EthTag]:[ESI]:[IPlen]:[VTEP-IP]:[Frag-id]
```

```
EVPN type-2 prefix: [2]:[EthTag]:[MAClen]:[MAC]:[IPlen]:[IP]
```

```
EVPN type-3 prefix: [3]:[EthTag]:[IPlen]:[OrigIP]
```

```
EVPN type-4 prefix: [4]:[ESI]:[IPlen]:[OrigIP]
```

```
EVPN type-5 prefix: [5]:[EthTag]:[IPlen]:[IP]
```

| Network | Next Hop | Metric | LocPrf | Weight | Path |
|---|-----------------|--------|--------|--------|------|
| Route Distinguisher: 192.0.0.1:2 | | | | | |
| *> [3]:[0]:[32]:[192.0.0.1] | 192.0.0.1 | | | 32768 | i |
| | ET:8 RT:100:102 | | | | |
| ... | | | | | |
| *>i[2]:[0]:[48]:[00:50:79:66:68:01] | 192.0.0.2 | 100 | | 0 | i |
| | RT:100:101 ET:8 | | | | |
| ... | | | | | |
| *>i[2]:[0]:[48]:[00:50:79:66:68:02]:[32]:[10.1.3.100] | 192.0.0.3 | 100 | | 0 | i |
| | RT:100:101 ET:8 | | | | |
| ... | | | | | |
| *>i[3]:[0]:[32]:[192.0.0.3] | 192.0.0.3 | 100 | | 0 | i |
| | RT:100:101 ET:8 | | | | |





Layer 3 VPN over EVPN with VXLAN

- As an alternative Layer3 VPN to MPLS VPN, it is possible to create a Layer3 VPN over an EVPn with VXLAN.



- Using announcements of Route Type-5.
- Route Type-5
 - Announces IP network prefixes.

| |
|--------------------------------|
| RD (8 Octets) |
| ESI (10 Octets) |
| Ethernet Tag ID (4 Octets) |
| IP Prefix Length (1 Octet) |
| IP Prefix (4 or 16 Octets) |
| GW IP Address (4 or 16 Octets) |
| MPLS Label (3 Octets) |

```
- Border Gateway Protocol - UPDATE Message
  Marker: ffffffffffffffffffffffffffffffffff
  Length: 121
  Type: UPDATE Message (2)
  Withdrawn Routes Length: 0
  Total Path Attribute Length: 98
- Path attributes
  - Path Attribute - MP_REACH_NLRI
    - Flags: 0x90, Optional, Extended-Length, Non-transitive, Complete
    Type Code: MP_REACH_NLRI (14)
    Length: 45
    Address family identifier (AFI): Layer-2 VPN (25)
    Subsequent address family identifier (SAFI): EVPN (70)
    - Next hop: 192.0.0.1
    Number of Subnetwork points of attachment (SNPA): 0
  - Network Layer Reachability Information (NLRI)
    - EVPN NLRI: IP Prefix route
      Route Type: IP Prefix route (5)
      Length: 34
      Route Distinguisher: 00010a0101010004 (10.1.1.1:4)
      - ESI: 00:00:00:00:00:00:00:00:00:00
      Ethernet Tag ID: 0
      IP prefix length: 16
      IPv4 address: 10.1.0.0
      IPv4 Gateway address: 0.0.0.0
      VNI: 101
    - Path Attribute - ORIGIN: INCOMPLETE
    - Path Attribute - AS_PATH: empty
    - Path Attribute - MULTI_EXIT_DISC: 0
    - Path Attribute - LOCAL_PREF: 100
    - Path Attribute - EXTENDED_COMMUNITIES
      - Flags: 0xc0, Optional, Transitive, Complete
      Type Code: EXTENDED_COMMUNITIES (16)
      Length: 24
      - Carried extended communities: (3 communities)
        - Encapsulation: VXLAN Encapsulation [Transitive Opaque]
        - Route Target: 100:101 [Transitive 2-Octet AS-Specific]
        - EVPN Router's MAC: Router's MAC: 0c:32:45:6d:00:01 [Transitive EVPN]
```



References

- Building Data Centers with VXLAN BGP EVPN: A Cisco NX-OS Perspective (Networking Technology), 1st Edition, David Jansen, Lukas Krattiger, Shyam Kapadia, Cisco Press (March 31, 2017), ISBN-13: 978-1587144677.
- The Fast-Track Guide to VXLAN BGP EVPN Fabrics: Implement Today's Multi-Tenant Software-Defined Networks, 1st Edition, Rene Cardona, Apress (May 19, 2021), ISBN-13:978-1484269299.

