Network Fundamentals for Cloud

Nishit Narang
WILPD-CSIS

BITS Pilani
Pilani Campus

BITS Pilani
Pilani Campus

CC ZG503: Network Fundamentals for Cloud

Lecture No. 12: Data Center Networks (Contd.)

RECAP: DCN Networking Technology Evolution - NVO3 & Overlay Networks

- NVO3 technologies are overlay network technologies driven by IT vendors and aim to get rid of the dependency on the traditional physical network architecture

Overlay Network:

- A software-defined logical network built over an existing underlay network
- Is completely decoupled from the underlay network
- In modern DCN, created using NVO3 technology
  - Overlay / NVO3 technology is a tunnel encapsulation technology that encapsulates Layer 2 packets over tunnels and transparently transmits the encapsulated packets
  - In a DCN environment, it enables layer 2 communication between large-scale VMs on the DCN

NVO3 technologies include VXLAN and NVGRE. VXLAN is used by the majority of enterprises

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

RECAP: Use of Tunneling for VPNs

- A tunnel is often used to encapsulate a packet of one protocol into a packet of another protocol to carry it over an intermediate network that supports the latter protocol
- Virtual Private Networks (VPNs) use tunneling for:
  - Hiding the inner IP packet and associated IP addresses
  - Allowing Private IP addresses to be used within the private network – the public IP address is added at network edge in the outer IP header

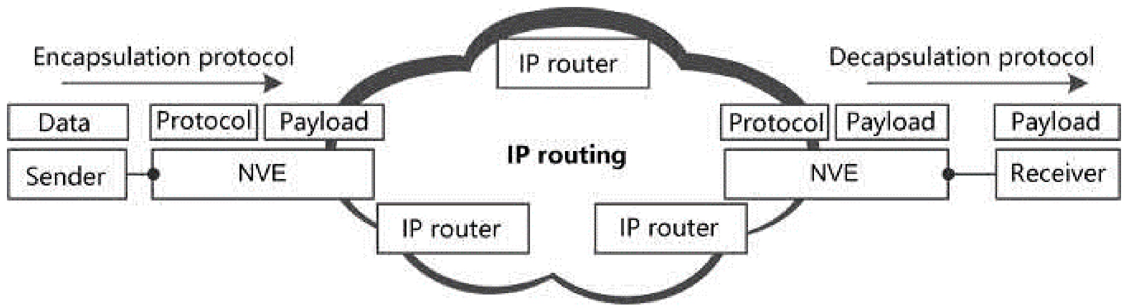
DCN Networking Technology Evolution (contd.)

- Network Virtualization Overlays (NVO3) Technologies
  - Since an overlay network is a virtual network topology constructed on top of a physical network, thus…
    - Each virtual network instance that is implemented as an overlay, requires that an original frame is encapsulated on a Network Virtualization Edge (NVE).
    - The encapsulation identifies the device that will perform decapsulation.
    - Before sending the frame to the destination endpoint, the device decapsulates the frame to obtain the original frame.
    - Intermediate network devices forward the encapsulated frame based on the outer encapsulation header and are oblivious to the original frame carried in the encapsulated frame.
    - The NVE can be a traditional switch or router, or a virtual switch in a hypervisor.
    - The endpoint can be a VM or physical server.
  - A virtual network identifier (e.g. VxLAN network identifier or VNI) can be encapsulated into an overlay header to identify a virtual network to which a data frame belongs.
  - Because a virtual DC supports both routing and bridging, the original frame in the overlay header can be a complete Ethernet frame containing a MAC address or an IP packet.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

DCN Networking Technology Evolution (contd.)

- Network Virtualization Overlays (NVO3) Technologies



- The sender in the figure is an endpoint, which may be a VM or physical server
- An NVE may be a physical switch or a virtual switch on a hypervisor
- The sender can be connected to an NVE directly or through a switching network
- NVEs are connected through a tunnel

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

DCN Networking Technology Evolution (contd.)

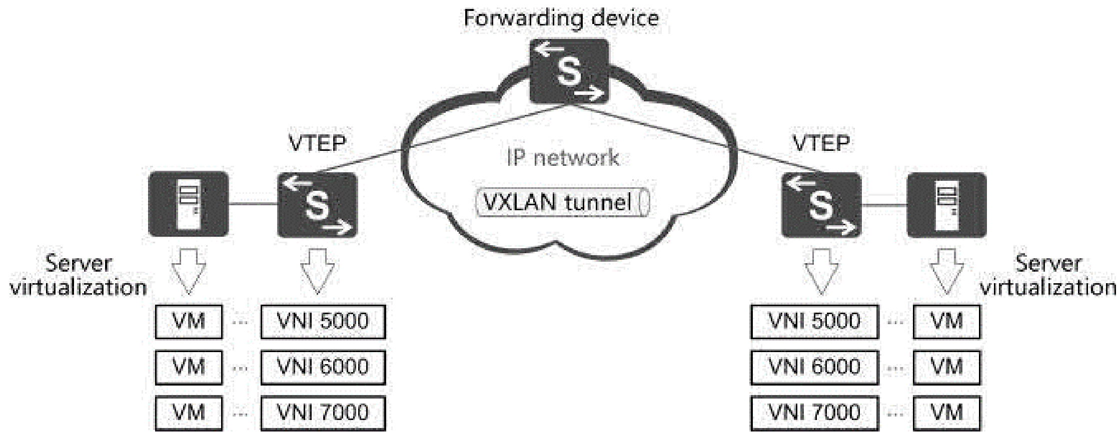
- L2MP Vs NVO3 Technologies
  - To some extent, NVO3 and L2MP technologies are similar.
  - They both build an overlay network on the physical network
  - The difference is that L2MP technologies add a new forwarding identifier to the original Layer 2 network, thereby requiring that chips on hardware devices support L2MP technologies.
  - In contrast, NVO3 technologies reuse the current IP forwarding mechanism and only add a new logical network that does not depend on the physical network environment on the traditional IP network.
  - The logical network is not perceived by physical devices, and its forwarding mechanism is the same as the IP forwarding mechanism.
  - In this way, the threshold of NVO3 technologies is greatly lowered, and this is why NVO3 technologies have become popular on DCNs in a few years

Typical NVO3 technologies include VXLAN, Network Virtualization Using Generic Routing Encapsulation (NVGRE), and Stateless Transport Tunneling (STT), among which VXLAN is the most popular one.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

VXLAN Basics

- VXLAN is an NVO3 technology that enables Layer 2 forwarding over a Layer 3 network by using L2 over L4 (MAC-in-UDP) encapsulation
- Defined by the IETF, it allows VMs to migrate over a large Layer 2 network and isolates tenants in a DC



Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

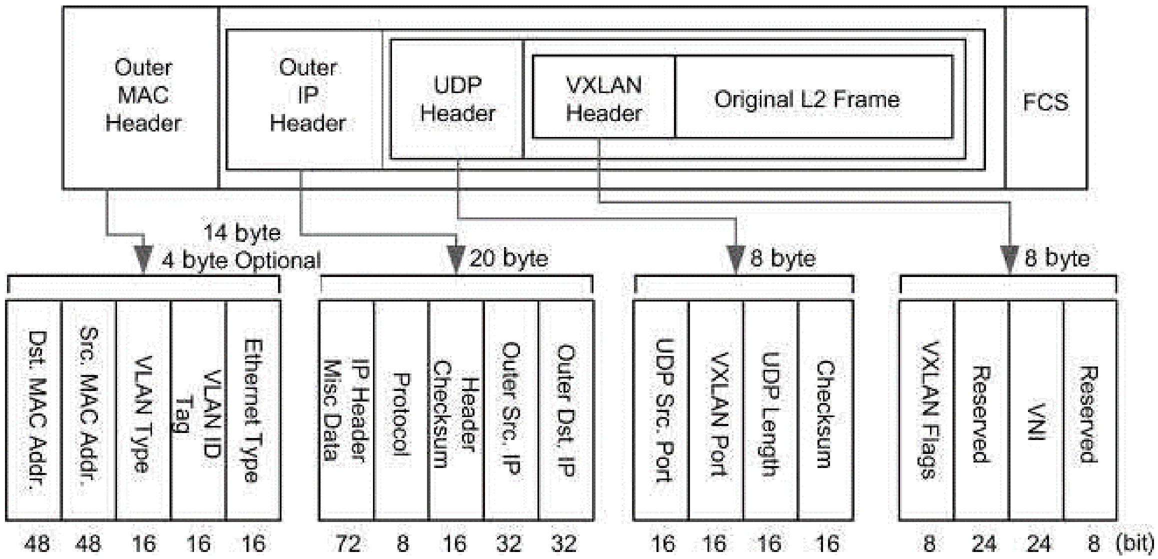
VXLAN Benefits

- **VLAN flexibility in multitenant segments:** It provides a solution to extend Layer 2 segments over the underlying network infrastructure so that tenant workload can be placed across physical pods in the data center.
- **Higher scalability:** VXLAN uses a 24-bit segment ID known as the VXLAN network identifier (VNID), which enables up to 16 million VXLAN segments to coexist in the same administrative domain.
- **Improved network utilization:** VXLAN solves the Layer 2 STP limitations. VXLAN packets are transferred through the underlying network based on its Layer 3 header and can take complete advantage of Layer 3 routing, equal-cost multipath (ECMP) routing, and link aggregation protocols to use all available paths.

Source: [Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press]

VXLAN Packet Format

- The transport protocol over the physical data center network is UDP/IP
- With MAC-in-UDP encapsulation, VXLAN tunnels the Layer 2 network over the Layer 3 network.

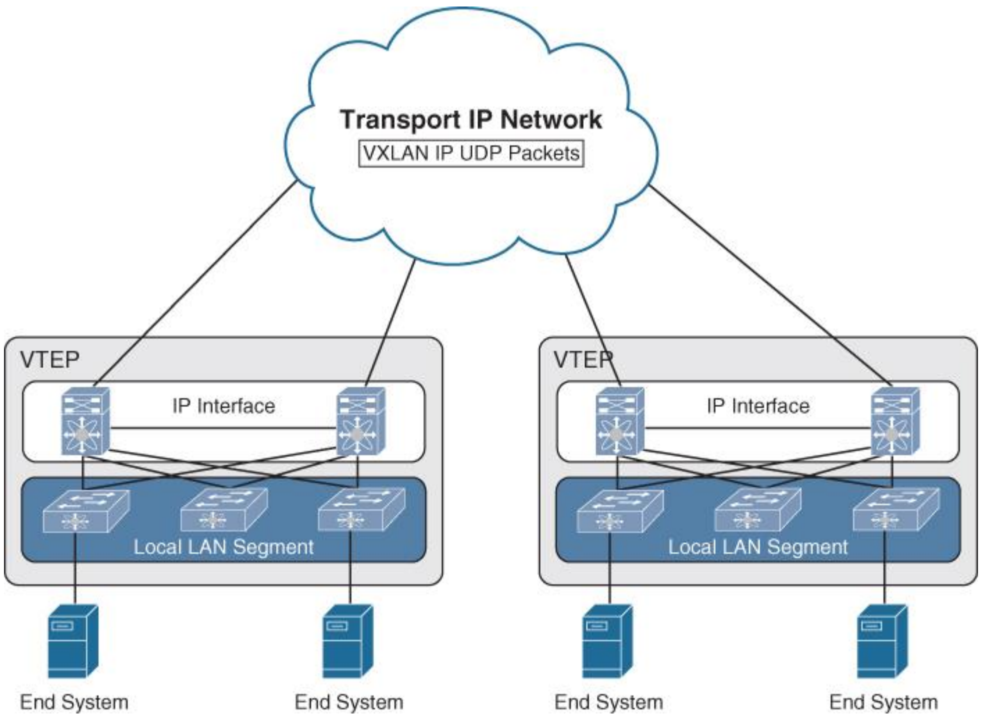


- In the UDP header,
  - the destination port number has a fixed value of 4789
  - The source port number is the hash value of the original Ethernet frame
- Outer IP (/MAC) header
  - source IP (/MAC) address specifies the IP (/MAC) address of the VTEP, where the source VM belongs.
  - destination IP address indicates the IP address of the VTEP where the destination VM belongs
  - destination MAC address is the MAC address of the next-hop device on the path to the destination VTEP.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

VXLAN Tunnels and VTEP

- VXLAN tunnel endpoint (VTEP) maps tenants' end devices to VXLAN segments performs VXLAN encapsulation and de-encapsulation
- Each VTEP function has two interfaces:
  - one is a switch interface on the local LAN segment to support local endpoint communication, and
  - the other is an IP interface to the transport IP network
- A VTEP device discovers the remote VTEPs for its VXLAN segments and learns remote MAC Address-to-VTEP mappings through its IP interface



Source: [Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press]

VNI

- A virtual network identifier (VNI) is a value that identifies a specific virtual network in the data plane
- It is typically a 24-bit value part of the VXLAN header, which can support up to 16 million individual network segments.
  - Valid VNI values are from 4096 to 16,777,215.
- There are two main VNI scopes:
- **Network-wide scoped VNIs:** The same value is used to identify the specific Layer 3 virtual network across all network edge devices
  - A uniform VNI per VPN is a simple approach → eases network operations
- **Locally assigned VNIs:** In an alternative approach supported as per RFC 4364, the identifier has local significance to the network edge device that advertises the route
  - uses the same existing semantics as an MPLS VPN label

Source: [Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press]

VXLAN Overlay Network Types

- Classified as one of three types:
  - Network overlay: All VTEPs are deployed on physical switches.
  - Host overlay: All VTEPs are deployed on vSwitches.
  - Hybrid overlay: Some VTEPs are deployed on physical switches with others deployed on vSwitches

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

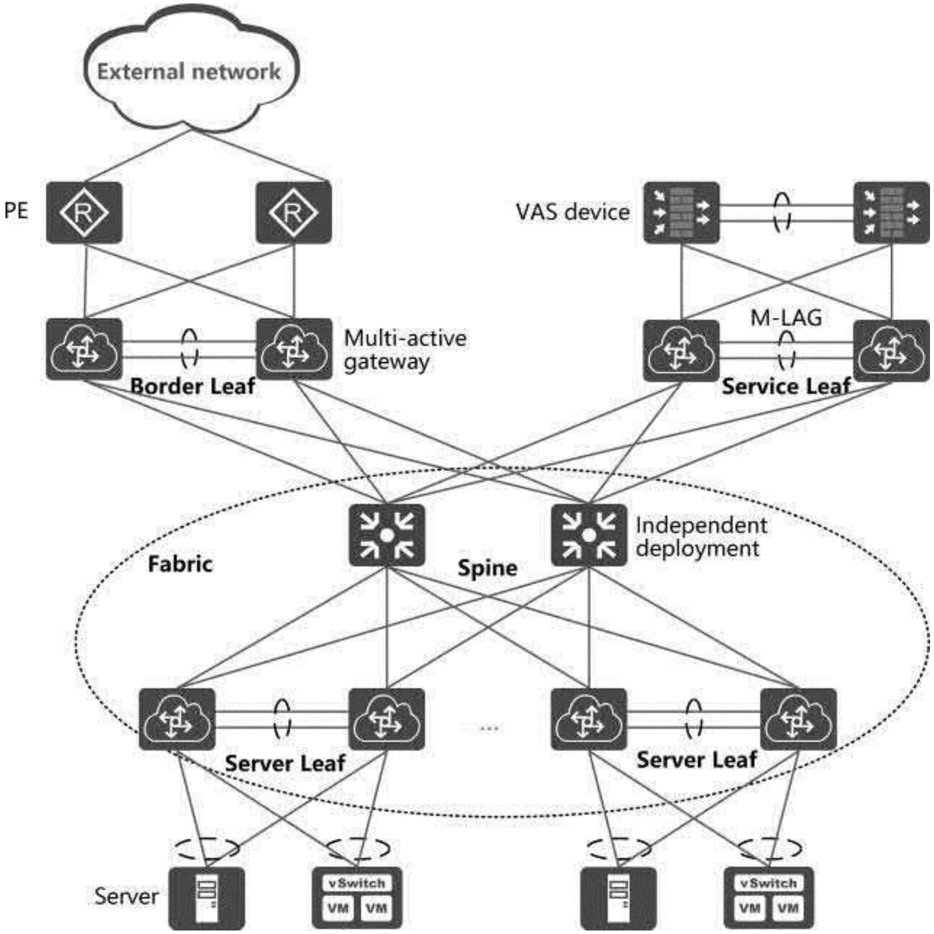
VXLAN Control Plane

- Two widely adopted control planes are used with VXLAN:
  - the VXLAN Flood and Learn Multicast-Based Control Plane and
  - the VXLAN MPBGP EVPN Control Plane.

Source: [Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press]

Control Plane Protocols in Data Center Networks

Constructing DCN Underlay Network



(a) Role	Function
(b) Fabric	Network failure domain that is managed by an SDN controller. It contains one or more spine-leaf architectures
Spine	Core node on a VXLAN fabric network. It provides high-speed IP forwarding and connects to leaf nodes through high-speed interfaces
Leaf	Access node on a VXLAN fabric network. It connects various network devices to the VXLAN fabric network
Service leaf	Functional node that connects VAS devices, such as firewalls and LBs, to a VXLAN fabric network
Server leaf	Functional node that connects virtual and physical servers to a VXLAN fabric network
Border leaf	Functional node that connects to routers or transmission devices outside a DC to forward traffic from an external network to a VXLAN fabric network in a DC

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

Constructing DCN Underlay Network (contd..)

- Leaf nodes and spine nodes are connected through Layer 3 routed interfaces
- They communicate at Layer 3 by configuring a dynamic routing protocol. OSPF or BGP is recommended.
- ECMP is recommended for implementing load balancing and link backup
  - In this case, leaf nodes forward data traffic to spine nodes through multiple ECMP paths, guaranteeing reliability while ensuring that network bandwidth improves.

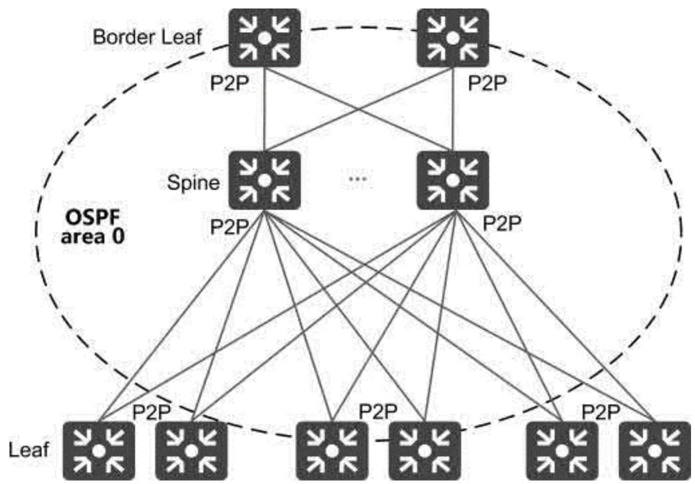
Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

Routing Protocol Selection

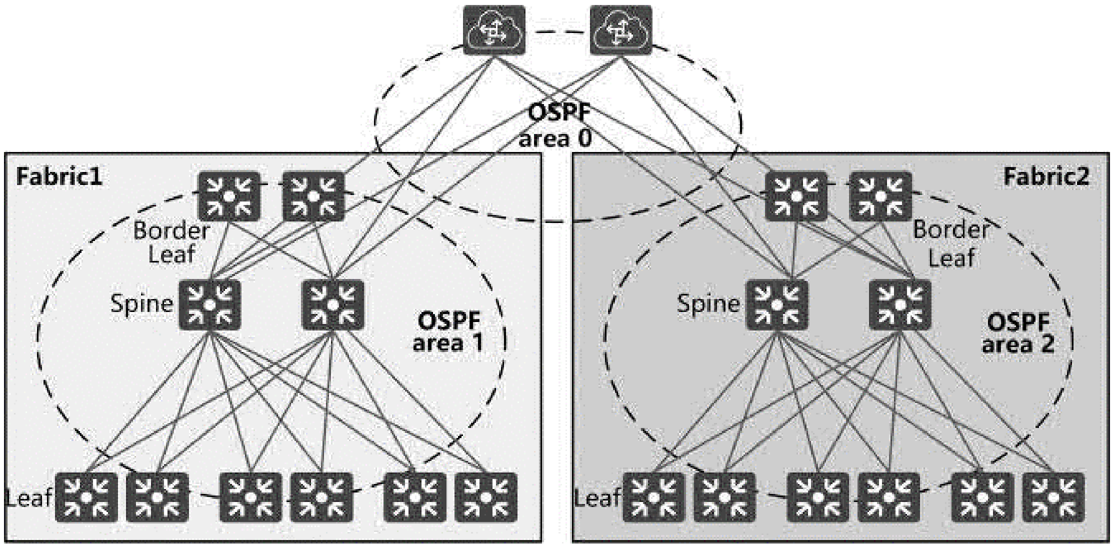
- In most cases, either OSPF or EBGP can be used on an underlay network
- OSPF is preferred for most small to medium sized networks
  - When the number of leaf nodes is less than 100, OSPF is recommended on the underlay network
- If the scale of the network is large, EBGP is recommended as the underlay network needs to be partitioned into areas, and flexible control of BGP is required
  - As the size of the network increases, and/or the number of prefixes to be advertised increases, BGP becomes the go-to routing protocol
  - For advertising fewer than 32,000 prefixes, OSPF is a fine choice as a routing protocol. For larger prefixes, BGP should be preferred.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021 **AND** Dinesh G. Dutt. Cloud Native Data Center Networking: Architecture, Protocols and Tools, O'Reilly 2020

OSPF Deployment Scenarios



Recommended OSPF planning for a single fabric network



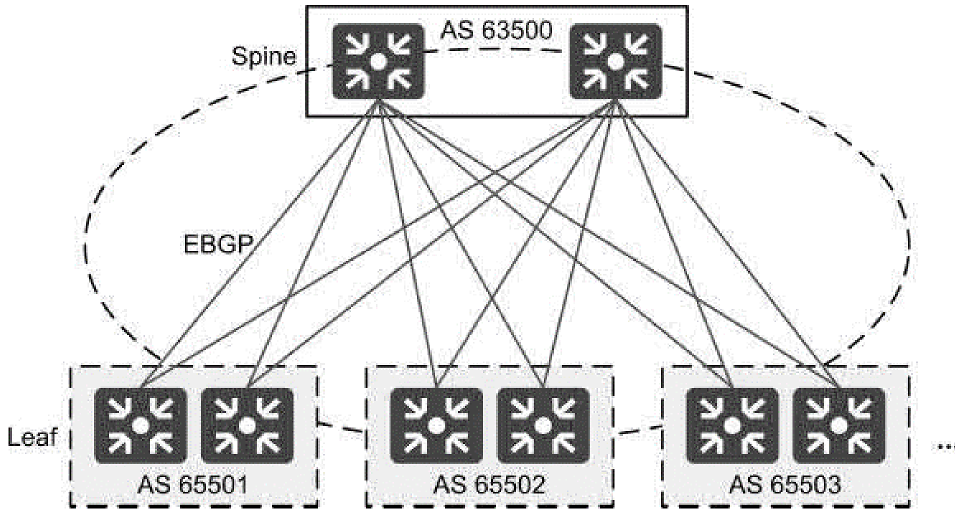
Recommended OSPF planning for multi-fabric deployment (single VXLAN domain).

If multiple fabric networks form two VXLAN domains on the overlay network (that is, two DCNs are managed through two sets of management interfaces, but they need to be interconnected), then:

- OSPF be deployed on each fabric network, and
- interconnected devices between fabric networks can exchange routes through BGP

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

EBGP Deployment in DCN Underlay Networks



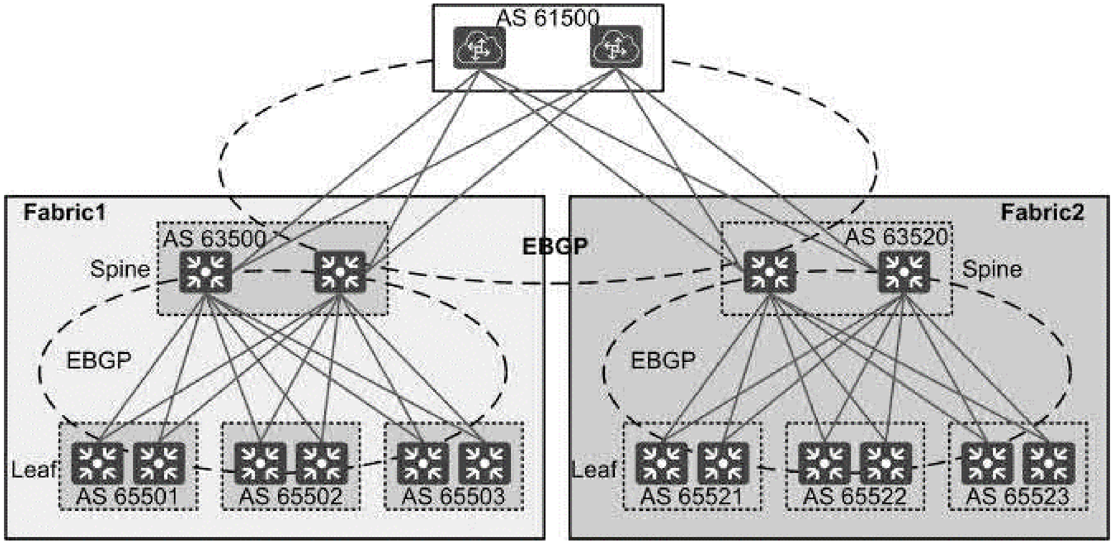
Recommended EBGP planning for a single fabric network.

On a single fabric network:

- spine nodes are added to the same autonomous system (AS),
- each group of leaf nodes is added to an AS
- EBGP peer relationships are established between leaf nodes and all spine nodes.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

EBGP Deployment in DCN Underlay Networks



Recommended EBGP planning for multiple fabric networks

Similar concept as Single Fabric Network

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

Overlay Networks: VXLAN Control Plane

- Two widely adopted control planes are used with VXLAN:
  - the VXLAN Flood and Learn Multicast-Based Control Plane and
  - the VXLAN MPBGP EVPN Control Plane.

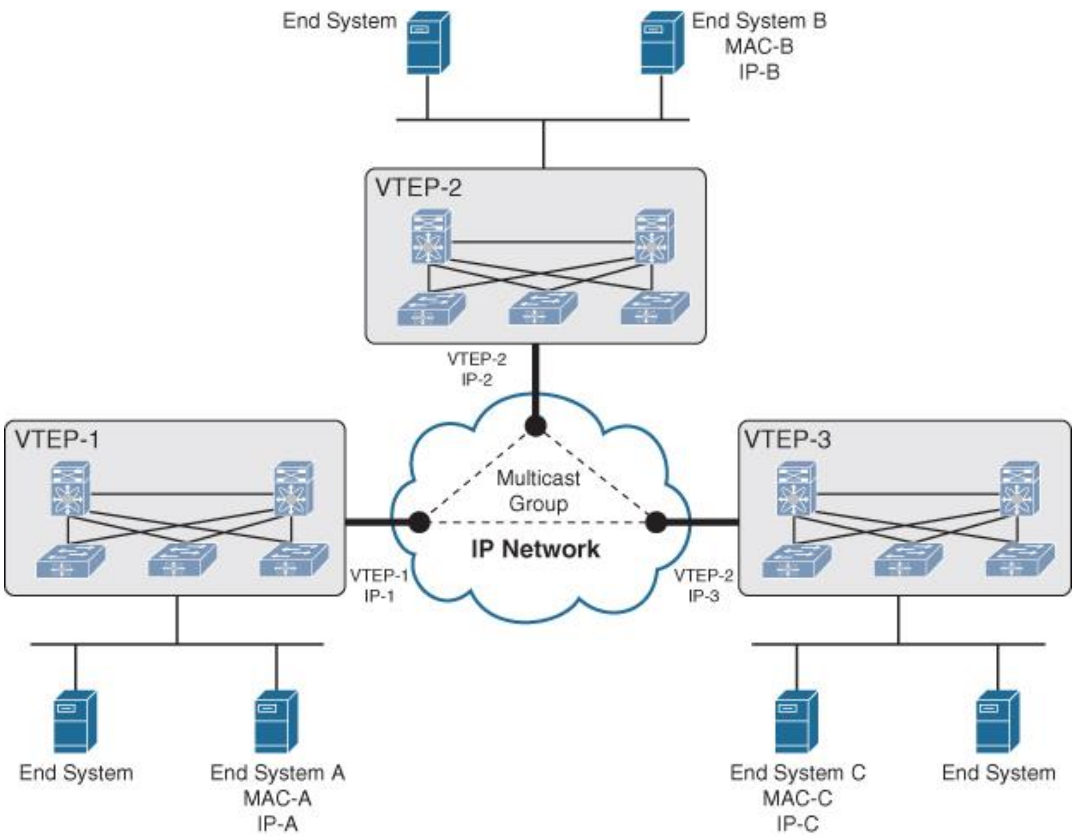
Source: [Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press]

VXLAN Flood and Learn Multicast-Based Control Plane

- Switches utilize existing Layer 2 flooding mechanisms and dynamic MAC address learning to
  - Transport broadcast, unknown unicast, and multicast (BUM) traffic
  - Discover remote VTEPs
  - Learn remote-host MAC addresses and MAC-to-VTEP mappings for each VXLAN segment
- IP multicast is used to reduce the flooding scope of the set of hosts that are participating in the VXLAN segment
  - Each VXLAN segment, or VNID, is mapped to an IP multicast group in the transport IP network
  - Each VTEP device is independently configured and joins this multicast group as an IP host through the Internet Group Management Protocol (IGMP).

Source: Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press

VXLAN Flood and Learn Multicast-Based Control Plane (Contd.)



Source: [Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press]

VXLAN Flood and Learn Multicast-Based Control Plane (Contd.)

As an example, if End System A wants to talk to End System B, it does the following:

1. End System A generates an ARP request trying to discover the End System B MAC address.

2. When the ARP request arrives at SW1, it will look up its local table, and if an entry is not found, it will encapsulate the ARP request over VXLAN and send it over the multicast group configured for the specific VNI.

3. The multicast RP receives the packet, and it forwards a copy to every VTEP that has joined the multicast group.

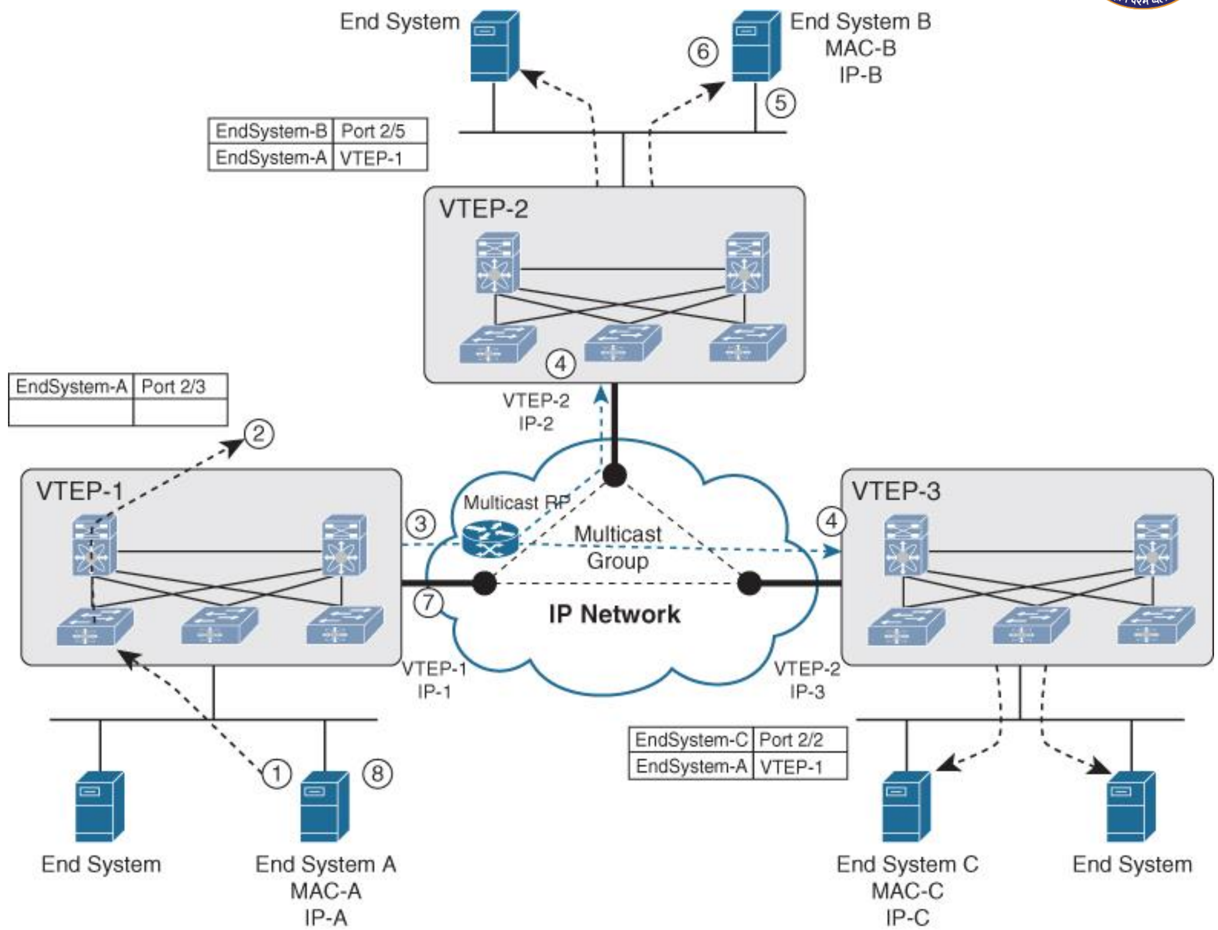
4. Each VTEP receives and de-encapsulates the VXLAN packet and learns the System A MAC address pointing to the remote VTEP address.

5. Each VTEP forwards the ARP request to its local destinations.

6. End System B generates the ARP reply. When SW2 VTEP2 receives it, it looks up its local table and finds an entry with the information that traffic destined to End System A must be sent to VTEP1 address. VTEP2 encapsulates the ARP reply with a VXLAN header and unicasts it to VTEP1.

7. VTEP1 receives and de-encapsulates the packet and delivers it to End System A.

8. When the MAC address information is learned, additional packets are fed to the corresponding VTEP address.



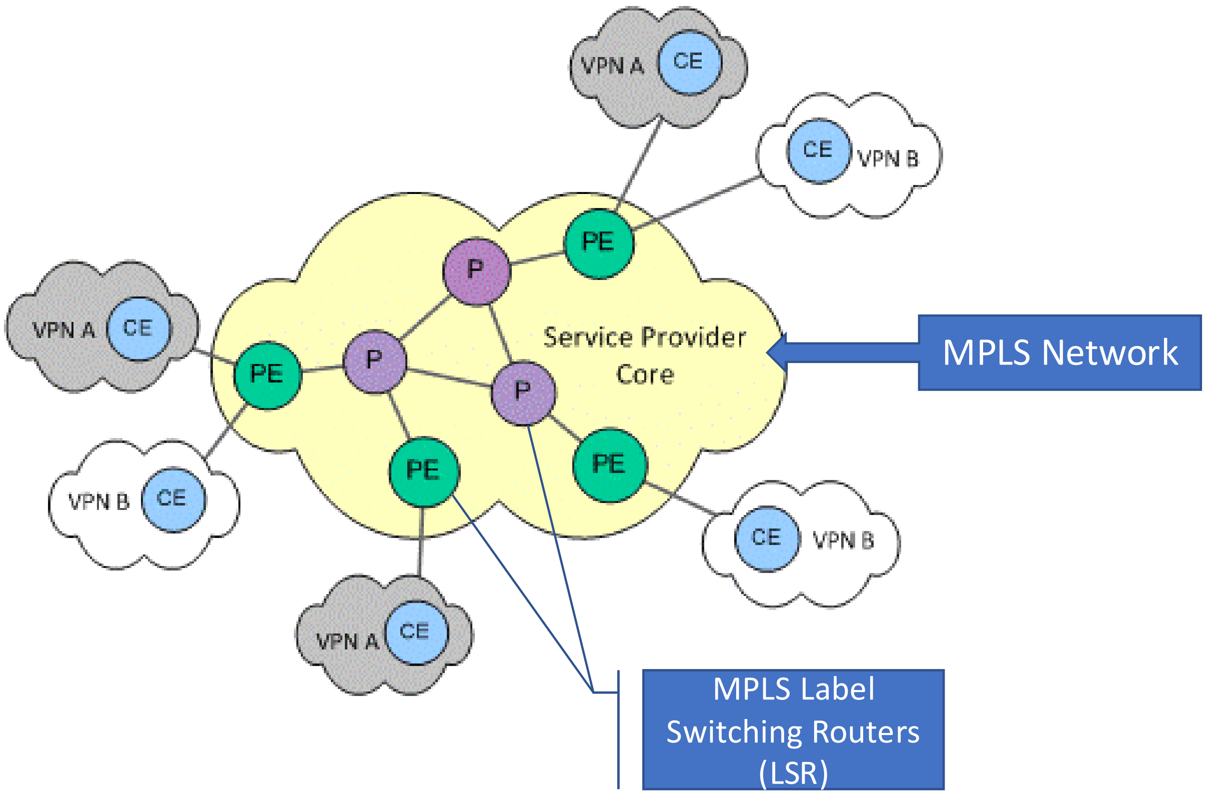
Source: Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press

VXLAN MPBGP EVPN Control Plane

- The Flood-and-Learn method creates significant flooding traffic resulting in network expansion difficulties
- As a solution to these problems, EVPN is introduced on the VXLAN control plane
  - EVPN = Ethernet Virtual Private Network OR Ethernet VPN
  - By referring to the BGP/MPLS IP VPN mechanism, EVPN defines several types of BGP EVPN routes by extending BGP.
  - The PE node role described in BGP MPLS EVPN is equivalent to the VTEP/network virtualization edge (NVE) device
  - It advertises BGP routes on the network to implement automatic VTEP discovery and host address learning.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021 and Virtual Extensible LAN (VXLAN) Overview > Implementing Data Center Overlay Protocols | Cisco Press

VXLAN MPBGP EVPN Control Plane: Reference to Layer 3 MPLS VPNs



The various routers within the VPN communicate using BGP, an IP routing protocol that defines how routes can be distributed. BGP transports information about CE routers only to members of the same VPN, ensuring security.

Source: https://docs.oracle.com/cd/E35292_01/doc.72/e47715/con_vpn.htm#autoId0

EVPN Advantages

- Using EVPN on the control plane offers the following advantages:
  - VTEPs can be automatically discovered and VXLAN tunnels can be automatically established, overall simplifying network deployment and expansion.
  - EVPN can advertise Layer 2 MAC addresses and Layer 3 routing information simultaneously.
  - Flooding traffic is reduced on the network.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

MP-BGP

- Traditional BGP-4 uses Update packets to exchange routing information between peers
  - An Update packet can advertise a type of reachable routes with the same path attributes, placed in Network Layer Reachability Information (NLRI) fields.
- BGP-4 can manage only IPv4 unicast routing information
- As a solution, Multiprotocol Extensions for BGP (MP-BGP) was developed as a means to support multiple network layer protocols, including IPv6 and multicast.
  - MP-BGP extends NLRI fields based on BGP-4.
  - After extension, the description of the address family is added to the NLRI fields to differentiate network layer protocols.
  - These include the IPv6 unicast address family and VPN instance address family.
- The EVPN NLRI defines different types of BGP EVPN routes

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

BGP EVPN Route Types

- The BGP EVPN route types are as follows:
  - Type 2 route — MAC/IP route: is used to advertise the MAC addresses, ARP entries, and routing information of hosts
  - Type 3 route — inclusive multicast route: is used for automatic discovery of VTEPs and dynamic establishment of VXLAN tunnels

Route Distinguisher	RD of an EVPN instance.
Ethernet Segment Identifier	Unique ID for defining the connection between local and remote devices.
Ethernet Tag ID	VLAN ID.
MAC Address Length	Length of the host MAC address carried in the route.
MAC Address	Host MAC address carried in the route.
IP Address Length	Mask length of the host IP address carried in the route.
IP Address	Host IP address carried in the route.
MPLS Label1	Layer 2 VNI carried in the route.
MPLS Label2	Layer 3 VNI carried in the route.

Prefix

Route Distinguisher	RD of an EVPN instance.
Ethernet Tag ID	VLAN ID. Here, the value is 0.
IP Address Length	Mask length of the local VTEP's IP address carried in the route.
Originating Router's IP Address	Local VTEP's IP address carried in the route.

PMSI attribute

Flags	Flag bit. This field is inapplicable in VXLAN scenarios.
Tunnel Type	Tunnel type carried in the route. The value can only be 6.
MPLS Label	Layer 2 VNI carried in the route.
Tunnel Identifier	Tunnel identifier carried in the route.

NLRI format of Type 2 and Type 3 Routes

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

BGP EVPN Route Types (Contd.)

- The BGP EVPN route types are as follows:
  - Type 5 route — IP prefix route: is used to advertise imported external routes or advertise routing information of hosts

Route Distinguisher	RD of an EVPN instance.
Ethernet Segment Identifier	Unique ID for defining the connection between local and remote devices.
Ethernet Tag ID	VLAN ID.
IP Prefix Length	Length of the IP prefix carried in the route.
IP Prefix	IP prefix carried in the route.
GW IP Address	Default gateway address.
MPLS Label	Layer 3 VNI carried in the route.

6.9 NLRI format in a type 5 route.

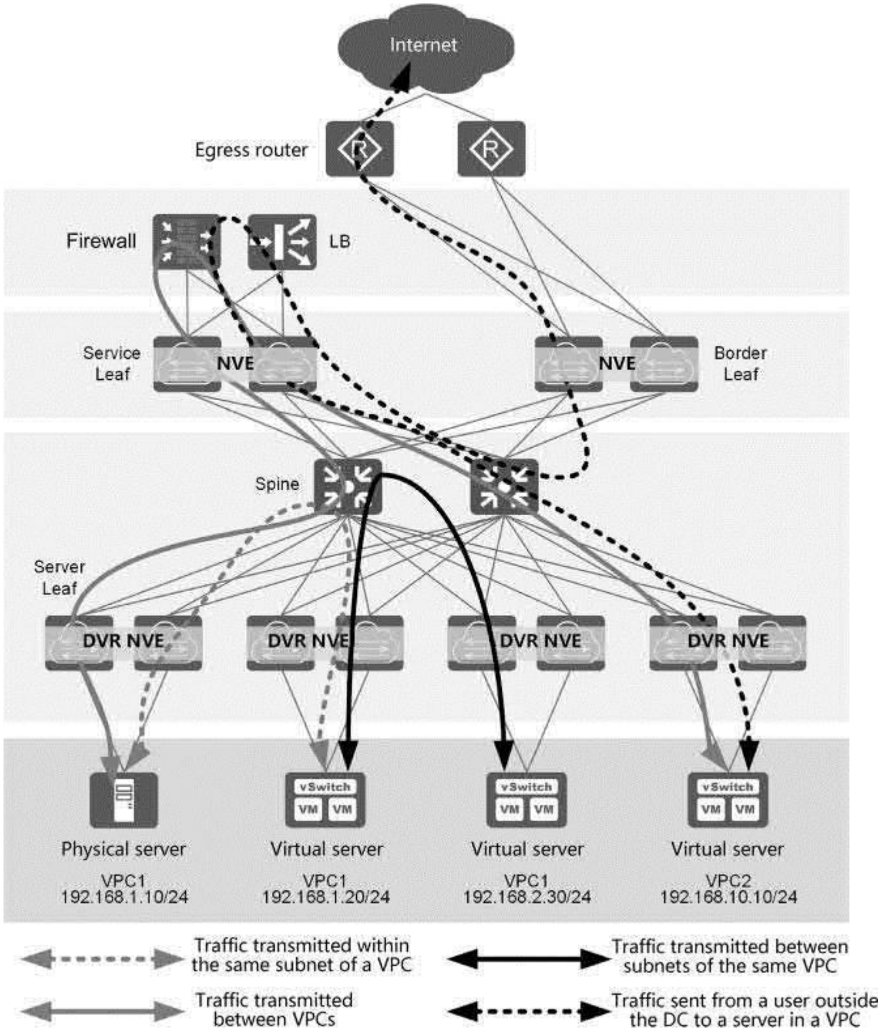
Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

For completeness – the VXLAN Data Plane!

- Depending on the traffic flow direction and scope, DCN traffic can be classified into east-west traffic (transmitted within a DC) and north-south traffic (sent across the DC)
  - Traffic transmitted within the same subnet of a VPC is forwarded by a TOR switch after Layer 2 VXLAN encapsulation.
  - Traffic transmitted between subnets of the same VPC is forwarded by a TOR switch based on Layer 3 routes. This is done after Layer 3 VXLAN encapsulation.
  - Traffic transmitted between VPCs is forwarded across subnets, and isolation for security purposes is required. Therefore, to meet this, the Traffic needs to pass through a firewall and reach the Layer 3 VXLAN gateway.
  - Traffic sent from a user outside the DC to a server in a VPC passes through the Intrusion Prevention System (IPS) or firewall, LB, VXLAN gateway, and TOR switch before reaching the server.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

VXLAN Data Plane



Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

Thank You!