BITS Pilani

Pilani Campus

Network Fundamentals for Cloud

Nishit Narang
WILPD-CSIS

innovate achieve lead

BITS Pilani
Pilani Campus

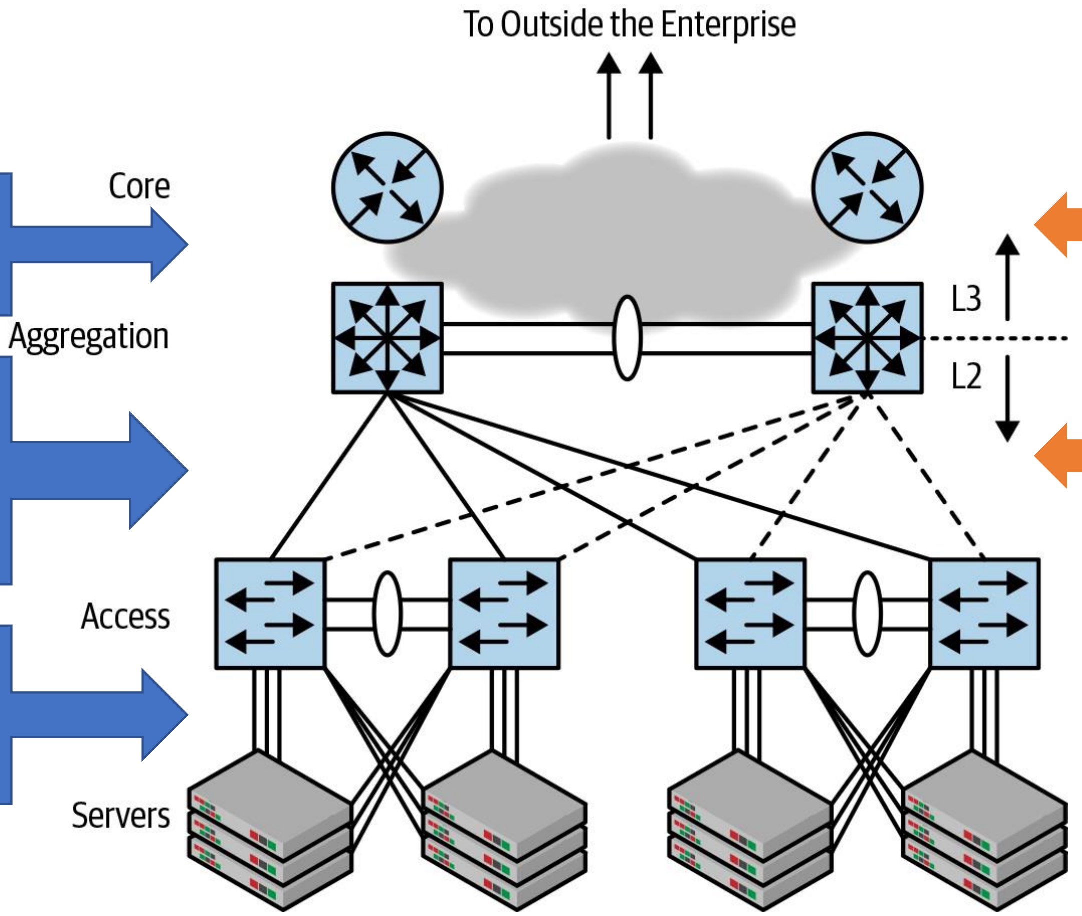
CC ZG503: Network Fundamentals for Cloud

Lecture No. 10: Data Center Networks (Contd.)

RECAP: DCN Evolution

- Traditional network topology
  - *access-aggregation-core*
  - Became prominent around year 2000
  - Considered fast, cheap and easy to administer
  - well suited to the north-south traffic pattern of client-server application architecture
  - Not suited, however, to the server-server traffic pattern of DCNs
- Modern DCN topologies:
  - The structure of the new world is the Clos topology (*named after one of its inventors, Charles Clos*)
  - Basic Clos topology is also called the ***leaf-spine*** topology
  - ***Fat Tree*** topology, a special instance of the Clos topology is extremely popular

RECAP:Traditional Network Topology



Aggregation to Core connects for external traffic

Access Switches connected to Aggregation Switches at the next hierarchy level

Servers connected to Access Switches at the lowest level

L3 Routing

L2 Bridging

- Silicon-switched packet forwarding
- Proprietary enterprise network stacks
- Promise of zero-configuration bridging

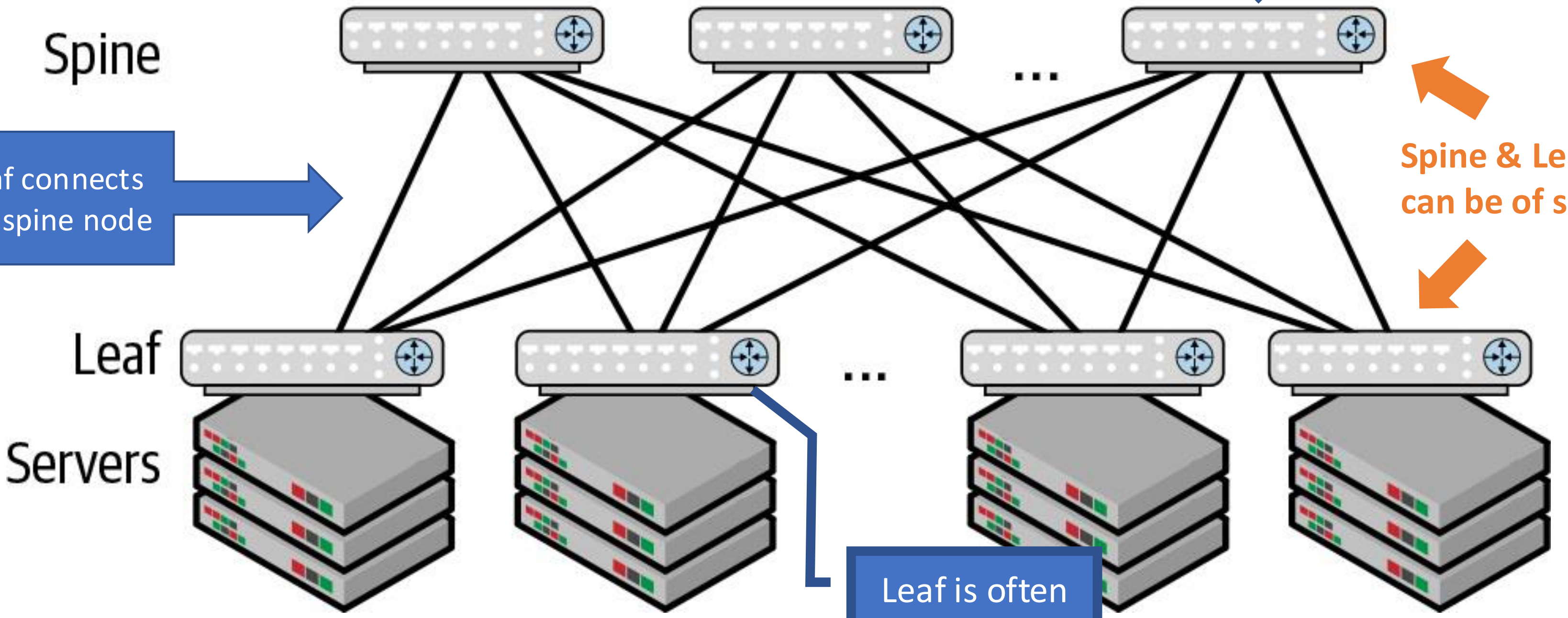
Source: Cloud Native Data Center Networking by Dinesh G. Dutt

RECAP: Challenges with Acc-Agg-Core Topologies

- Lack of scalability for DCN traffic patterns / applications
  - Flooding → ***flood-and-learn*** model of self-learning bridges doesn't scale!
  - VLAN limitations → 12-bit VLAN ID => 4096 VLANs, a paltry value at the scale of the cloud
  - Burden on Aggregation switches (2) to respond to all ARP messages
  - STP limitations → more east-west traffic => more aggregation switches. Unpredictable / unusable topologies emerged due to link/node failures.
- Complexity
  - Unless the access-agg-core network is carefully designed, congestion can quite easily occur in such networks → over-subscription of network bandwidth
- Failure Impact
  - access-agg-core model is prone to very coarse-grained failures; In other words, failures with large blast radiuses.
    - For example, the failure of a single link halves the available bandwidth
- Inflexibility: It is not possible to have the same VLAN be present across two different pairs of aggregate switches

RECAP: Clos Network Topology

- Basic Clos Topology → ***leaf-spine*** topology



Spines connect the leaves to one another

Each leaf connects to every spine node

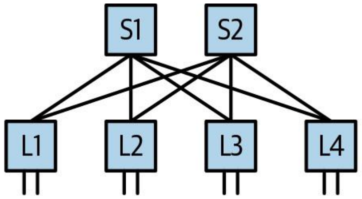
Spine & Leaf switches can be of same type

Leaf is often called the ToR switch

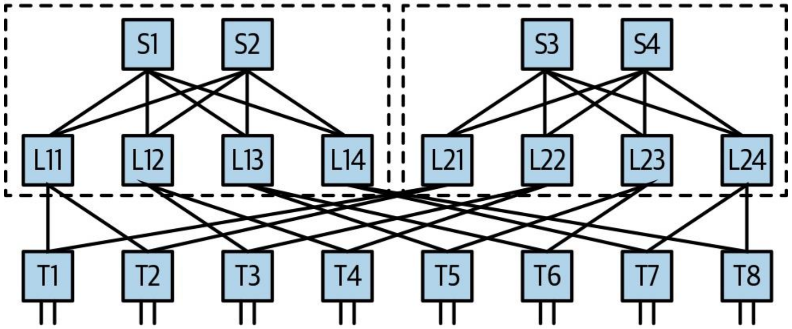
Source: Cloud Native Data Center Networking by Dinesh G. Dutt

RECAP: Scaling Clos Topology

Examples with four-port switches

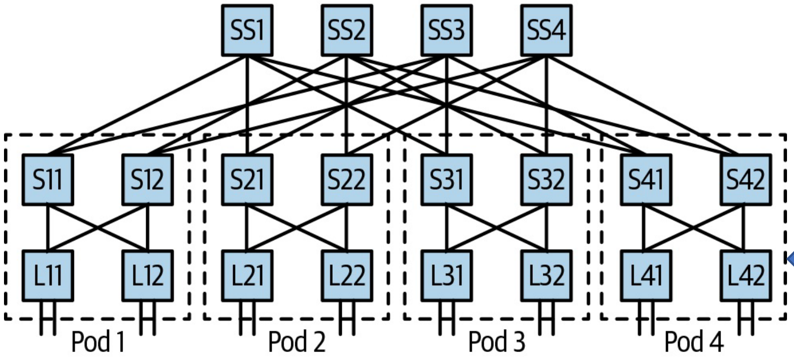


(a) Two-tier Clos using four-port switches



(b) Virtual chassis-based three-tier Clos using four-port switches

Model popularized by **Facebook**



(c) Pod-based three-tier Clos using four-port switches

Model used by Microsoft and Amazon

Source: Cloud Native Data Center Networking by Dinesh G. Dutt

innovate achieve lead

BITS Pilani
Pilani Campus

DCN Evolution & Technologies

DCN Networking Technology Evolution

- xSTP technologies to eliminate loops in the L2 Network broadcast domain
  - STP had many issues (as summarized earlier).
  - Led to the evolution of Virtual Chassis Technologies
- Virtual Chassis Technology
  - Virtual chassis technologies implement N:1 virtualization
  - Integrates the control planes of multiple devices to form a unified logical device
    - I.e. Different physical devices share the same control plane, which is equivalent to creating a cluster for physical network devices
  - This logical device has a unified management IP address and also works as one node in various Layer 2 and Layer 3 protocols
    - Link aggregation allows this logical device to connect to each physical or logical node at the edge through only one logical link, solving the problem of dual-homing of terminals that inevitably causes loops on the network.
    - Therefore, the network topology after the integration is loop-free for xSTP, which indirectly avoids problems of xSTP
    - Master election and a master/standby switchover can also be performed in the cluster

Several virtual chassis technologies exist in the industry, such as Cisco’s VSS, Huawei’s CSS, etc

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

DCN Networking Technology Evolution (contd.)

- Disadvantages of Virtual Chassis Technology
  - Limited scalability: the scalability of any virtual chassis technology is limited by the performance of the master switch, as it provides the control plane for the entire virtual chassis system
  - Reliability: Because the control plane is on the master switch, packet loss may last for a long time or the entire system may stop running if the master switch fails
  - Upgrade Challenges: Control plane integration also makes it difficult to upgrade a virtual chassis system. Common restart and upgrade operations cause interruption of the control plane, resulting in packet loss for a long time.
  - Bandwidth waste: Dedicated links in a virtual chassis system are used for status exchange and data forwarding between devices

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

DCN Networking Technology Evolution (contd.)

- Layer 2 Multi-Pathing (L2MP) Technologies
  - L2MP technologies attempt to address the below twin challenges of xSTP and Virtual Chassis technologies:
    - They cannot support large DCNs with massive amounts of data.
    - Their link utilization is low.
  - It is recommended that link- state routing protocols widely used on Layer 3 networks be employed
  - These protocols not only support a large number of devices, but they are also loop-free and have high link utilization
    - Example: OSPF and IS-IS → they support ECMP load balancing and use the Shortest Path First (SPF) algorithm
  - <u>The basic principle of L2MP technologies is to introduce mechanisms of routing technologies used on Layer 3 networks to Layer 2 networks</u>

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

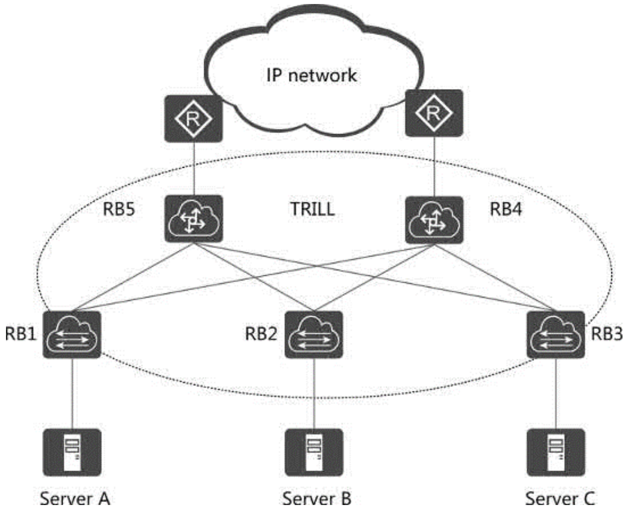
DCN Networking Technology Evolution (contd.)

- Layer 2 Multi-Pathing (L2MP) Technologies (contd.)
  - A link-state routing protocol usually requires that each node on a network be addressable
  - Each node uses the link-state routing protocol to calculate the network topology and then calculates the forwarding database based on the network topology
  - Therefore, L2MP technologies need to add an addressable identifier, which is similar to an IP address on an IP network, to each device on a network
  - TRILL is a standard L2MP protocol defined by the Internet Engineering Task Force (IETF) that became popular.

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

DCN Networking Technology Evolution (contd.)

- TRILL: Basic Concepts
  - Stands for **Transparent Interconnection of Lots of Links**
  - Is implemented by devices called TRILL switches
  - TRILL combines techniques from bridging and routing, and is the application of link-state routing to L2 networks
  - To apply link-state routing protocols to Ethernet networks, a frame header needs to be added to an Ethernet header for the addressing of the link-state routing protocols
  - TRILL uses MAC in TRILL in MAC encapsulation
    - I.e. In addition to the original Ethernet header, a TRILL header that provides an addressing identifier and an outer Ethernet header used to forward a TRILL packet on an Ethernet network are added



0 15 31

Outer Destination MAC Address		Outer Ethernet header
	Outer Source MAC Address	
Ethertype = 802.1Q	Outer VLAN Tag Information	
Ethertype = TRILL	V | R | M | Op-Length | Hop-Count	TRILL header
Egress RB Nickname	Ingress RB Nickname	
Options		
Inner Destination MAC Address		Inner Ethernet header
	Inner Source MAC Address	
Ethertype = 802.1Q	Inner VLAN Tag Information	
Ethertype of Original Ethernet Payload		Payload
Original Ethernet Payload		
New FCS		

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

DCN Networking Technology Evolution (contd.)

- Disadvantages of L2MP Technologies
  - Limited number of tenants: Similar to xSTP, TRILL uses VLAN IDs to identify tenants. Because the VLAN ID field has only 12 bits, a TRILL network supports only about 4000 tenants.
    - a solution to the tenant problem was considered at the beginning of TRILL design. A field was reserved in the TRILL header for tenant identification, but the problem has not yet been resolved because the protocol has not been continuously evolved.
  - Increased deployment costs: L2MP technologies introduce new forwarding identifiers or add new forwarding processes, which inevitably requires the upgrade of forwarding chips.
  - Mechanism-related challenges: The Operations, Administration, and Maintenance (OAM) mechanism and multicast mechanism of TRILL have not been defined into formal standards, restricting further protocol evolution

Introduction of NVO3 technologies (like VXLAN) eventually led to the downfall of L2MP technology

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

Meanwhile, what is changing in the DCN?

- As DCNs scale and traffic patterns change to more east-west traffic….
- …DCN topologies evolve towards Clos Topology and its variants….
- ….DCN protocols evolve towards use of L3 technologies (IP & associated networking/control plane protocols)
  - Refer introduced case studies

NVO3 & Overlay Networks

- Virtual chassis, L2MP, and multi-chassis link aggregation technologies can solve problems of xSTP technologies. However, these technologies are fundamentally traditional network technologies and are still hardware-centric.
- 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
  - Allows the underlay network to be flexibly expanded
  - Facilitates SDN architecture deployment
    - SDN controller is not required to consider the underlay network architecture and can flexibly deploy services on the overlay network
- 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

Where else have you seen the use of tunnels??

- 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

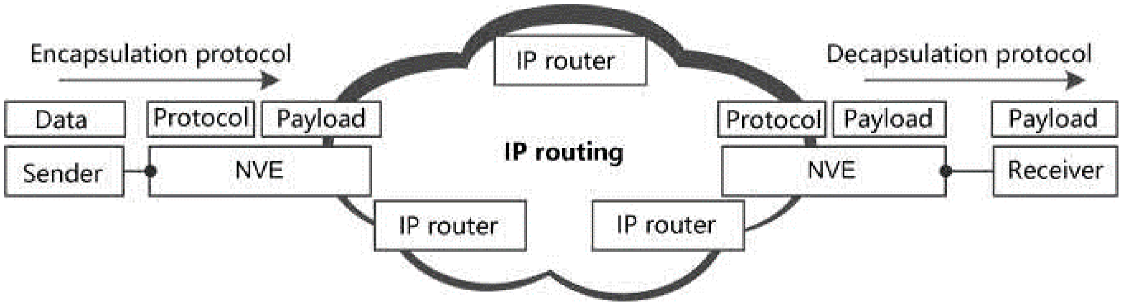
DCN Networking Technology Evolution (contd.)

- 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 VXLAN network identifier (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.)

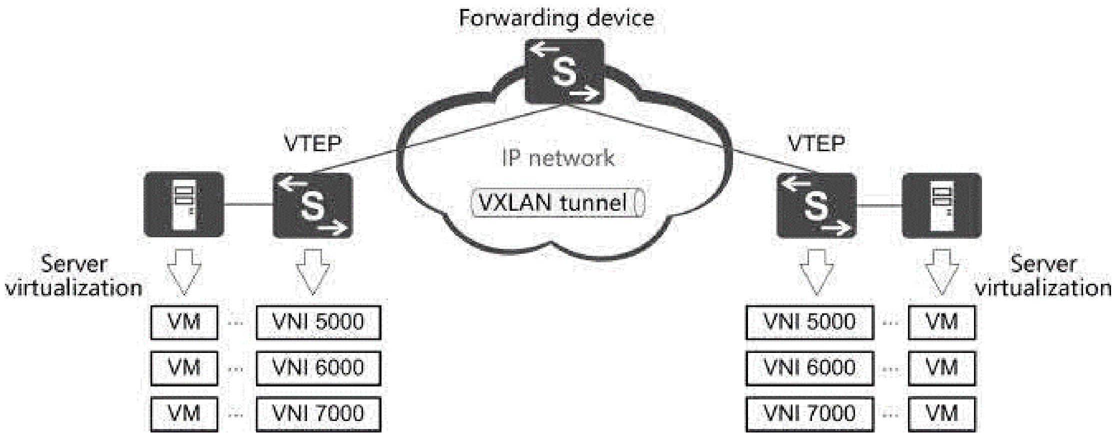
- L2MP Vs NVO3 Technologies
  - To some extent, NVO3 and L2MP technologies are similar.
  - They both build an overlay network on the physical network
  - The difference is that L2MP technologies add a new forwarding identifier to the original Layer 2 network, thereby requiring that chips on hardware 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

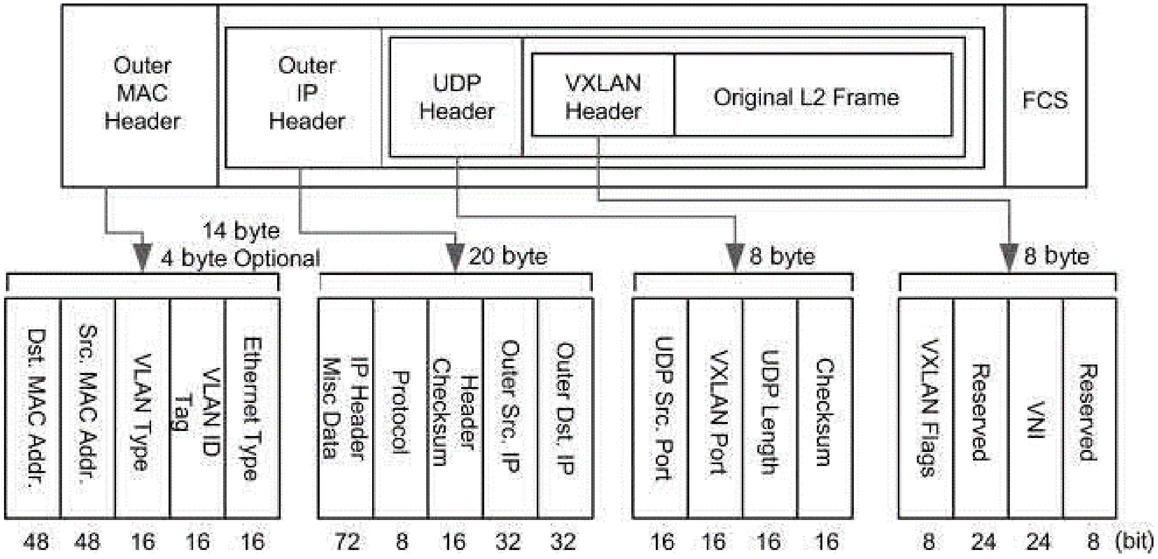
VXLAN Benefits

- **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.

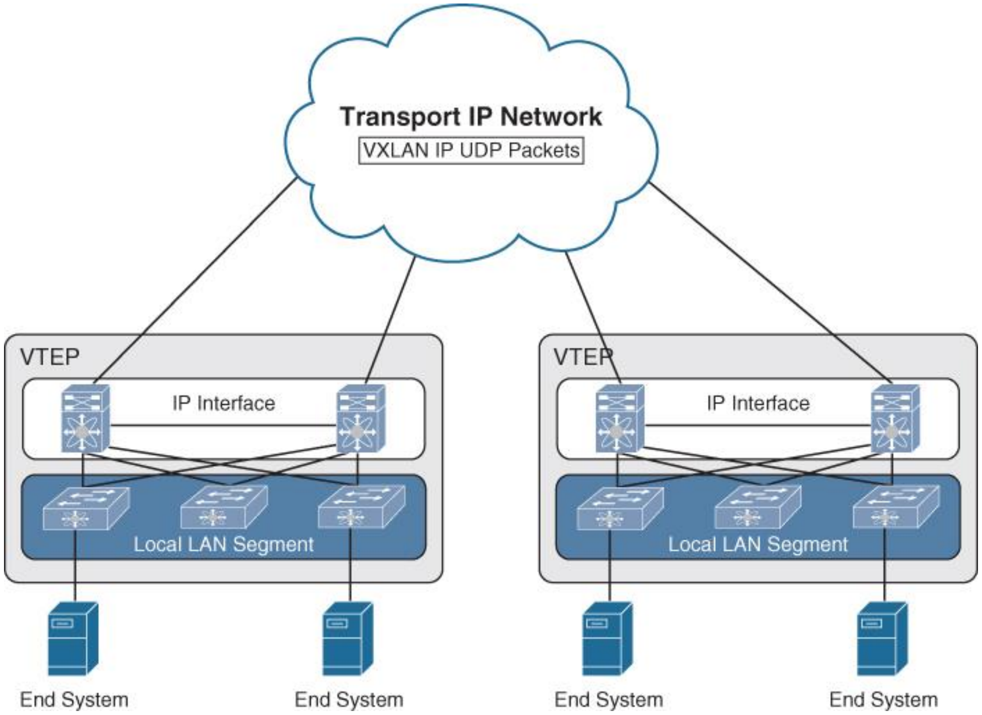


- In the UDP header,
  - the destination port number has a fixed value of 4789
  - The source port number is the hash value of the original Ethernet frame
- Outer IP (/MAC) header
  - source IP (/MAC) address specifies the IP (/MAC) address of the VTEP, where the source VM belongs.
  - destination IP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

More on this later…..

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

Thank You!