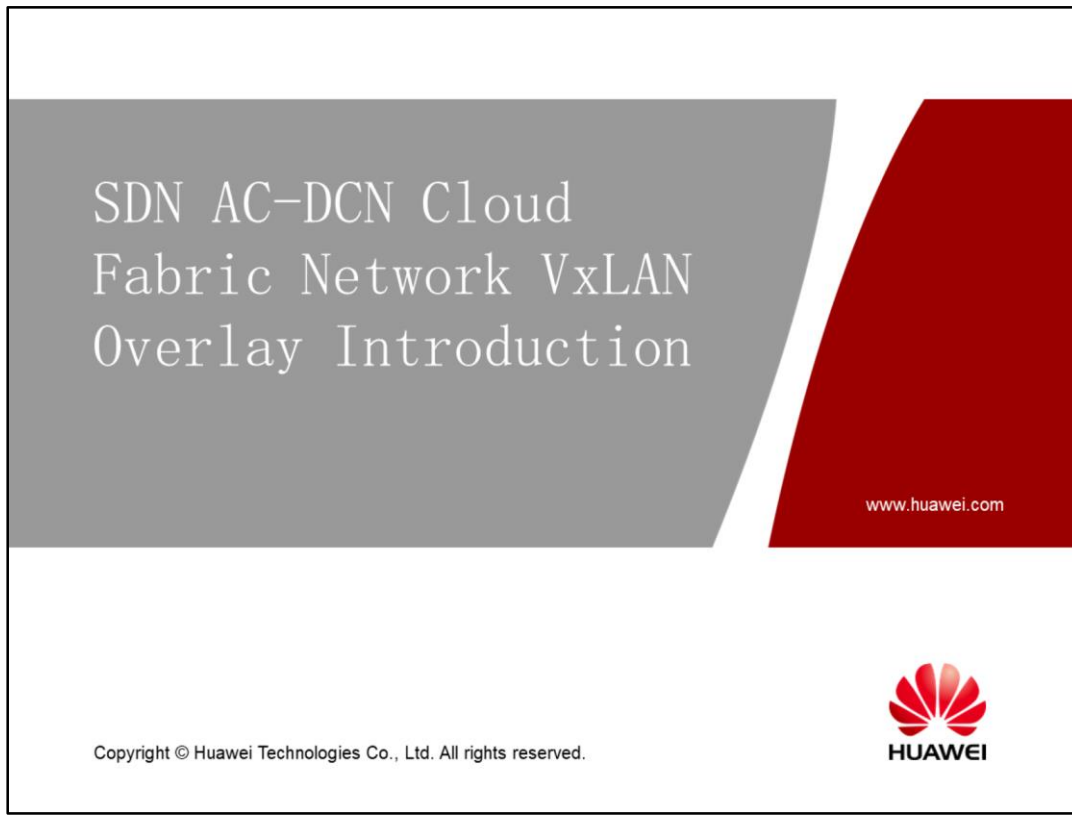
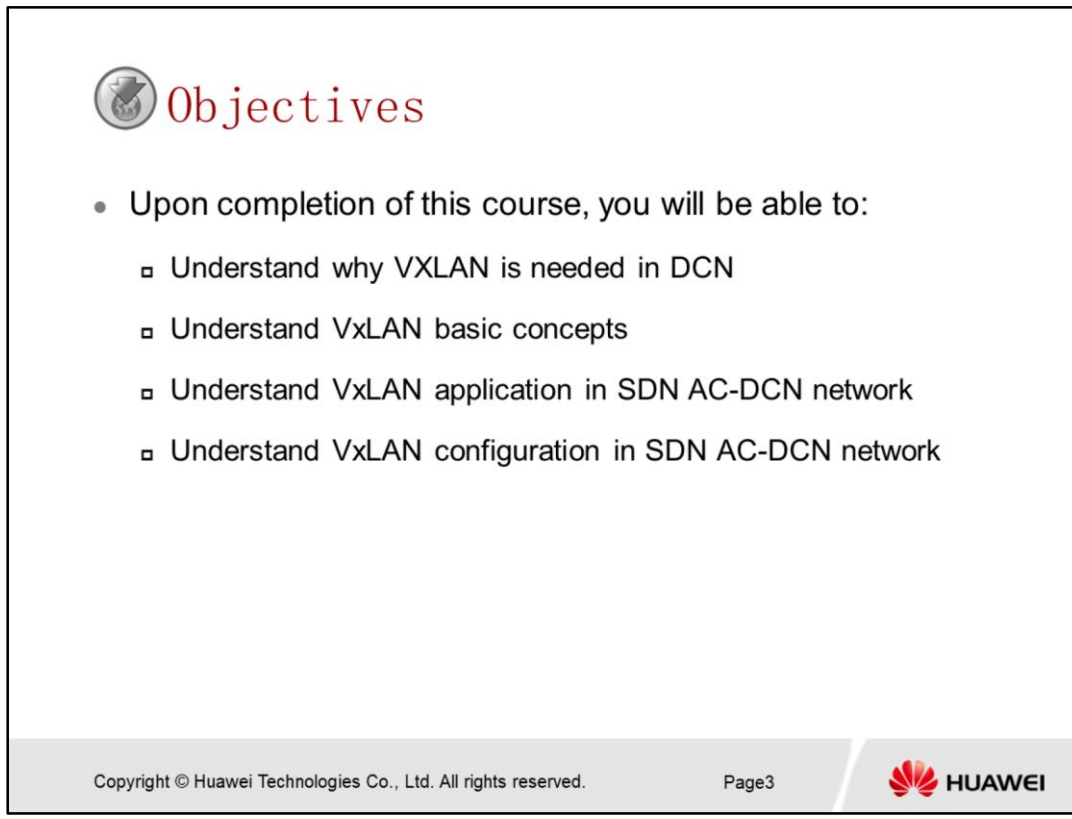
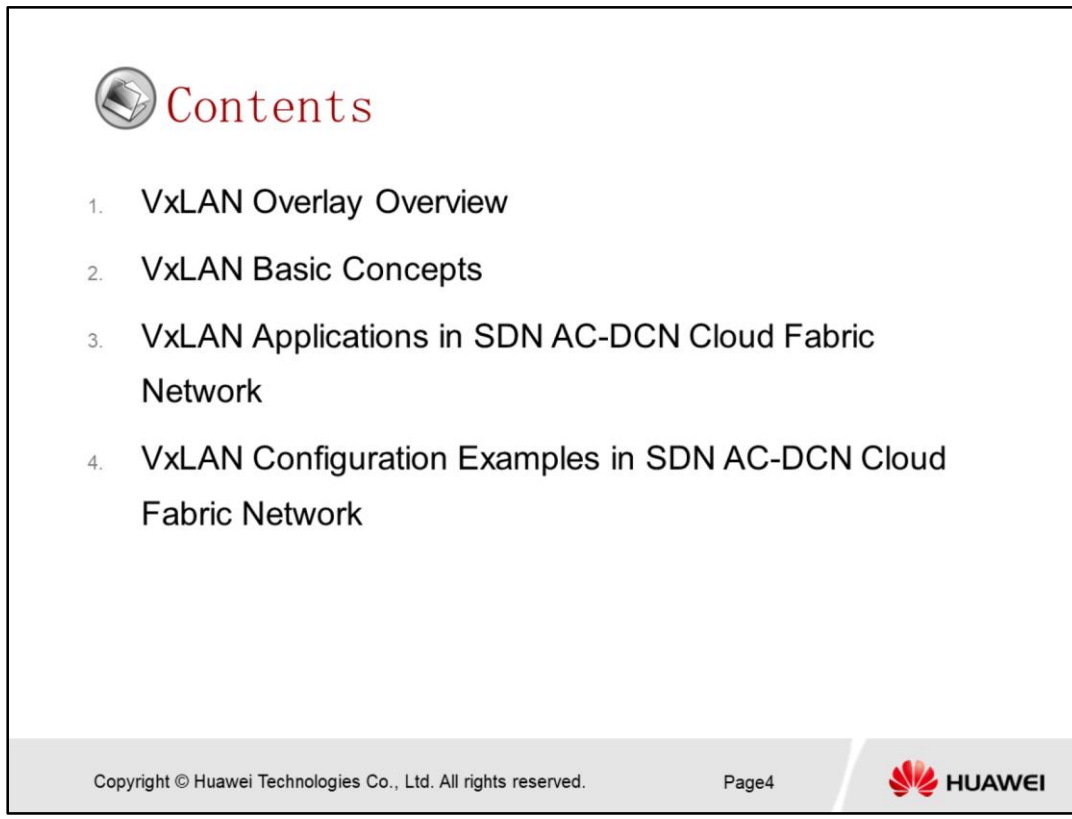
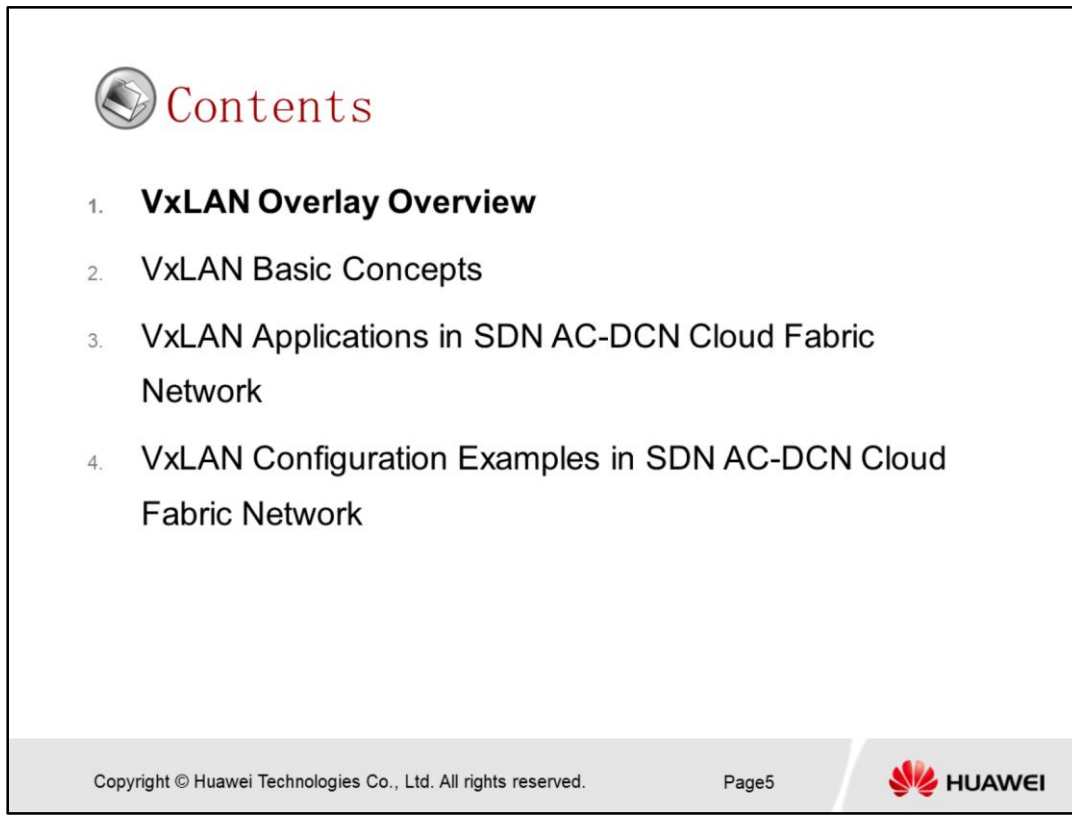


Foreword

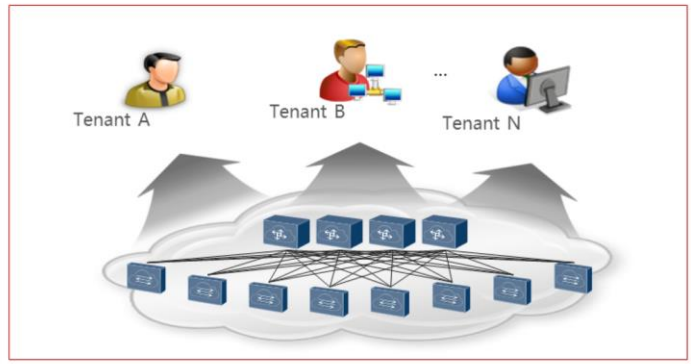
- VxLAN is a very important overlay technology used throughout the whole AC-DCN network; thus, understanding the concepts of VxLAN and its applications and configuration in AC-DCN network is crucial before we go into the end-to-end service deployment through the AC-DCN underlay deployment.







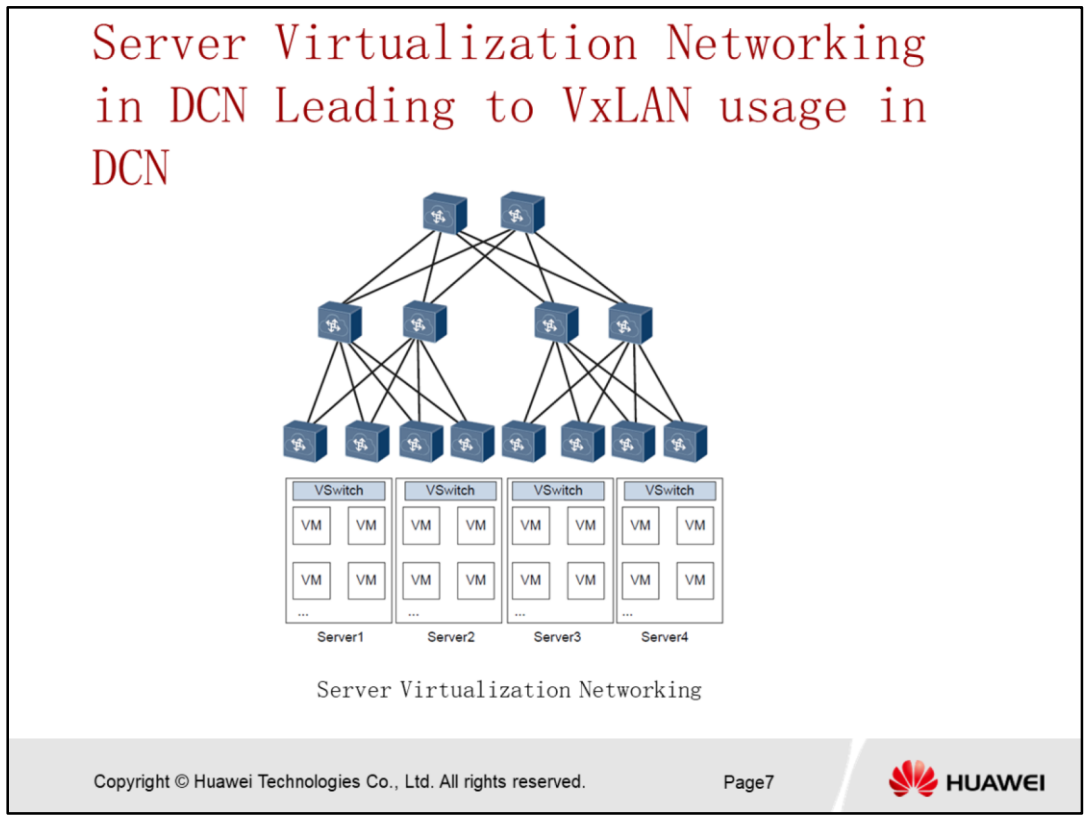
Why VxLAN?



- VxLAN is a technology which is much needed in the DCN network nowadays prior to the demand of data center multi-tenants scenarios. The traditional VLAN technology which can only support up to maximum 4096 VLANs is definitely not sufficient in indentifying user on Layer network in DCN. Thus, VxLAN, serves as a overlay tunnel technology is implemented to solve this requirement.

- The VLAN tag field, as defined in IEEE 802.1Q, has only 12 bits, and can only identify a maximum of 4096 VLANs, making it insufficient for identifying users on large Layer 2 networks;

- VXLAN uses a VXLAN network identifier (VNI) field similar to the VLAN ID field defined in IEEE 802.1Q. The VNI field has 24 bits and can identify a maximum of 16M VXLAN segments.



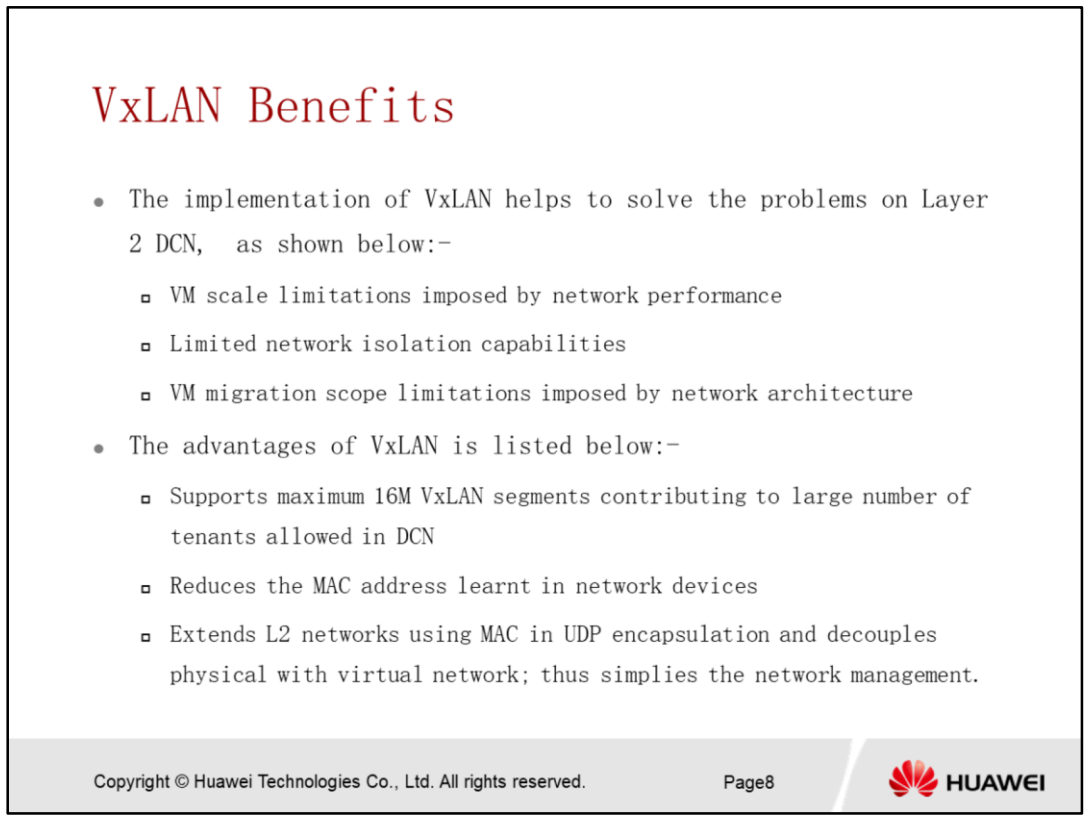
- On the network shown in the diagram above, one server is virtualized into multiple virtual machines (VMs), each of which acts as a host. However, the exponential increase in the number of hosts leads to the following problems on a virtual network:

1. **The number of VMs is limited by network performance.**
   - On a large Layer 2 network, data packets are forwarded based on MAC address entries. Therefore, the number of VMs supported on the network depends on the MAC address table size.

2. **Network isolation capabilities are limited.**
   - Most networks use VLANs or virtual private networks (VPNs) for network isolation. However, these two network isolation technologies have the following limitations on large scale virtualized networks:
     - The VLAN tag field, as defined in IEEE 802.1Q, has only 12 bits, and can only identify a maximum of 4096 VLANs, making it insufficient for identifying users on large Layer 2 networks.
     - VLANs or VPNs cannot support dynamic network adjustment on traditional Layer 2 networks.

3. **VM migration scope is limited by the network architecture.**
   - After VMs are started, they may need to be migrated from one server to

another due to server resource problems (for example, CPU overload or insufficient memory). To ensure uninterrupted services during VM migration, the IP and MAC addresses of VMs must remain unchanged. To meet this requirement, the service network must be a Layer 2 network that provides multipath redundancy and reliability.



- VXLAN addresses the above problems on large Layer 2 networks as follows:

1. VM scale limitations imposed by network performance

- VXLAN encapsulates data packets sent from VMs into UDP packets and encapsulates IP and MAC addresses used on the physical network into outer headers. The network is onlyaware of the encapsulated parameters. This greatly reduces the number of MAC addressentries required on large Layer 2 networks.

2. Limited network isolation capabilities

- VXLAN uses a VXLAN network identifier (VNI) field similar to the VLAN ID fielddefined in IEEE 802.1Q. The VNI field has 24 bits and can identify a maximum of 16M [(2^24-1)/1024^2] VXLAN segments.

3. VM migration scope limitations imposed by network architecture

- When VXLAN is used to construct a large Layer 2 network, VM IP and MAC addresses can remain unchanged after VM migration.

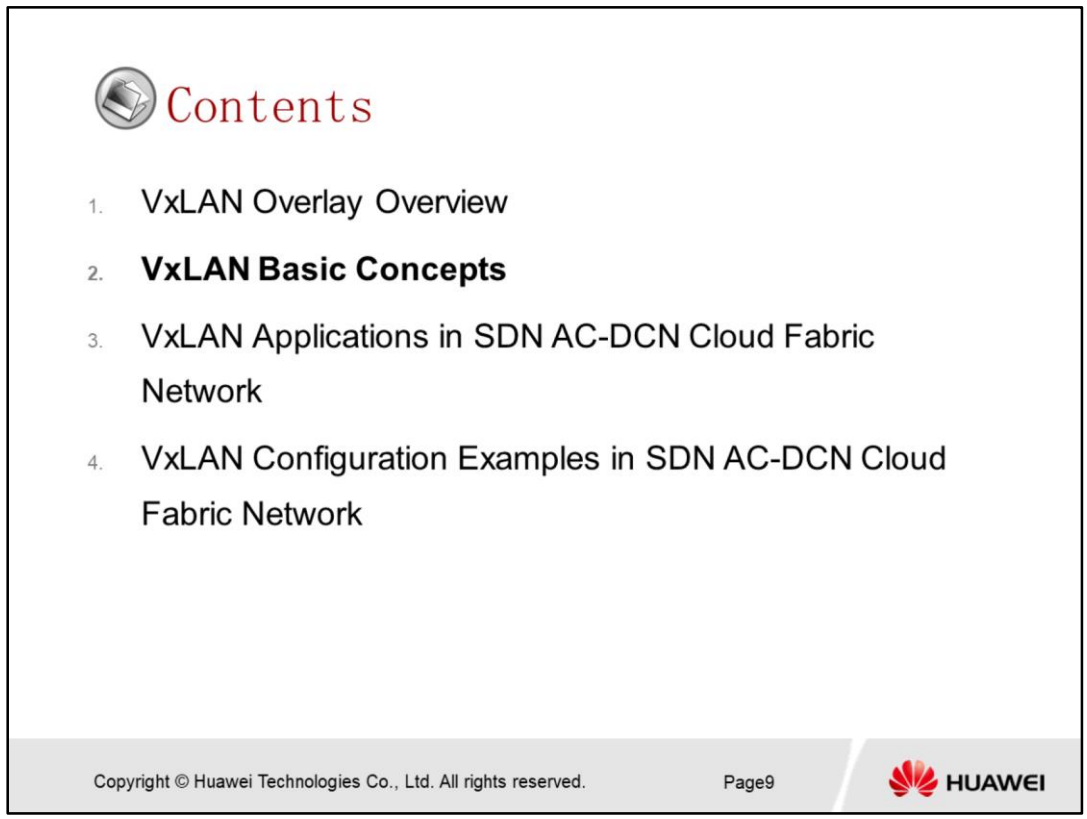
- **The advantages of VxLAN are listed below:-**

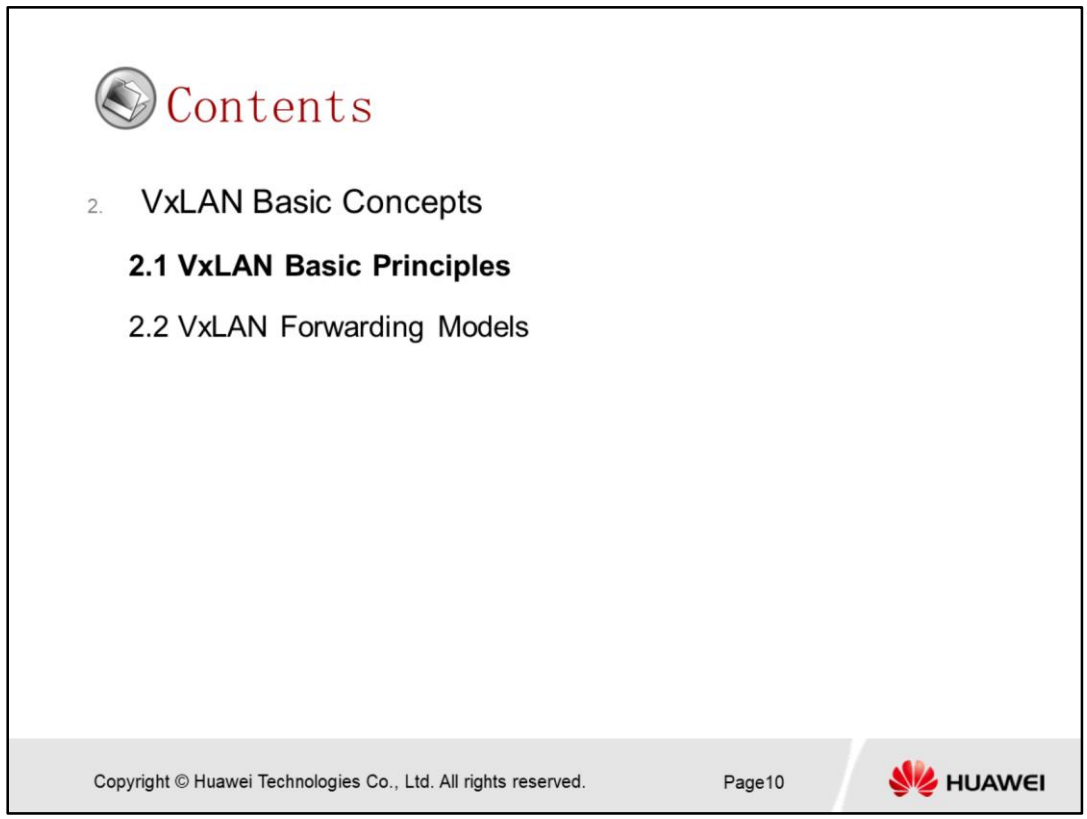
1. Supports a maximum of 16M VXLAN segments with 24-bit VNIs, so a data center can accommodate a large number of tenants.

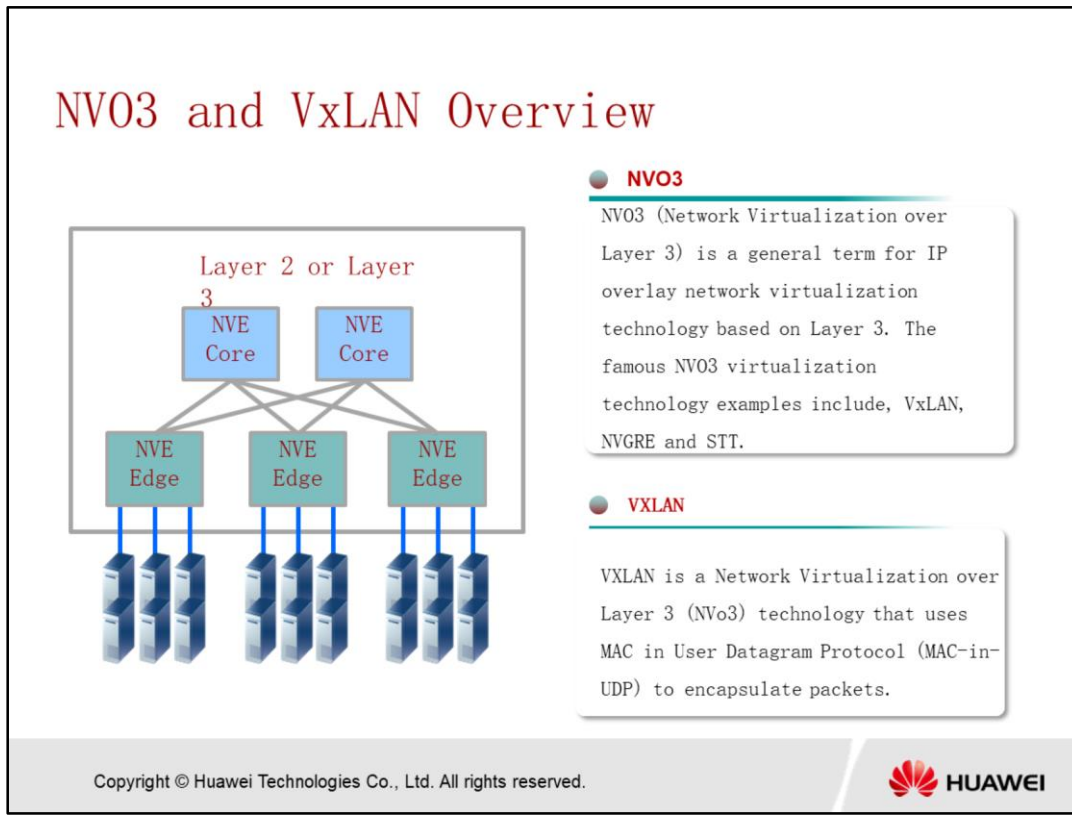
2. Reduces the number of MAC addresses that network devices need to learn and enhances network performance because only devices at the edge of the VXLAN network need to identify VM MAC addresses.

3. Extends Layer 2 networks using MAC-in-UDP encapsulation and decouples physical and virtual networks. Tenants can plan their own virtual networks, without being limited by the physical network IP addresses or broadcast domains. This greatly simplifies network

management.

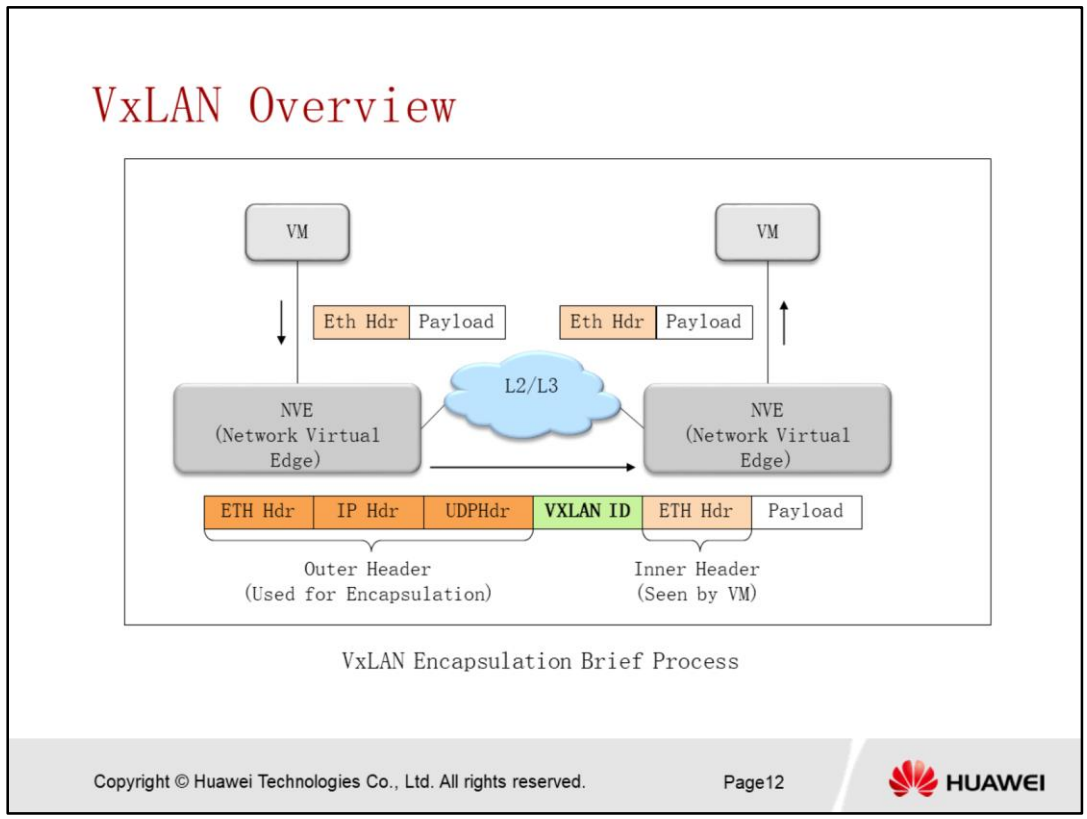




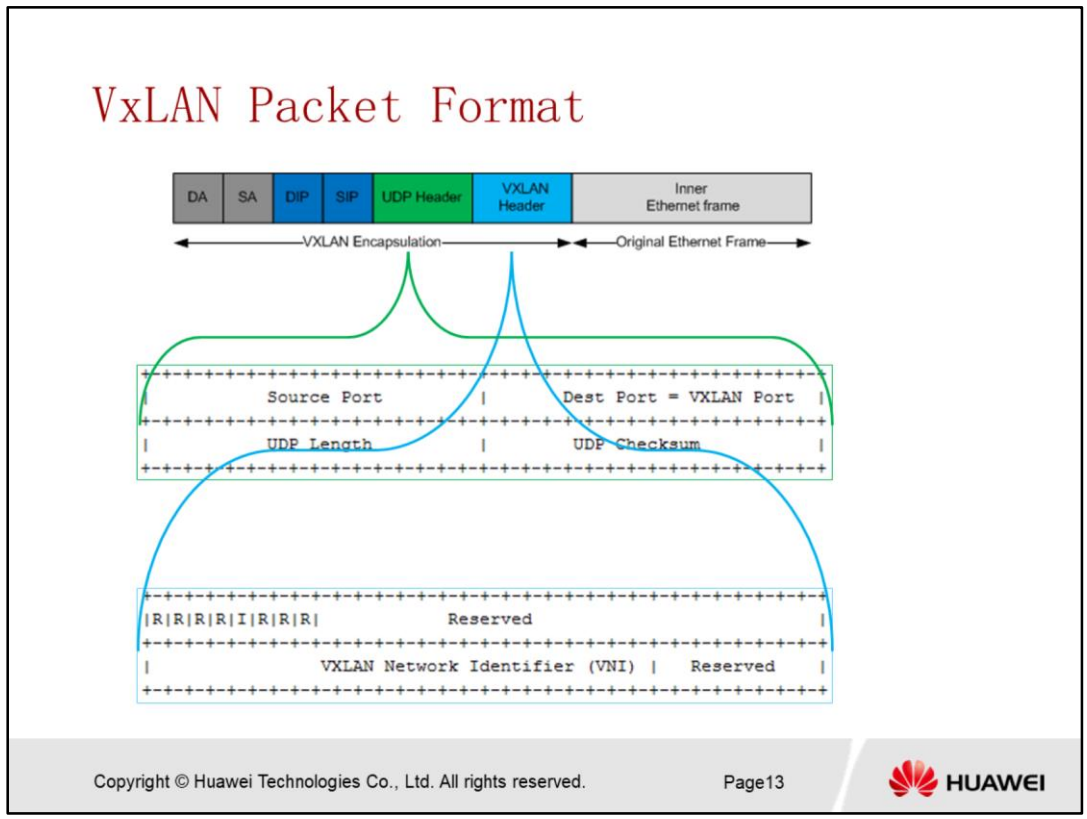


- NVO3 (Network Virtualization over Layer 3) is a general term for IP overlay network virtualization technology based on Layer 3. The famous NVO3 virtualization technology examples include, VxLAN (Virtual extensible Local Area Network) , NVGRE (Network Virtualization using Generic Routing Encapsulation) and STT (Stateless Transport Tunneling Protocol).
- Devices running NVO3 is called NVE (Network Virtualization Edge), which is located at the edge of the overlay network to realize L2 and L3 virtualization function.

- VxLAN is considered as the most widely used NVO3 technologies nowdays.



- VxLAN (Virtual Extensible LAN) is a Network Virtualization over Layer 3 (NVo3) technology that uses MAC in User Datagram Protocol (MAC-in-UDP) to encapsulate packets.
- In SDN Cloud Fabric solution, VxLAN technology is realized through AC in order to build the overlay network over L3 fabric. It uses MAC over UDP encapsulation to achieve the requirements of multi-tenants scenarios in data center network virtualization solution.



- The explanation of the VxLAN packet format is as below:-

- **VXLAN header**
  - Flags (8 bits): The value is 00001000. (R flag must be set to 0 and I flag must be set to 1)
  - VNI (24 bits): used to identify a VXLAN segment.
  - Reserved fields (24 bits and 8 bits): must be set to 0.

- **Outer UDP header**
  - The destination UDP port number is 4789. The source port number is the hash value calculated using parameters in the inner Ethernet frame header.

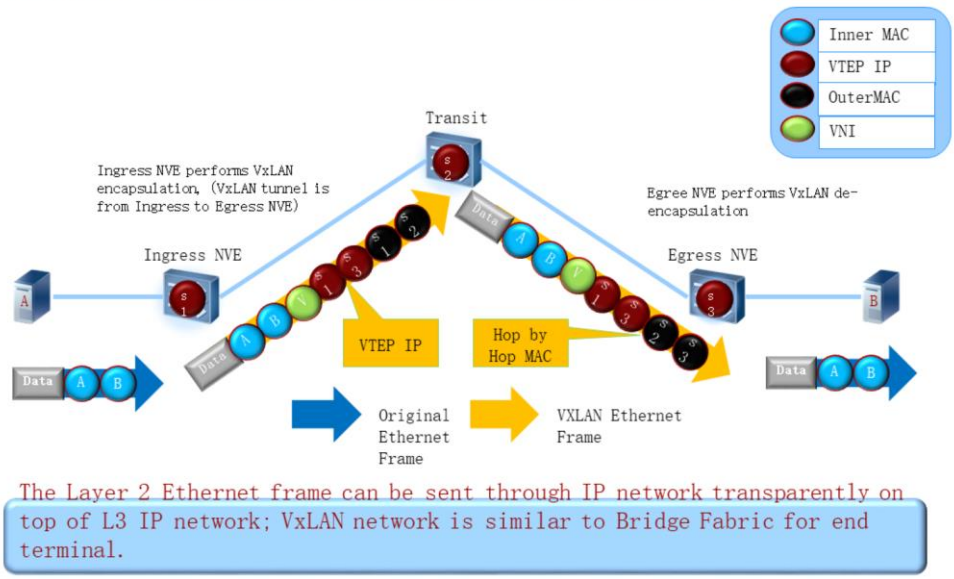
- **Outer IP header**
  - In the outer IP header, the source IP address is the IP address of the VTEP where the sender VM resides; the destination IP address is the IP address of the VTEP where the destination VM resides.

- **Outer Ethernet header**
  - SA: specifies the MAC address of the VTEP where the sender VM resides.
  - DA: specifies the next-hop MAC address in the routing table of the VTEP

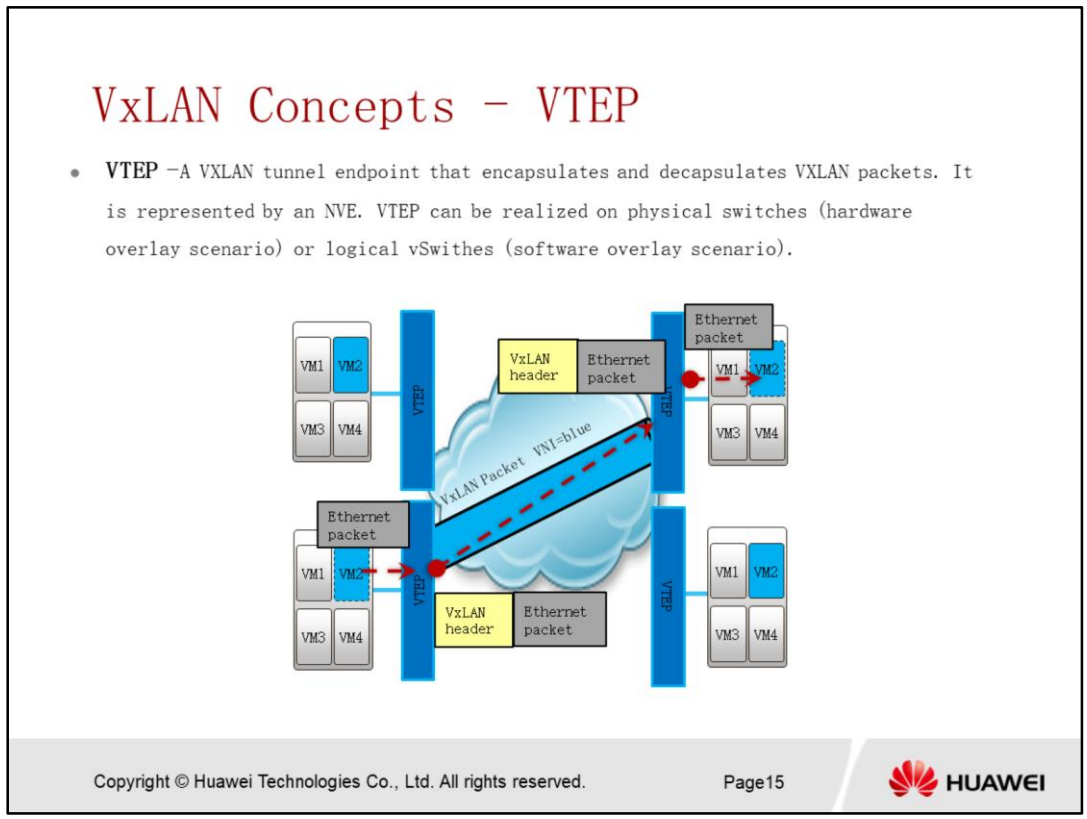
where the destination VM resides.

- VLAN Type: This field is optional. The value of this field is 0x8100 when the packet has a VLAN tag.
- Ethernet Type: specifies the type of the Ethernet frame. The value of this field is 0x0800 when the packet type is IP.

VxLAN Packet Encapsulation Process



The Layer 2 Ethernet frame can be sent through IP network transparently on top of L3 IP network; VxLAN network is similar to Bridge Fabric for end terminal.

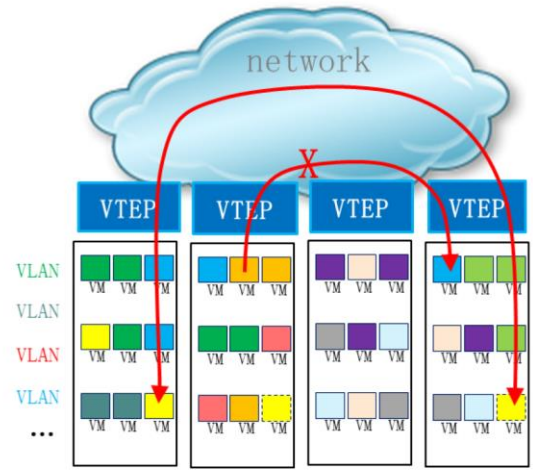


- VTEP stands for Virtual Tunnel End Point.
- VXLAN tunnel endpoints that are deployed on NVE nodes and responsible for VXLAN packet encapsulation and decapsulation. VTEPs are connected to the physical network and assigned IP addresses (VTEP IP) of the physical network. VTEP IP addresses are independent of the virtual network. A local VTEP IP address and a remote VTEP IP address identify a VXLAN tunnel.

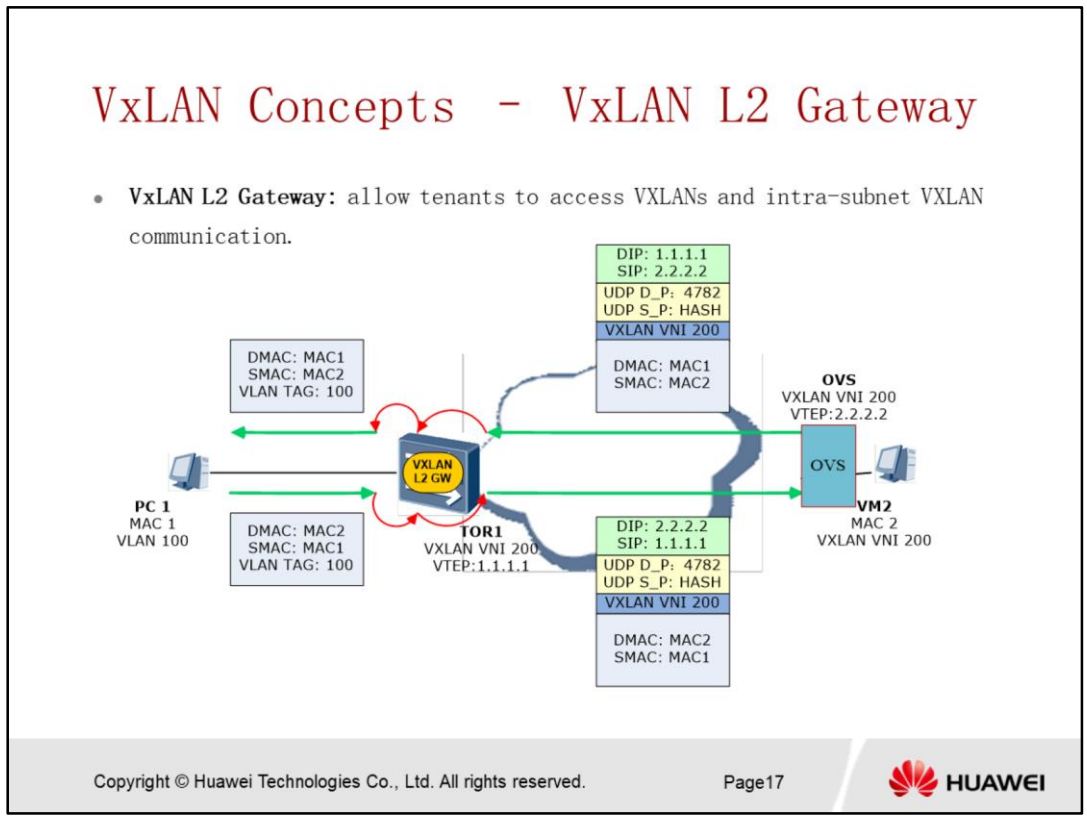
- There are 2 types of configuration for VTEP configuration, as per listed below:-
  - **VTEP on physical TOR (Hardware overlay)**
    - Advantage : High performance, line speed forwarding capability
    - Disadvantage： limited by chip manufacturers; TOR switch needs virtual perceptions.
  - **VTEP on vSwitch (Software overlay)**
    - Advantage : TOR switch does not need to virtual perception, achieve simply
    - Disadvantage: Consume the host hardware resources, impact the host performance

VxLAN Concepts – VNI

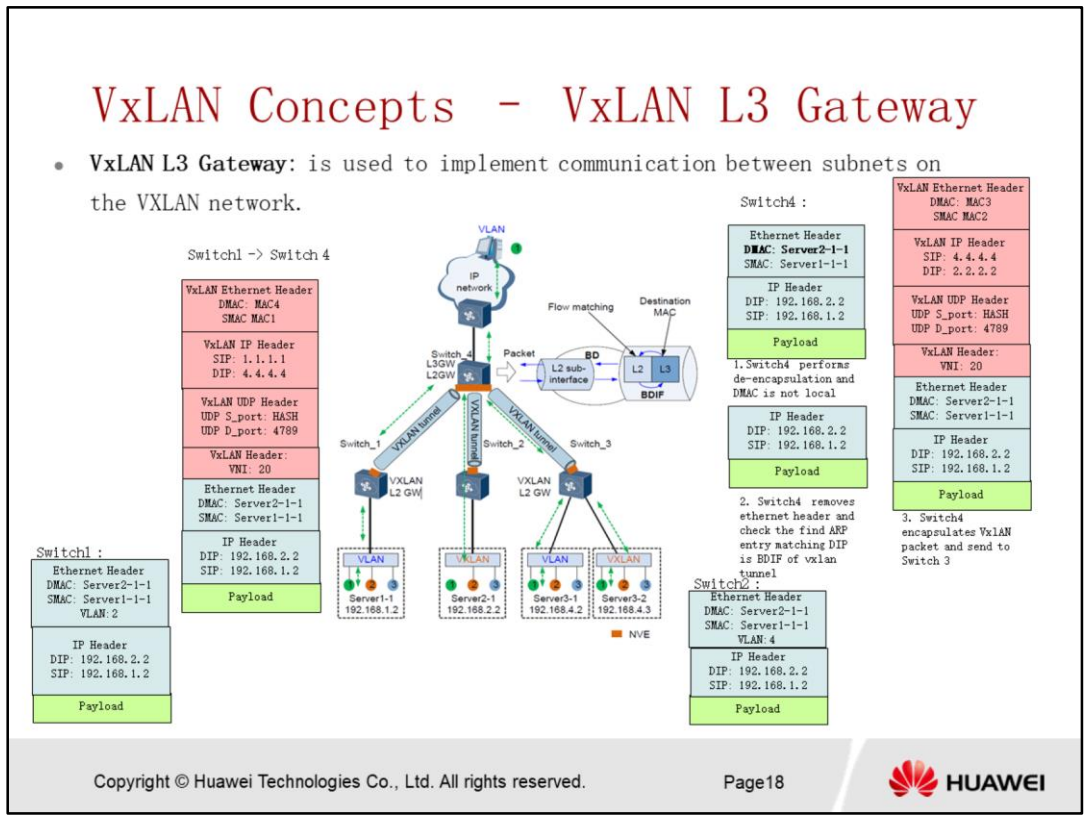
- **VNI**– A VXLAN segment identifier similar to a VLAN ID. VMs on different VXLAN segments cannot communicate directly at Layer 2.



- VNI is VXLAN network identifier that identifies a VXLAN segment.
  - VNI segment : 24 bits;
  - Within the same VM VNI can communicate directly;
  - Different VM VNI cannot communicate directly, must use Layer 3 Gateway to communicate.



- **Layer 2 gateway:**
- After a VXLAN Layer 2 gateway receives user packets, it forwards the packets in different processes based on the packets' destination MAC address type:
  - If the MAC address is a broadcast, unknown unicast, and multicast (BUM) address, the Layer 2 gateway follows the BUM packet forwarding process.
  - If the MAC address is a unicast address, the Layer 2 gateway follows the unicast packet forwarding process.



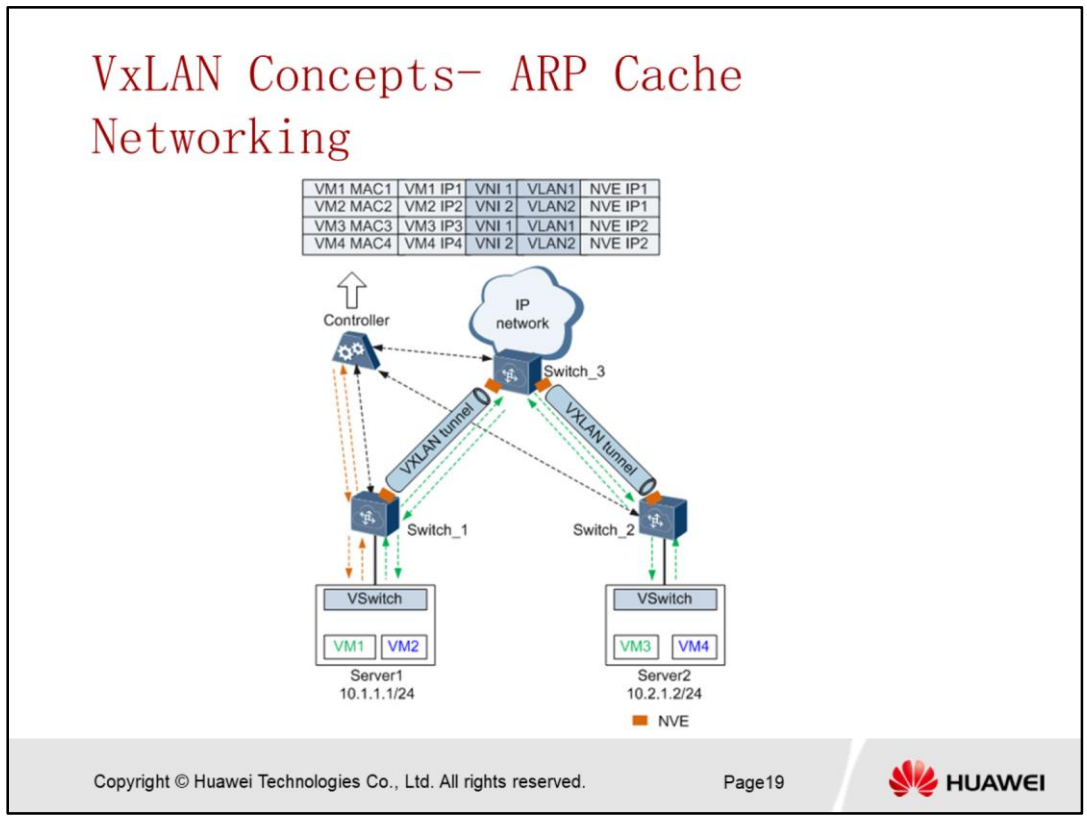
- VxLAN communication between different network segment, and the communication between VxLAN network and non VxLAN network requires IP routing. Thus, Bridge Domain (BD) needs to be established on L3 gateway; VNI is mapped to a BD in 1:1 ratio; BDIF interface is configured based on different BD to allow L3 communication. BDIF is similar to Vlanif in VLAN concept.

1. After Switch_4 (VXLAN Layer 2 gateway) receives a VXLAN packet, it decapsulates the packet and checks whether the destination MAC address in the inner packet is the gateway MAC address.
   - If so, Switch_4 forwards the packet to the Layer 3 gateway on the destination network segment. Go to Step **2.**
   - If not, Switch_4 searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain.

2. Switch_4 functions as a VXLAN Layer 3 gateway to remove the Ethernet header of the inner packet and parse the destination IP address. Switch_4 searches for the ARP entry matching the destination IP address and checks the destination MAC address, VXLAN tunnel's outbound interface, and VNI.
   - If the VXLAN tunnel's outbound interface and VNI are not found, Switch_4 forwards the packets at Layer 3.

- If the VXLAN tunnel's outbound interface and VNI are found, go to Step **3.**

1. Switch_4 functions as a VXLAN Layer 2 gateway to encapsulate the VXLAN packet again by adding the gateway's MAC address as the source MAC address in the Ethernet header.

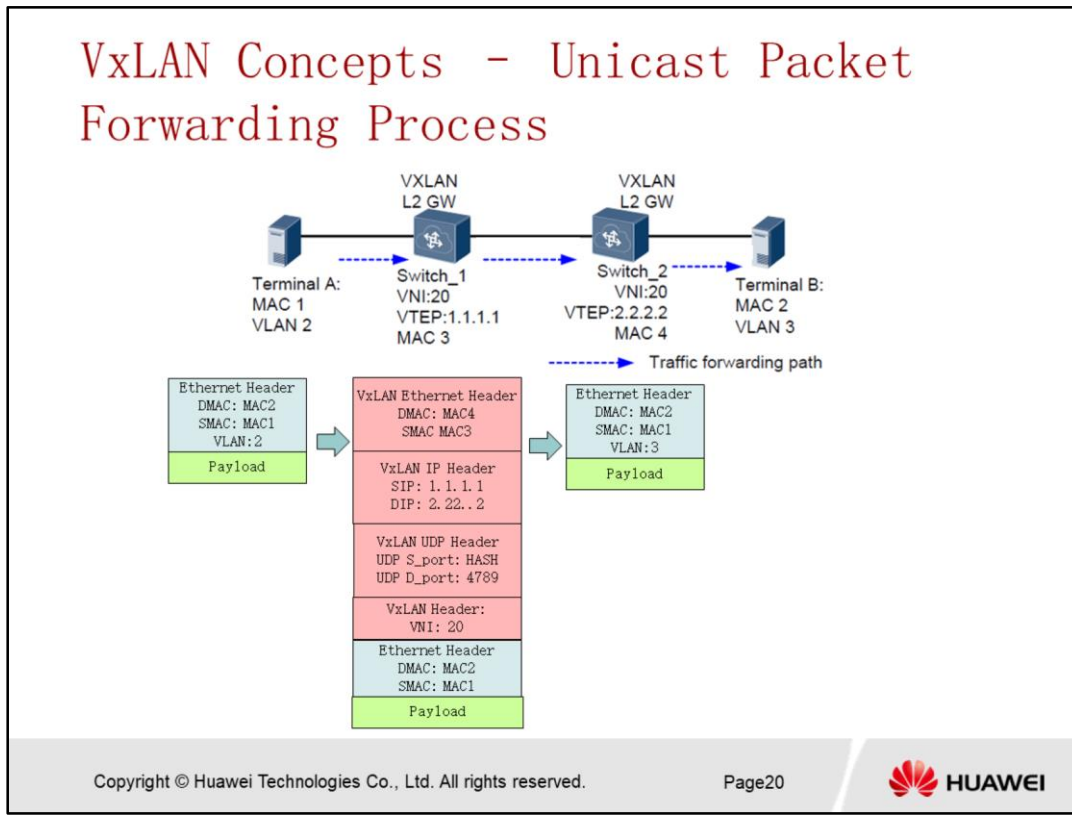


- Traditionally, a terminal needs to send a broadcast ARP request message before it communicates with another terminal for the first time. For example, on the network shown in Figure above, VM1 needs to send an ARP request message to VM3 when it needs to communicate with VM3 for the first time. The ARP request message is broadcast on the Layer 2 network. After receiving the ARP request message, VM3 sends a unicast ARP reply message to VM1.

- To prevent broadcast storms caused by broadcast ARP request messages, ARP cache can be enabled on the controller, as shown in the figure above. After that, the following process occurs when VM1 sends an ARP request message to request VM3's MAC address:

1. VM1 sends an ARP request message, with the source MAC address MAC1, source IP address IP1, destination MAC address FF-FF-FF, and destination IP address IP3.
2. After receiving the ARP request message, the NVE1 sends the message to the controller through an OpenFlow channel.
3. The controller searches the user information database based on IP3 and obtains the MAC address (MAC3) of VM3.
4. The controller sends an ARP reply message to the NVE1 through the OpenFlow channel.
5. After receiving the ARP reply message, the NVE1 sends the message to VM1

through the outbound interface specified by the controller, which is the inbound interface that receives the ARP request message



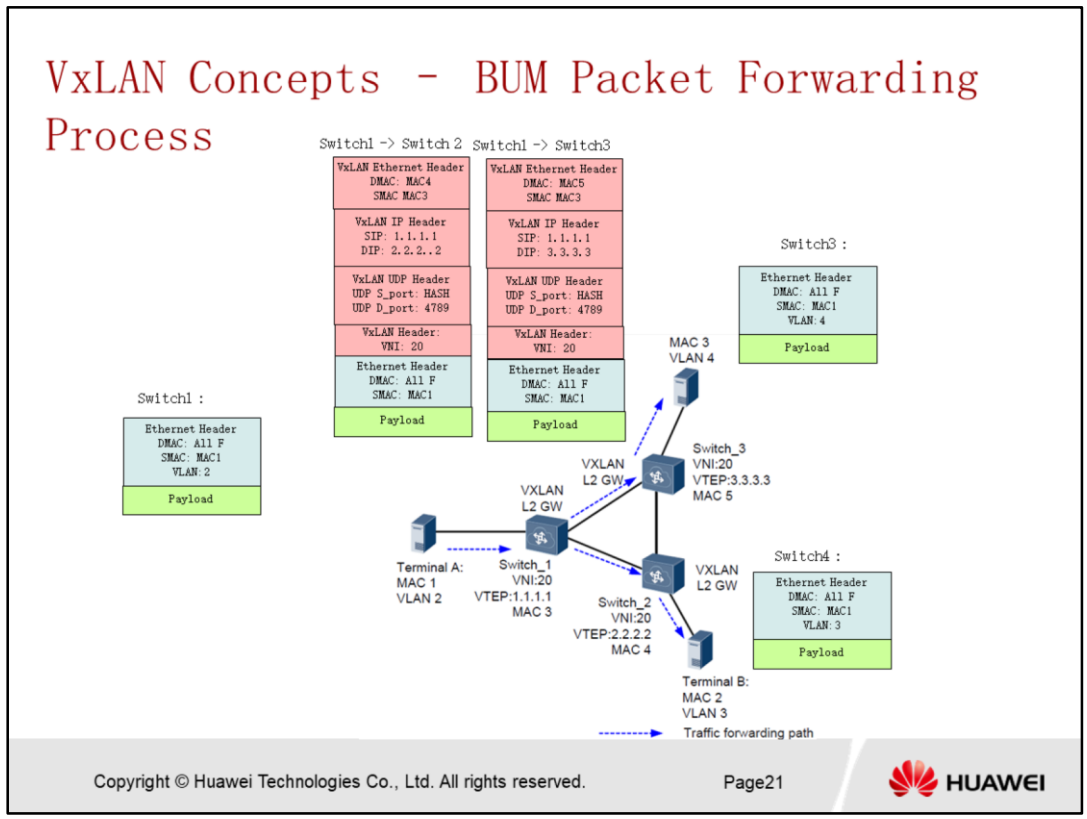
1. After Switch_1 receives a packet from terminal A, Switch_1 determines the Layer 2 broadcast domain of the packet based on the access interface and VLAN information carried in the packet, and checks whether the destination MAC address is a known unicast address.
    - If the destination MAC address is a known unicast address, Switch_1 checks whether the destination MAC address is a local MAC address.
        - If so, Switch_1 processes the packet.
        - If not, Switch_1 searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain. Go to Step 2.
    - If the destination MAC address is not a known unicast address, Switch_1 broadcasts the packet in the Layer 2 broadcast domain. Go to Step 2.

2. The VTEP on Switch_1 performs VXLAN tunnel encapsulation based on the outbound interface and encapsulation information, and forwards the packet.

3. After the VTEP on Switch_2 receives the VXLAN packet, it checks the UDP destination port number, source and destination IP addresses, and VNI of the packet to determine its validity. The VTEP obtains the Layer 2 broadcast domain based on the VNI and performs the destination MAC address is a known unicast address.
    - If the destination MAC address is a known unicast address, the VTEP searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain. Go to Step 4.

- If the destination MAC address is not a known unicast address, the VTEP checks whether the destination MAC address is a local MAC address.
  - If so, the VTEP sends the packet to Switch_2.
  - If not, the VTEP forwards the packet according to the BUM Packet Forwarding Process.

1. Switch_2 adds a VLAN tag to the packet based on the outbound interface and encapsulation information, and forwards the packet to terminal B.



- BUM = Broadcast Unknown Unicast and Multicast

1. After Switch_1 receives a packet from terminal A, Switch_1 determines the Layer 2 broadcast domain of the packet based on the access interface and VLAN information carried in the packet, and checks whether the destination MAC address is a BUM address.
   - If the destination MAC address is a BUM address, Switch_1 broadcasts the packet in the Layer 2 broadcast domain. Go to Step **2.**
   - If the destination MAC address is not a BUM address, Switch_1 forwards the packet according to the **Forwarding Process of Known Unicast Packets.**

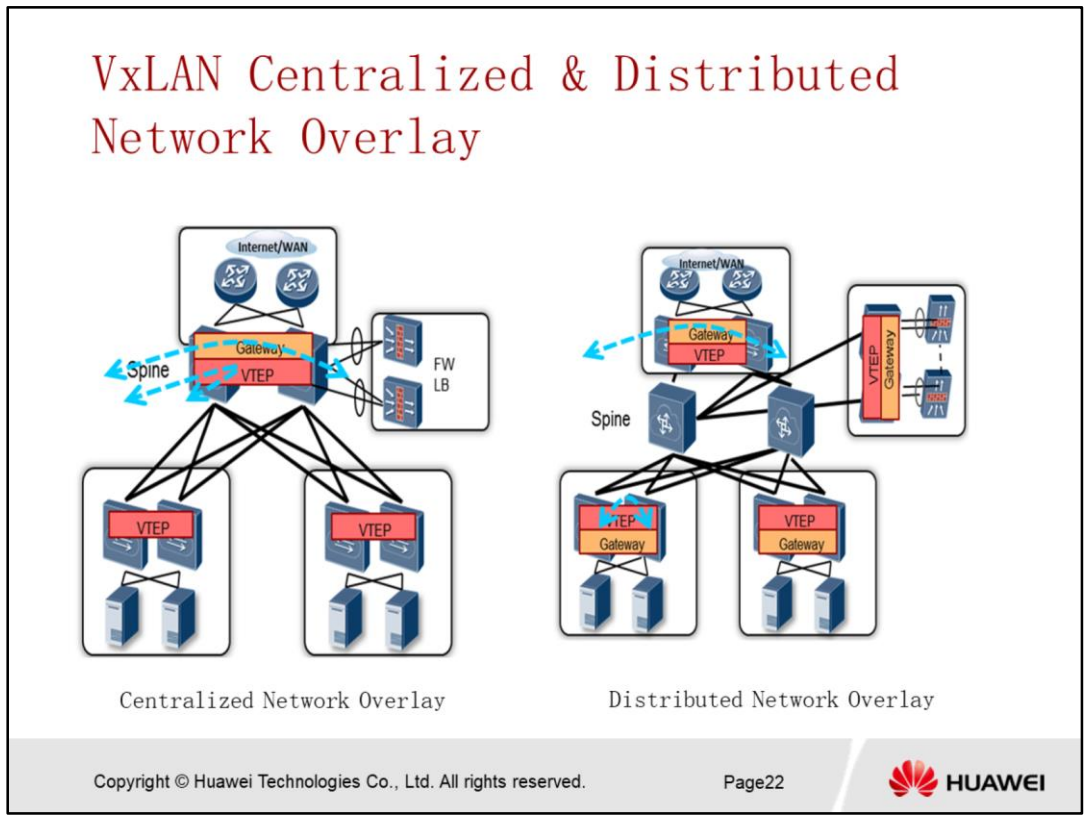
2. The VTEP on Switch_1 obtains the ingress replication list for the VNI based on the Layer2 broadcast domain, replicates the packet based on the list, and performs VXLAN tunnel encapsulation by adding the VXLAN header and outer IP header. Switch_1 then forwards the packet through the outbound interface.

3. After receiving the VXLAN packet, the VTEP on Switch_2 or Switch_3 checks the UDP destination port number, source and destination IP addresses, and VNI of the packet to determine its validity. The VTEP obtains the Layer 2 broadcast domain based on the VNI and performs VXLAN tunnel decapsulation to obtain the inner

Layer 2 packet. The VTEP then determines whether the destination MAC address is a BUM address.

- If the destination MAC address is a BUM address, the VTEP broadcasts the packet in the Layer 2 broadcast domain.
- If the destination MAC address is not a BUM address, the VTEP checks whether the destination MAC address is a local MAC address.
  - If so, the VTEP sends the packet to Switch_2 or Switch_3.
  - If not, the VTEP searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain. Go to Step **4.**

1. Switch_2 or Switch_3 adds a VLAN tag to the packet based on the outbound interface and encapsulation information, and then forwards the packet to terminal B or terminal C.



- **Centralized mode**
  - In VxLAN network. L3 gateway function is centralized on one or one group of switches
  - Leaf switches who is connected to firewall, load balancer, and various servers only serves as L2 gateway.

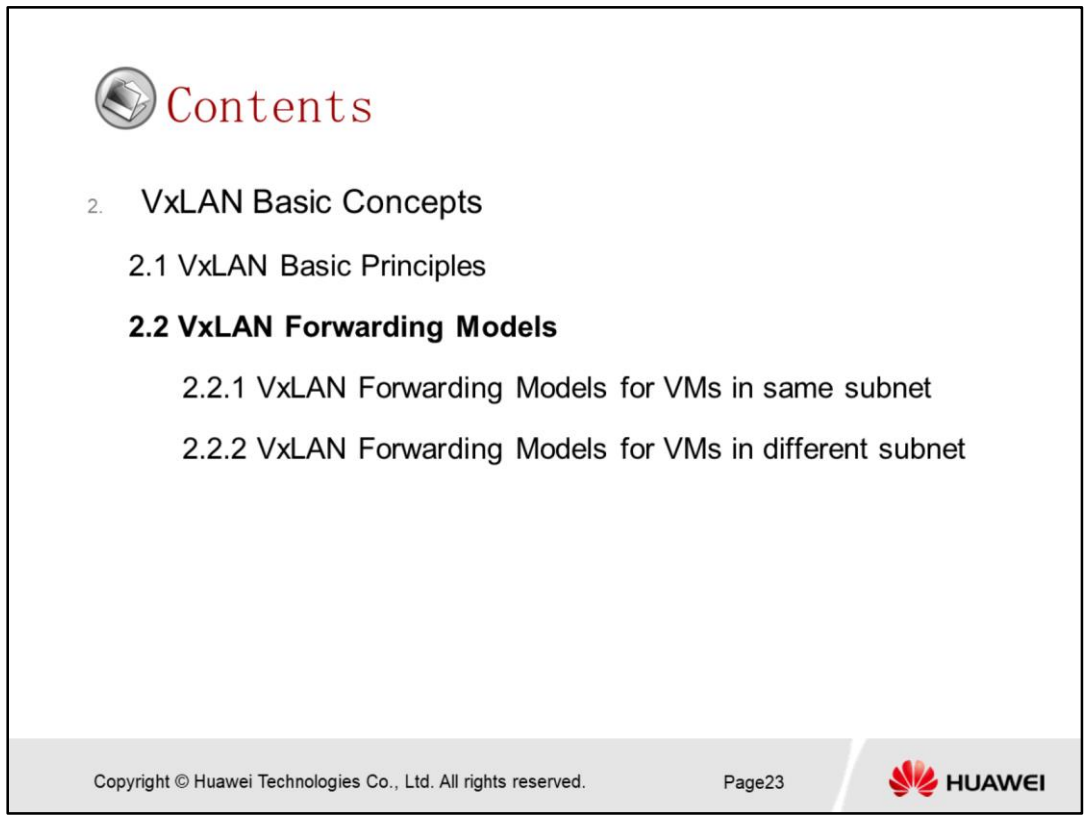
- **Distributed Mode:**
  - In Network Overlay distributed VxLAN network, all leaf physical switches are equipped with L3 gateway functions; Spine functions as traffic forwarding node, and does not function as VTEP;

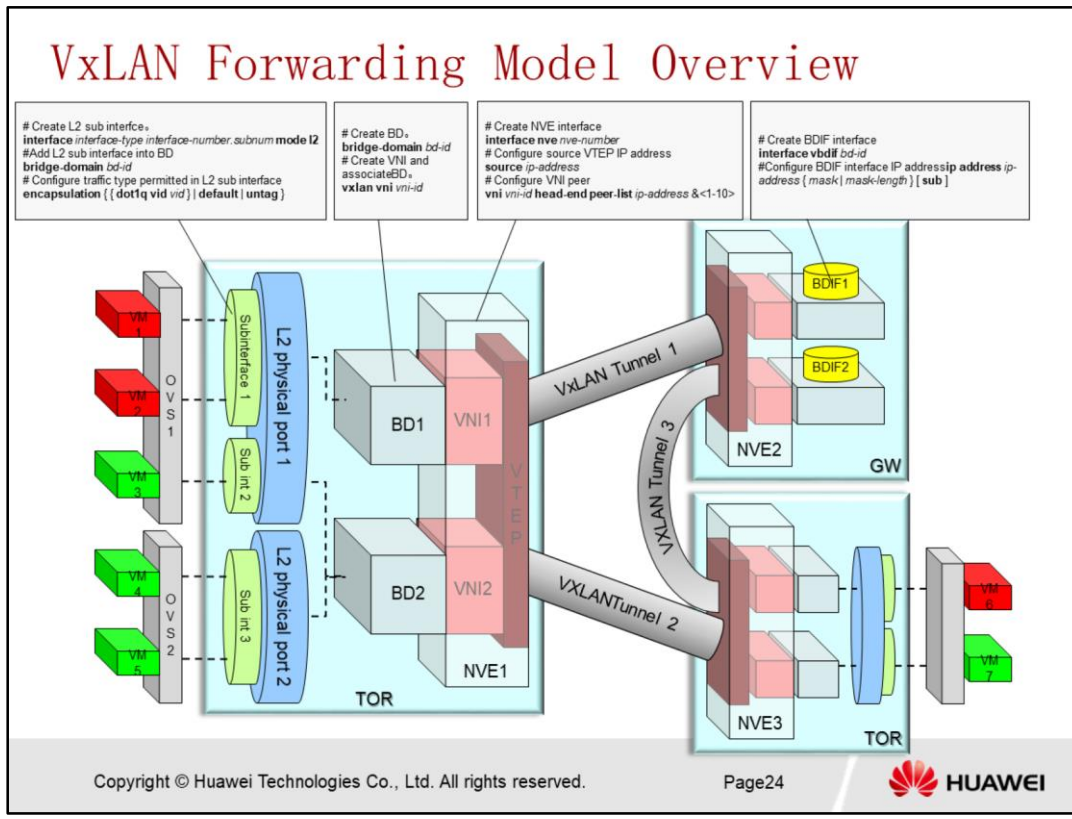
- The disadvantages of centralized network overlay is that traffic might be passing through the sub optimal path as all inter network routes must be passed through spine. Distributed mode can solve this problem as leaf is serving as L3 gateway too.

- For V3R3 AC DCN solutions, both hardware overlay in centralized mode and distributed mode are supported. Underlay physical design is same for both centralized and distributed gateway deployment; However, the overlay configuration might be different.

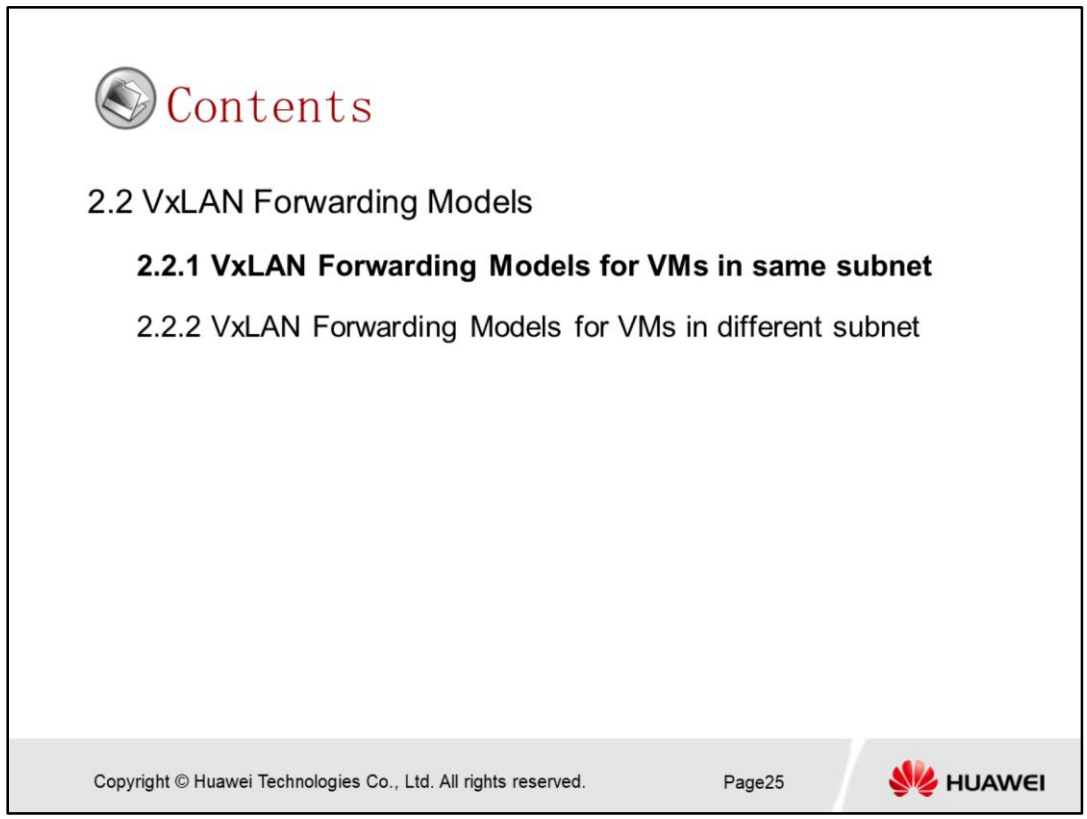
- Centralized hardware overlay will be used for the following discussion in the following

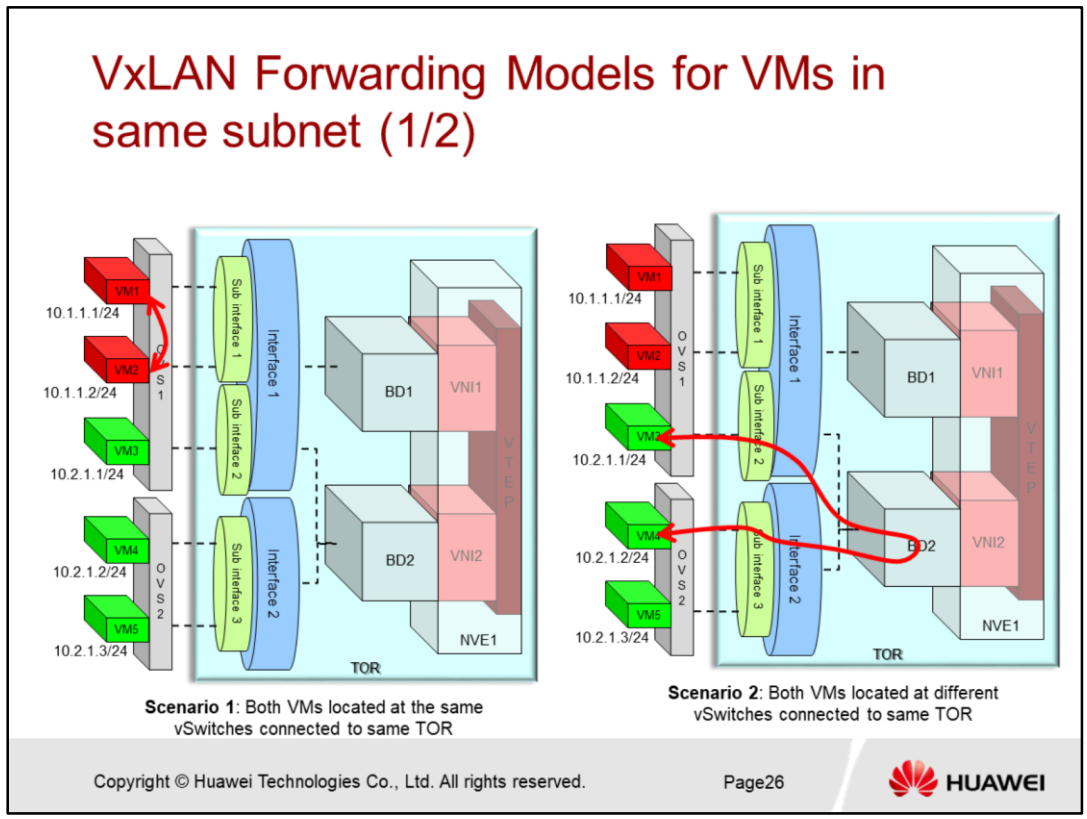
slides.



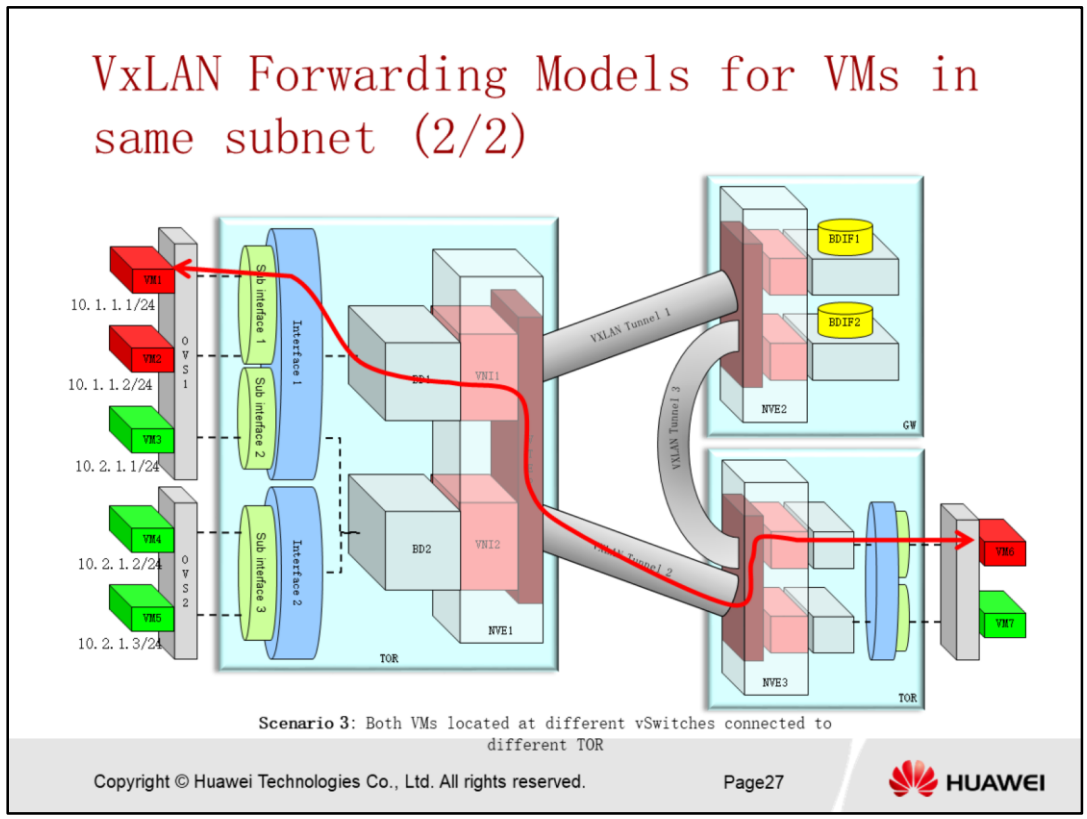


- The VxLAN forwarding model here is only discussing on the centralized gateway mode; the distributed gateway mode is not discussed in this slide here. Diagram above shows the basic configuration to be done on VxLAN configuration to prepare for VxLAN forwarding model.



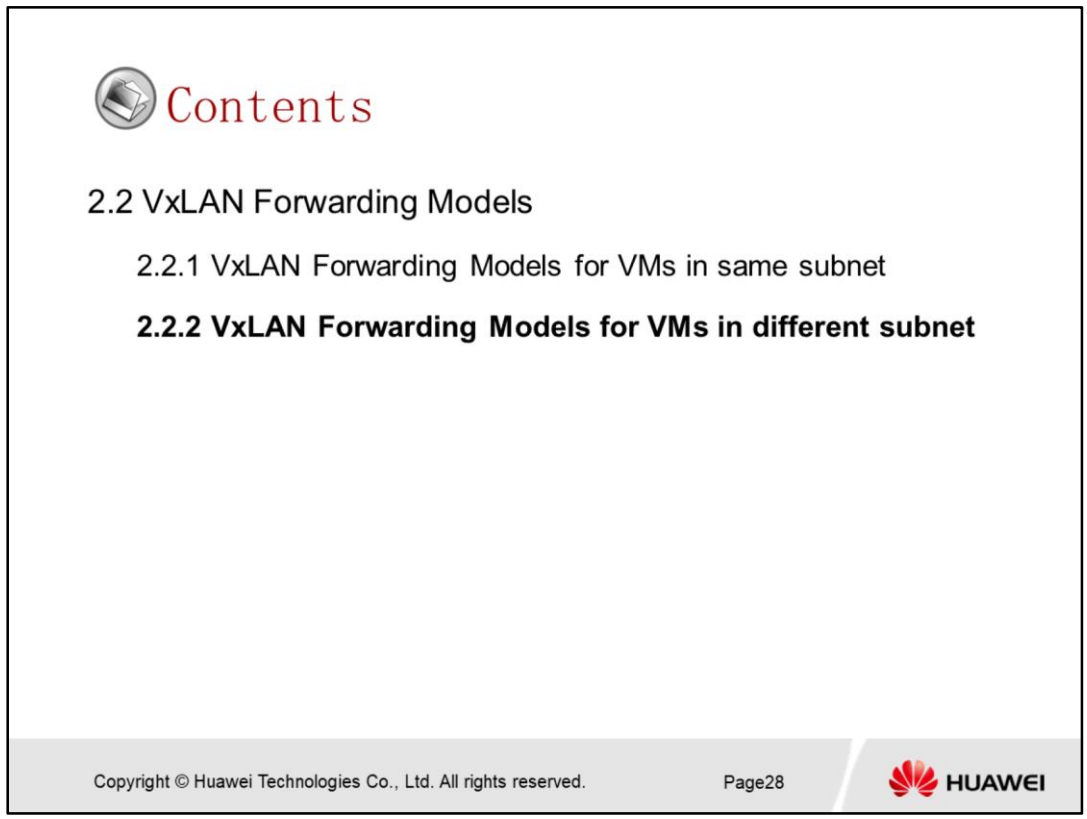


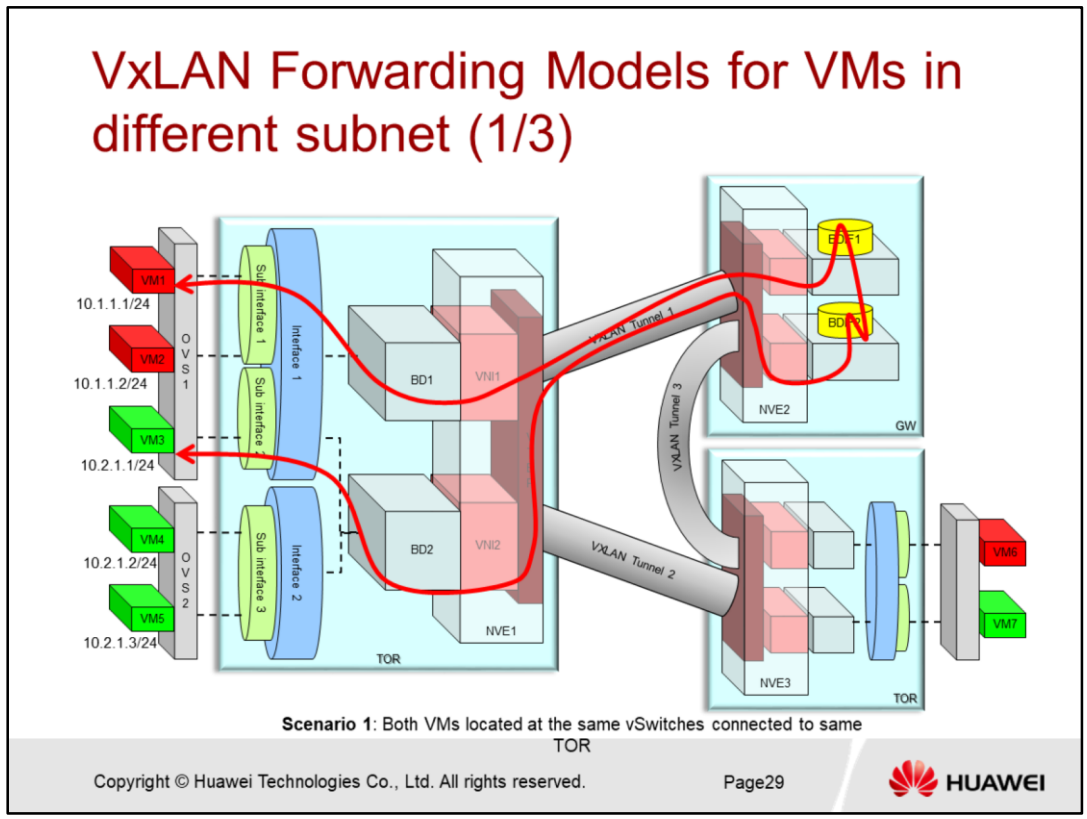
- Communication for VMs in the same subnet can be further divided into a 3 scenarios, depending on the locations of VMs and connections to TOR, as per listed below:-
    1. **Both VMs are located in the same vSwitches connected to the same TOR.**
        - In this case, the communication of the 2 VMs in the same network segment is just through L2 in OVS
    2. **Both VMs are located in different vSwitches but connected to the same TOR.**
        - TOR, serving as the NVE binds VM in the same network segments in the same bridge domain mapping to the same VNI (VNI is mapped to the user access VLAN as well). Thus, inter vSwitches communication can be achieved on TOR in the same bridge domain as they are mapping to the same VLAN.



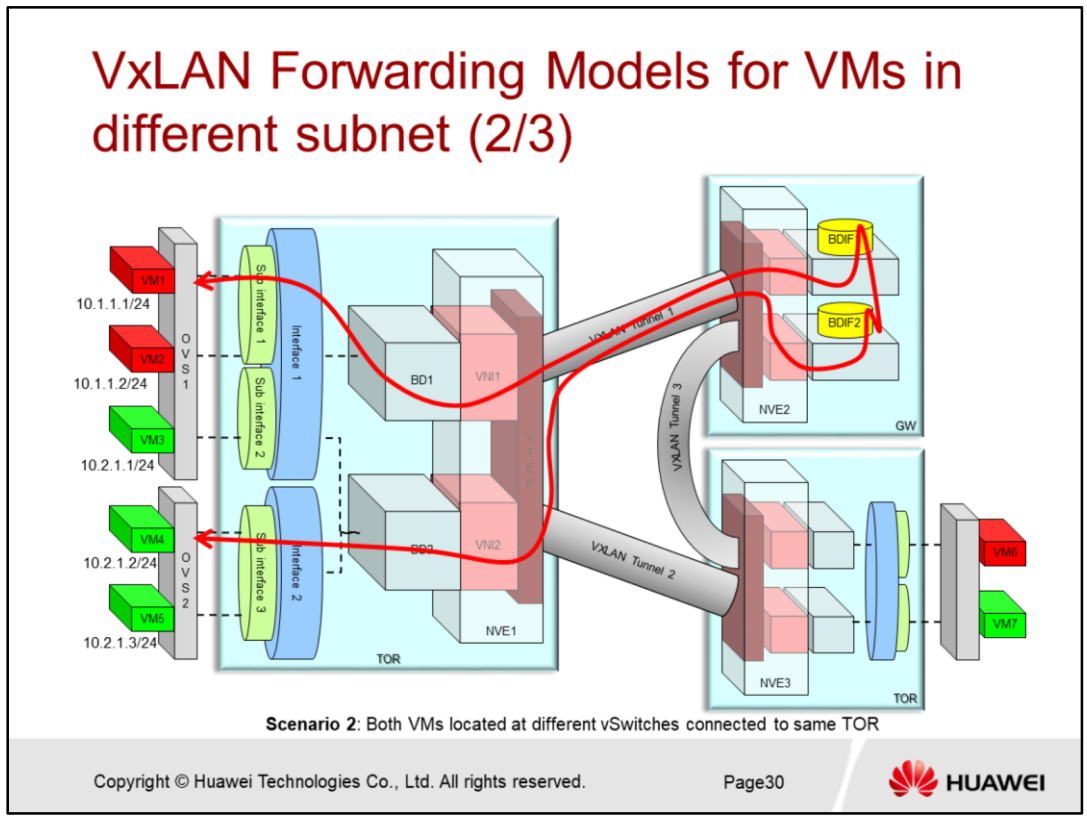
3. Both VMs re located at different vswitches connected to different TOR.

- As this is intra-segment communication, the communication does not need to pass through gateway but can be achieved by using the VxLAN tunnel built between 2 VxLAN L2 gateway, which is on both TOR serving as VTEP.



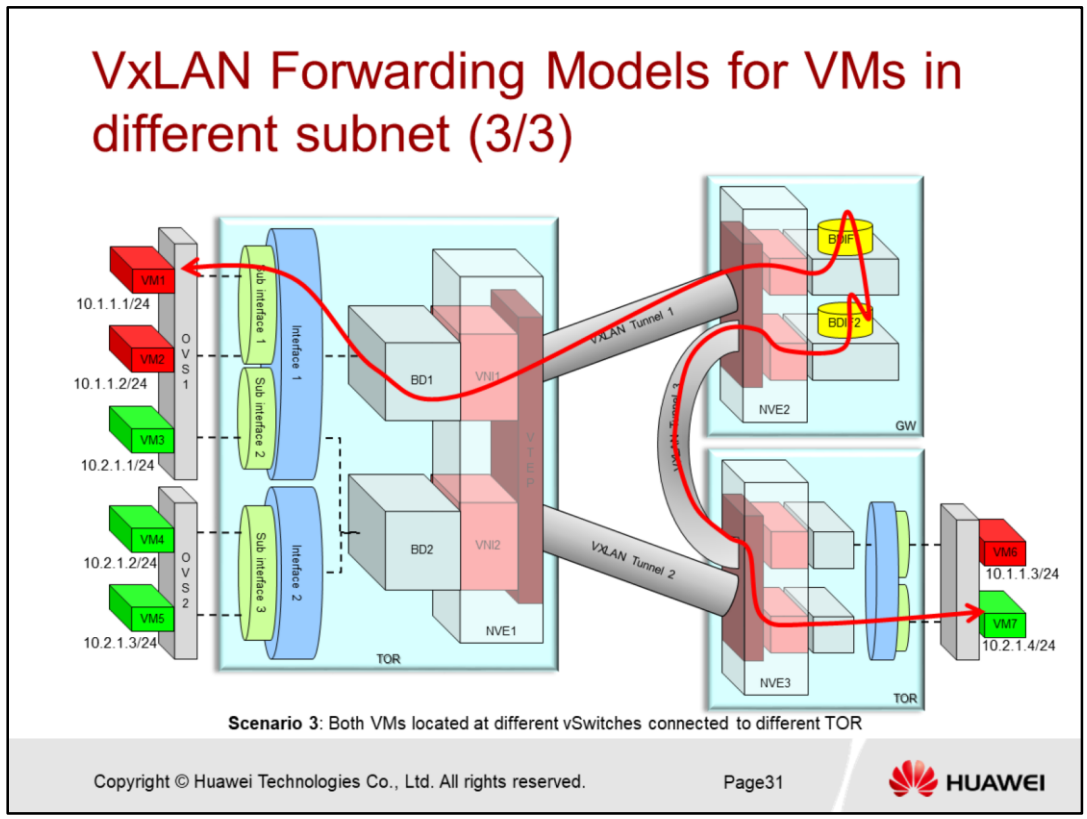


- Communication for VMs in the differnet subnet can be further divided into a 3 scenarios, depending on the locations of VMs and connections to TOR, as per listed below:-
    1. **Both VMs are located in the same vSwitches connected to the same TOR.**
        - As both VMs is in different network segment, VM1 will forwards it data to the L3 gateway, L3 gateway will route it to the VM3 through another Bdinterface to reach another network.



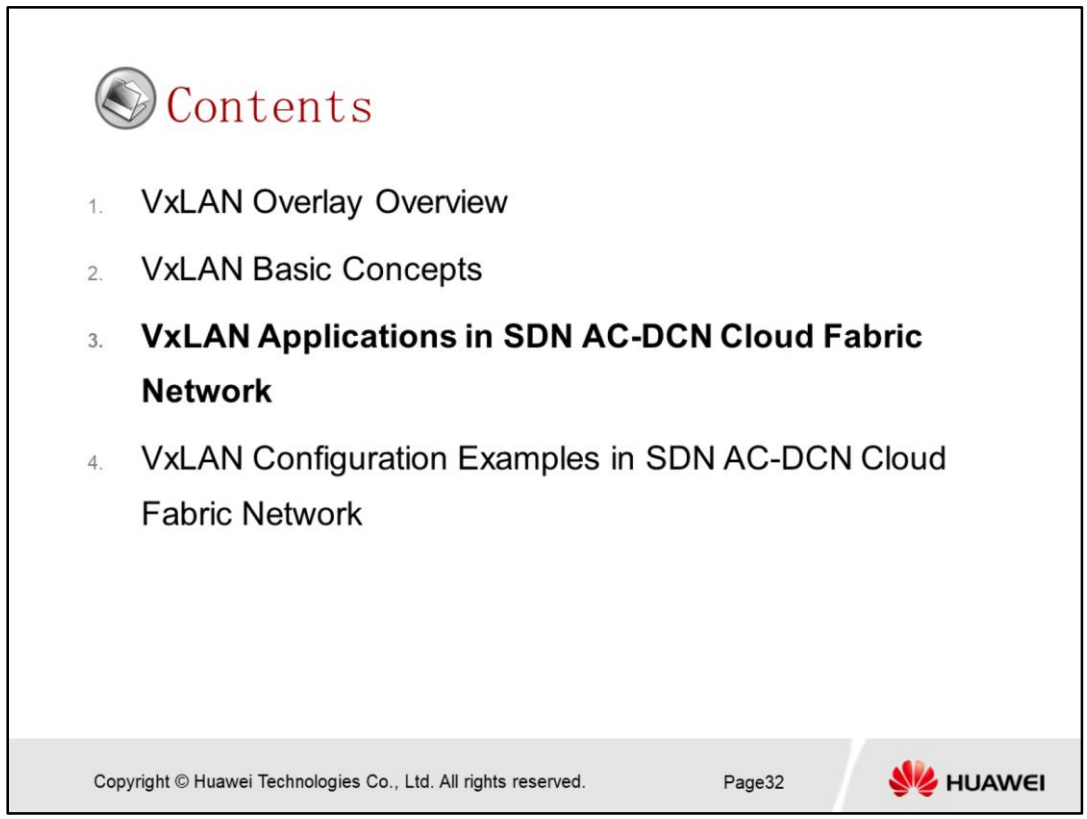
2.. Both VMs are located in different vSwitches but connected to the same TOR.

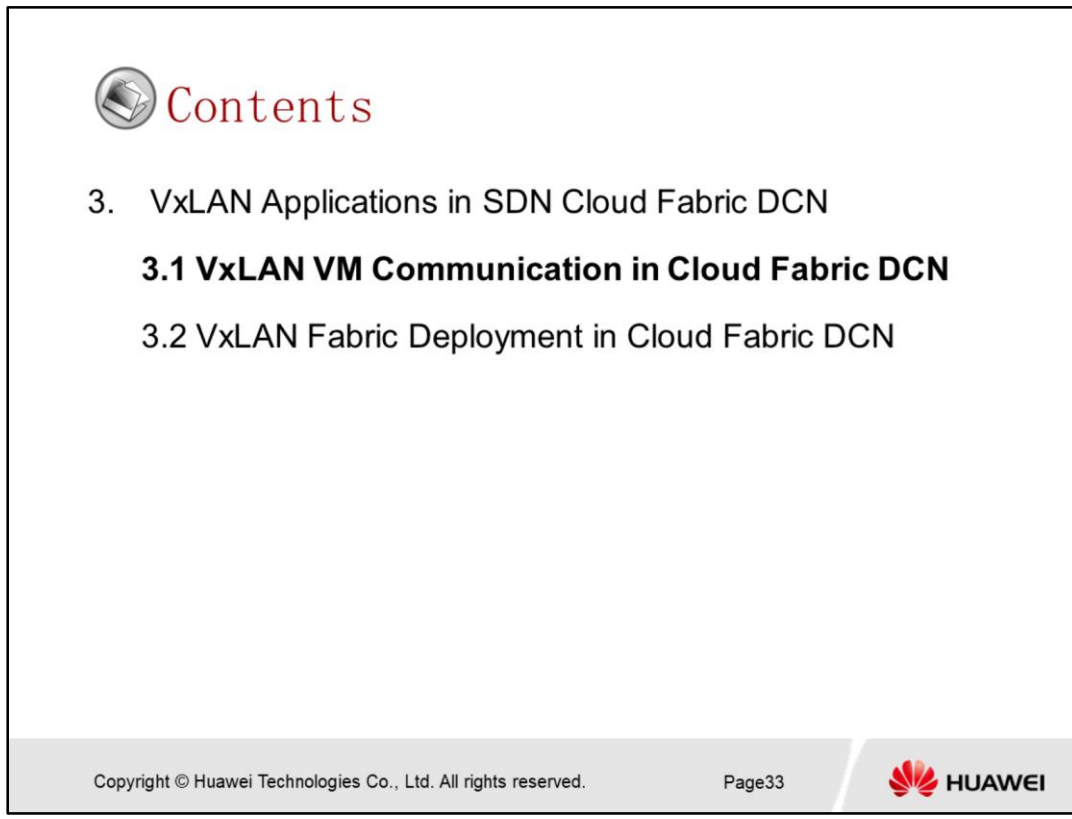
- As both VMs is in different network segment, VM1 will forwards it data to the L3 gateway, L3 gateway will route it to the VM3 through another Bdinterface to reach another network.

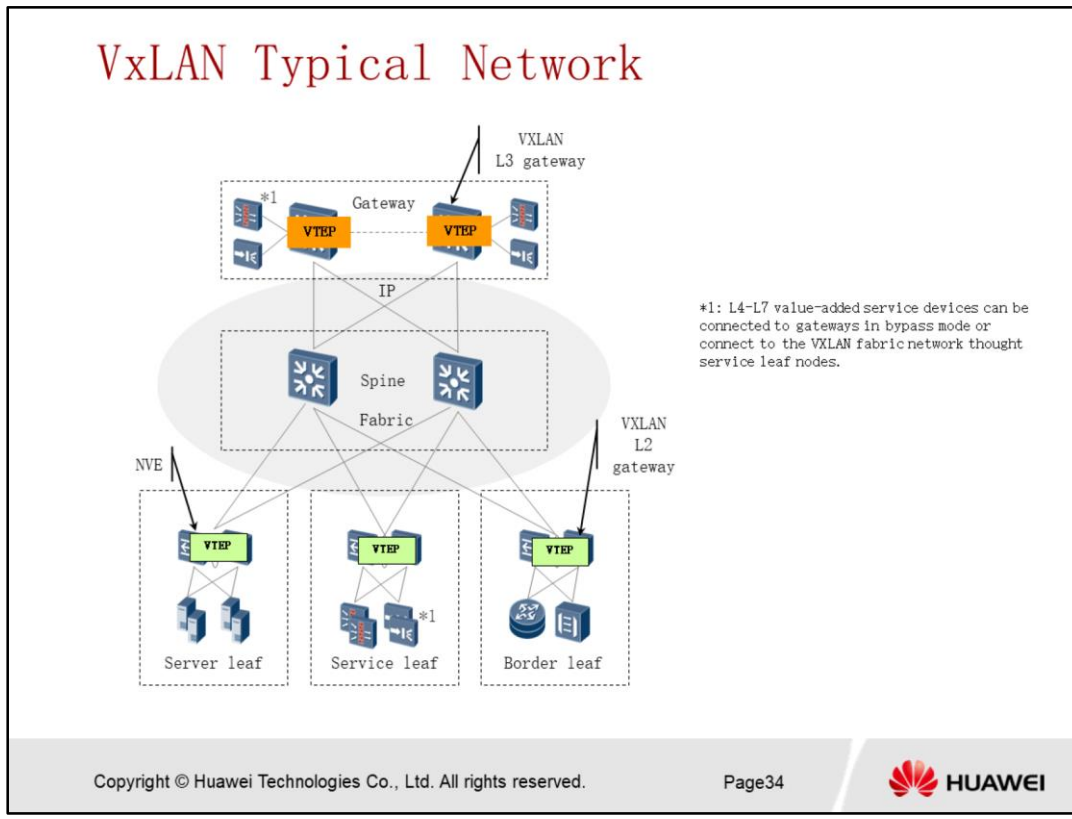


3. Both VMs are located at different vswitches connected to different TOR.

- As both VMs is in different network segment on different TOR, VM1 will forwards it data to the L3 gateway through VxLAN tunnel, L3 gateway will route it to the VM3 through another Bdinterface to reach another network located on another switches through another VxLAN tunnel.



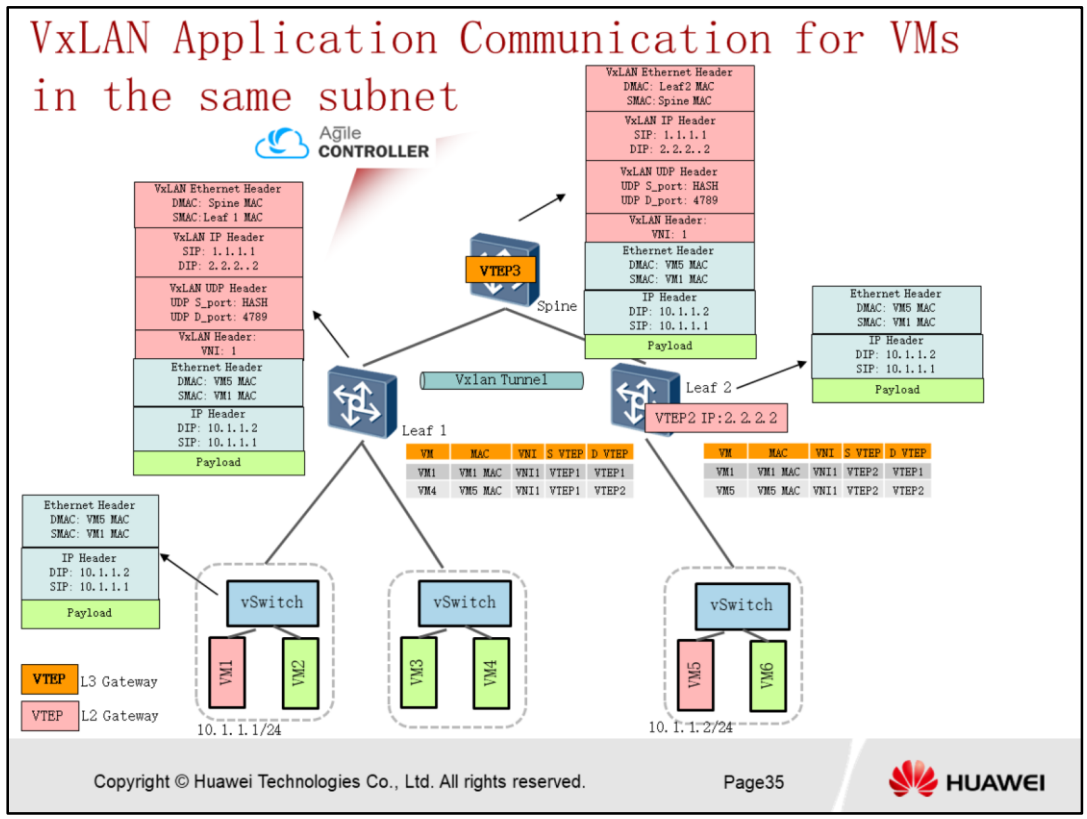




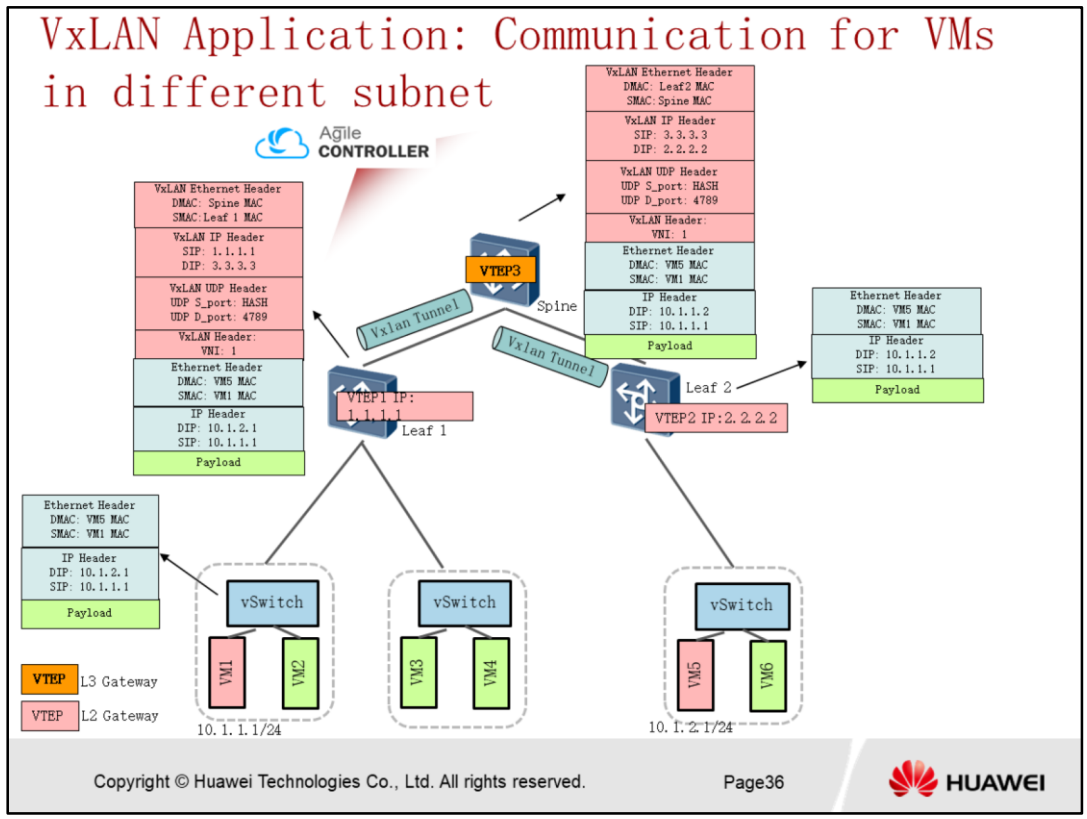
- The explanation and general terms of VxLAN network in DCN is listed below:-

1. **Fabric:** A basic physical network for a data center, which is composed of a group spine and leaf nodes.
2. **Spine:** Core node of a VXLAN fabric network, which uses high-speed interfaces to connect to functional leaf nodes and provides high-speed IP forwarding.
3. **Leaf:** An access node that is deployed on a VXLAN fabric network to connect various network devices to the VXLAN network.
4. **Service leaf:** A functional leaf node that connects L4-L7 value-added service devices, such as firewall and LB, to the VXLAN fabric network.
5. **Server leaf:** A functional leaf node that connects computing resources (virtual or physical servers) to the VXLAN network.
6. **Border leaf:** A functional leaf node that connects to a router or transmission device and forwards traffic sent from external networks to the data center.
7. **NVE:** Network virtualization edge, a network entity that implements network virtualization. NVE nodes establish an overlay virtual network on the underlay Layer 3 basic network.
8. **VTEP:** VXLAN tunnel endpoints that are deployed on NVE nodes and responsible for VXLAN packet encapsulation and decapsulation. VTEPs are connected to the physical network and assigned IP addresses (VTEP IP) of the physical network. VTEP IP addresses are independent of the virtual network. A local VTEP IP address and a remote VTEP IP address identify a VXLAN tunnel.

9. **VNI:** VXLAN network identifier that identifies a VXLAN segment. Traffic sent from one VXLAN segment to another must be forwarded by a VXLAN L3 gateway.

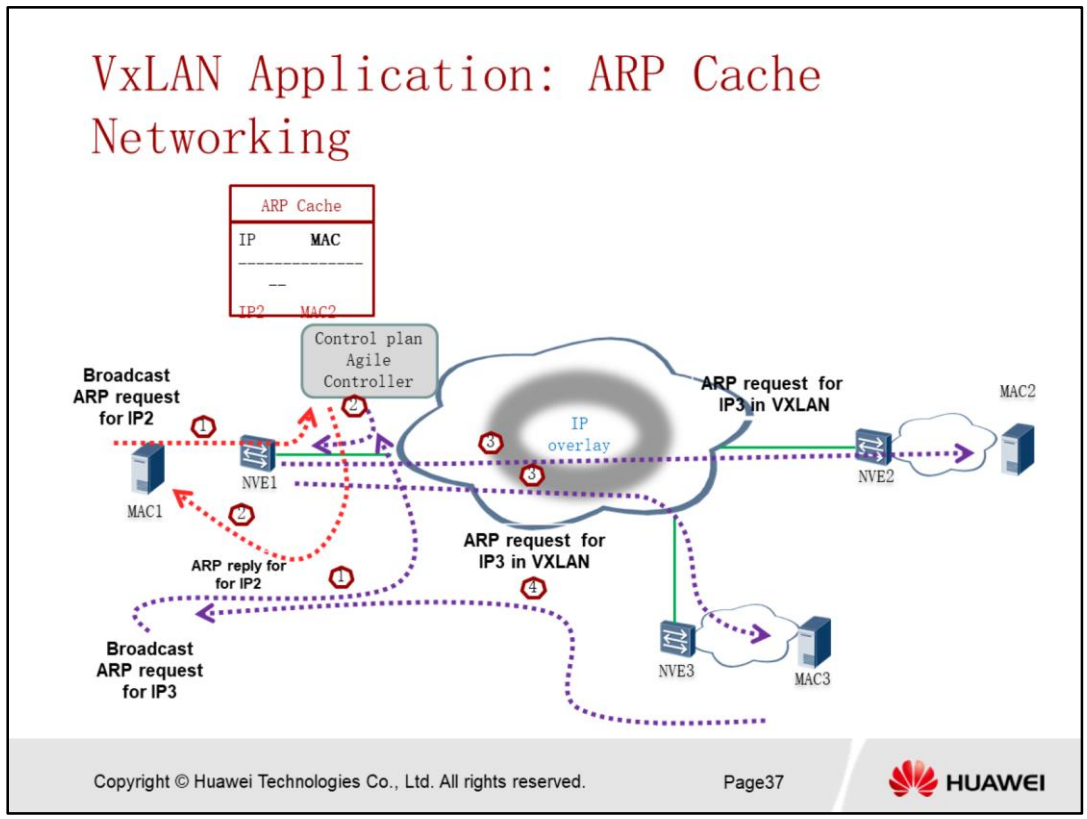


- As shown in the diagram above, Leaf switches serves as VxLAN L2 gateway and VxLAN tunnel is established between both Leaf1 and Leaf2 switches. Openflow channel is established between AC and forwarders through openflow protocol. VxLAN configuration is performed by administrator on AC using Netconf procotol, and the VxLAN configuration is deployed to forwarders through VxLAN channel configured.
- Through the VxLAN tunnel established, same segment VM communication will be performed through VTEP by using the MAC address mapping table.



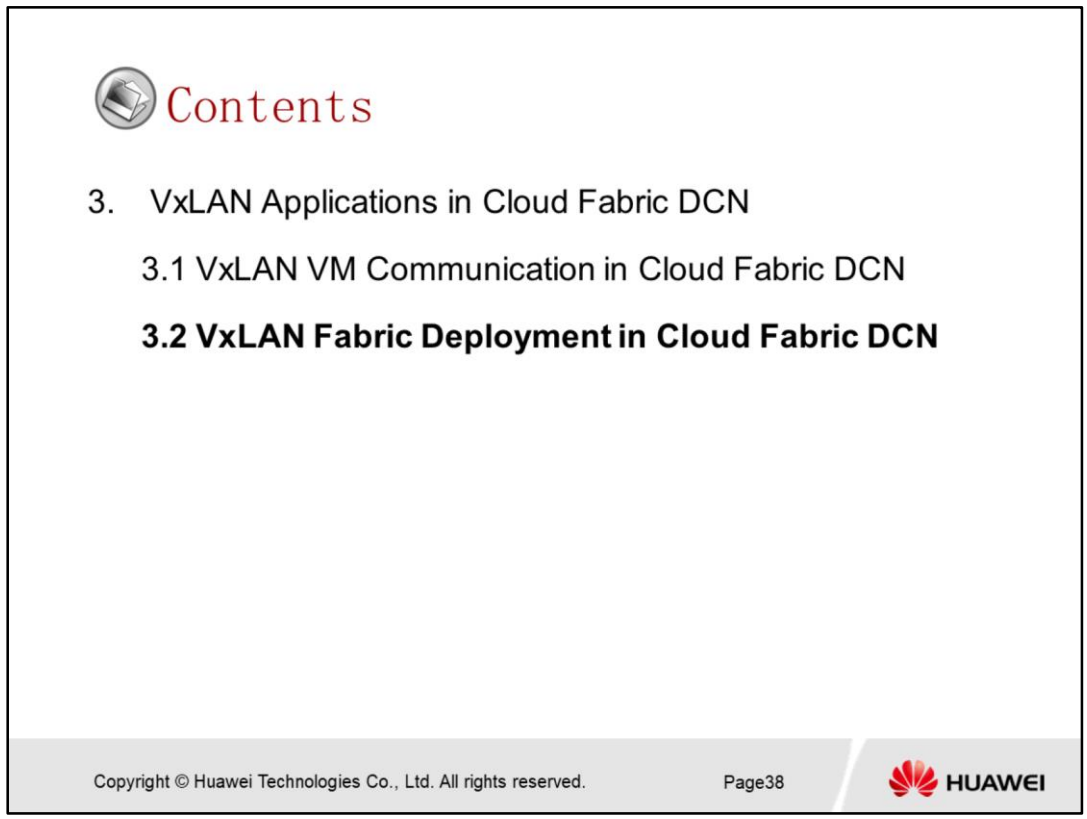
- As shown in the diagram above, Spine switch serves as VxLAN L3 gateway while leaf1 and leaf2 switches serves as VxLAN L2 gateway; VxLAN tunnel is built between switches and realizes inter-segment VM communication (VM1 to VM5) through L3 gateway. Openflow channel is established between AC and forwarder using Openflow protocol.
- Through Netconf protocol, administrator performs VxLAN configuration on AC and AC deploys VxLAN information to forwarder through Openflow channel.
- Logical BDIF interface configuration is performed on L3 gateway and ARP cache function is enabled on AC. Inter segment VM communication can be achieved by VxLAN L3 gateway and ARP cache proxy function.
- For example based on diagram above, VM1 is to communicate with VM5 in different network segment connected to different leaf switches. All VxLAN configuration has been completed by admin and configuration has been deployed from AC to switches using Openflow. When the L2 Ethernet frame reaches VTEP1, VTEP1 serves as L2 gateway will perform VxLAN encapsulation and perform forwarding based on VxlAN Ethernet Header. When VTEP3 receives the VxLAN frames, it performs VxLAN de-encapsulation and find out that the DMAC is not on local segment; It serves as VxLAN L3 gateway, removes the ethernet header, check the ARP entry matching entry of BDIF interface, and encapsulates back VxLAN and send it to VTEP2. When VxLAN frame reaches VTEP2, VxLAN header is removed and frame is forwarded based on

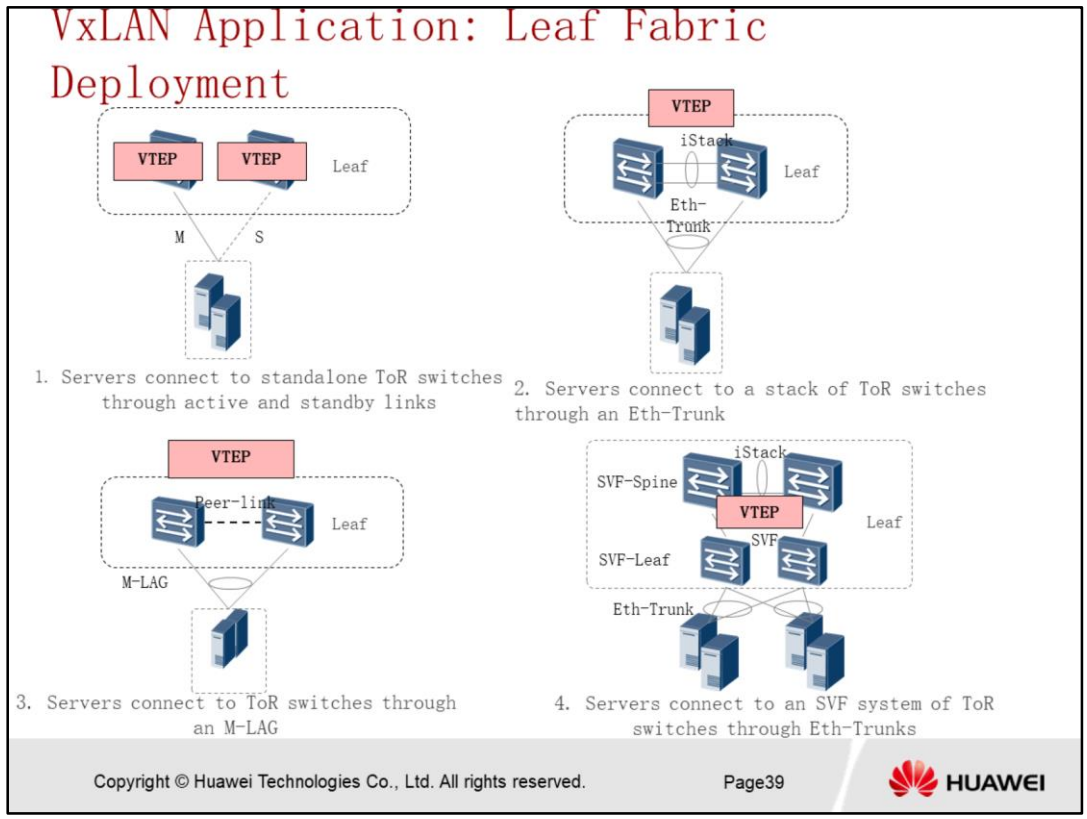
the original ethernet frame.



- AC obtains the VM detailed information and keeps ARP cache in the controller. When MAC1 broadcast an ARP request for IP2, ingress NVE sends this broadcast message will be sent to AC for processing; AC analyzes the ARP request and searches the ARP cache based on mapping of MAC address entry; if The MAC entry is existing in the cache, AC will reply with ARP unicast reply and the broadcast message is terminated.

- Else if the MAC entry it is not existing in ARP cache, AC will send to ARP request back to NVE1, NVE1 broadcasts to everyone to get reply from the real host. As shown in example, when MAC1 sends a broadcast ARP request for IP3 and the corresponding MAC address does not exist in ARP cache of AC, AC will send back to NVE1 and NVE1 broadcast to NVE2 and NVE3 to reach all hosts. MAC3 will reply in this case.





- There are basically 4 different types of VxLAN leaf fabric deployment in Cloud DCN, considering redundancy and protection level. Thus, a server is normally dual-homed to 2 leaf fabric physically. The 4 types of VxLAN leaf fabric deployment is listed below:

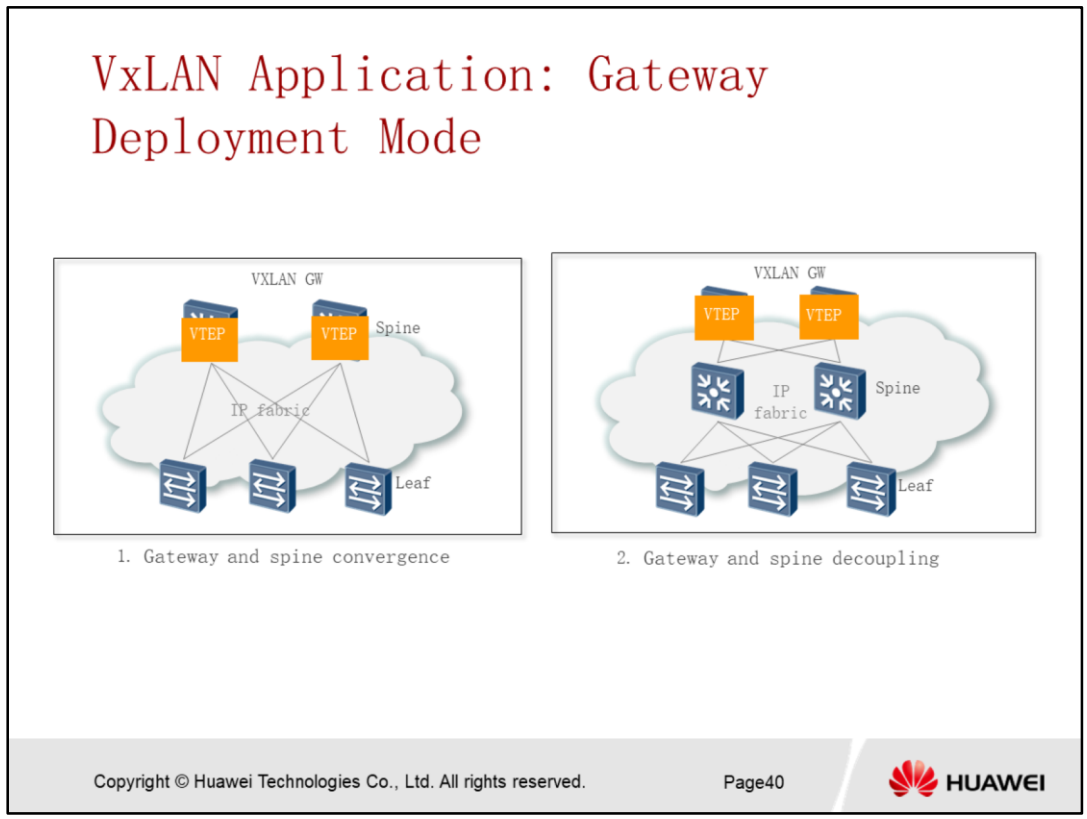
1. **Servers connect to standalone ToR switches through active and standby links**
   - A standalone ToR switch acts as a leaf node. Each server is connected to two ToR switches using active-standby NICs (NIC bonding). Only one NIC in a server sends and receives packets at a time, resulting in a low bandwidth efficiency. The VTEP IP address will change after an active/standby NIC switchover. In this case, the upstream VTEP needs to learn the forwarding entry from the BUM traffic sent from the server.

2. **Servers connect to a stack of ToR switches through an Eth-Trunk**
   - iStack technology virtualizes two ToR switches into one logical switch with a single control plane, which simplifies device management. The NICs of a server work in active-standby/load-balancing mode to improve bandwidth efficiency. However, the upgrade and maintenance operations for the logical device are complex.

3. **Servers connect to ToR switches through an M-LAG.**
   - Two ToR switches are connected using a peer-link and set up a Dynamic Fabric Service (DFS) group. The two switches act as one logical device but have their own independent control planes, simplifying upgrade and maintenance while improving system reliability. Their downlink ports set up an M-LAG for dual-homing of servers. NICs of a server work in active-standby/load-balancing mode. The configuration is complex because each ToR switch has an independent control plane.

4. **Servers connect to an SVF system of ToR switches through Eth-Trunks**

- Two high-performance ToR switches supporting VXLAN use iStack technology to set up a stack. Cost-effective ToR switches with SVF configured are connected to the stack to provide cost-effective 1G/10G VXLAN network access capability. Each server is dual-homed to two SVF leaf nodes, with the NICs working in active-standby/load-balancing mode.

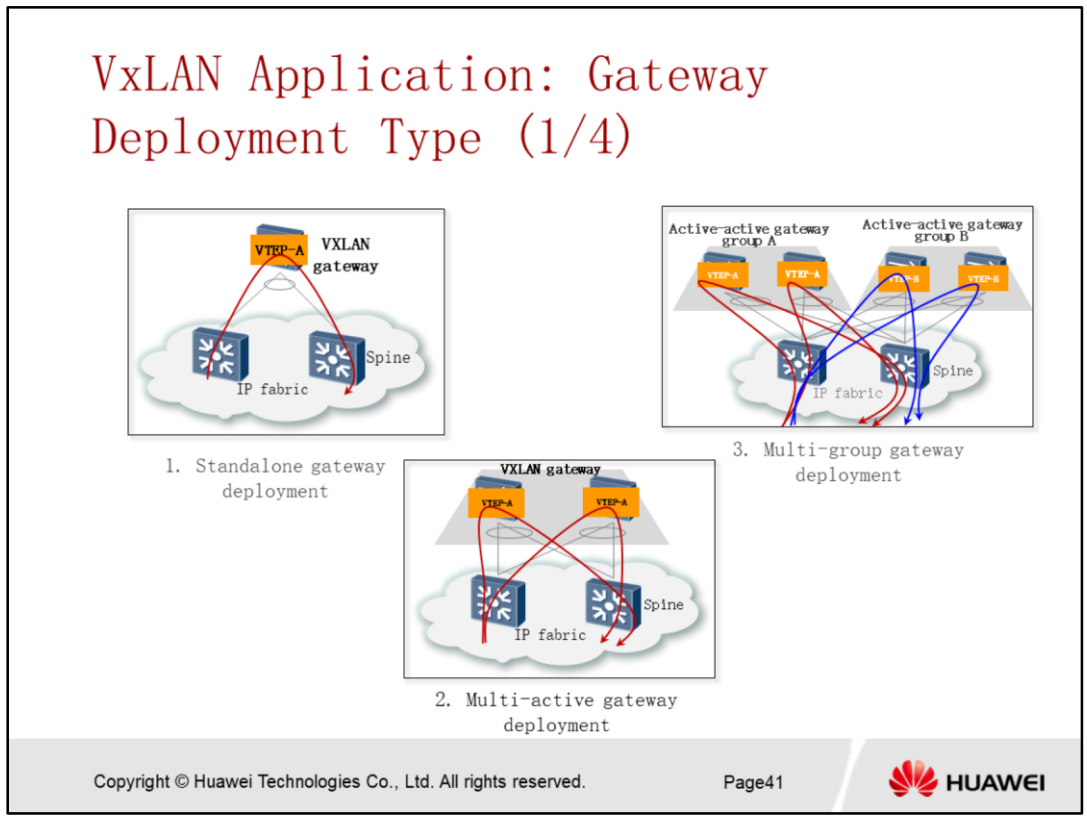


- For the VxLAN deployment in Cloud DCN scenario, the gateway deploymet can be divided into 2 physical deployment type:-

1. **Gateway and spine convergence**
   - The gateway and VxLAN termination point is deployed on the same spine switch; which means the exit gateway connecting to internet function also is realized on this spine switch. This realizes 2 layer architecture in the physical underlay deployment
   - The convergence deployment reduces the number of network devices and lowers the network deployment cost.
   - The gateway nodes are closely coupled with the spine nodes, making network expansion difficult. This deployment is applicable to a data center that does not need to be expanded in the near future.
   - Gateways cannot be deployed in multi-group mode.

2. **Gateway and spine decoupling**
   - Exit gateway and VxLAN gateway is realized on different equipments; this leads to 3 layers architecture in the underlay deployment.
   - The decoupling deployment facilitates network expansion. Expansion of the spine, leaf, or gateway nodes will not greatly affect the other nodes.
   - Multiple groups of gateways can be deployed on a large-sized network.

- Gateways can be deployed in multi-group, multi-active mode.

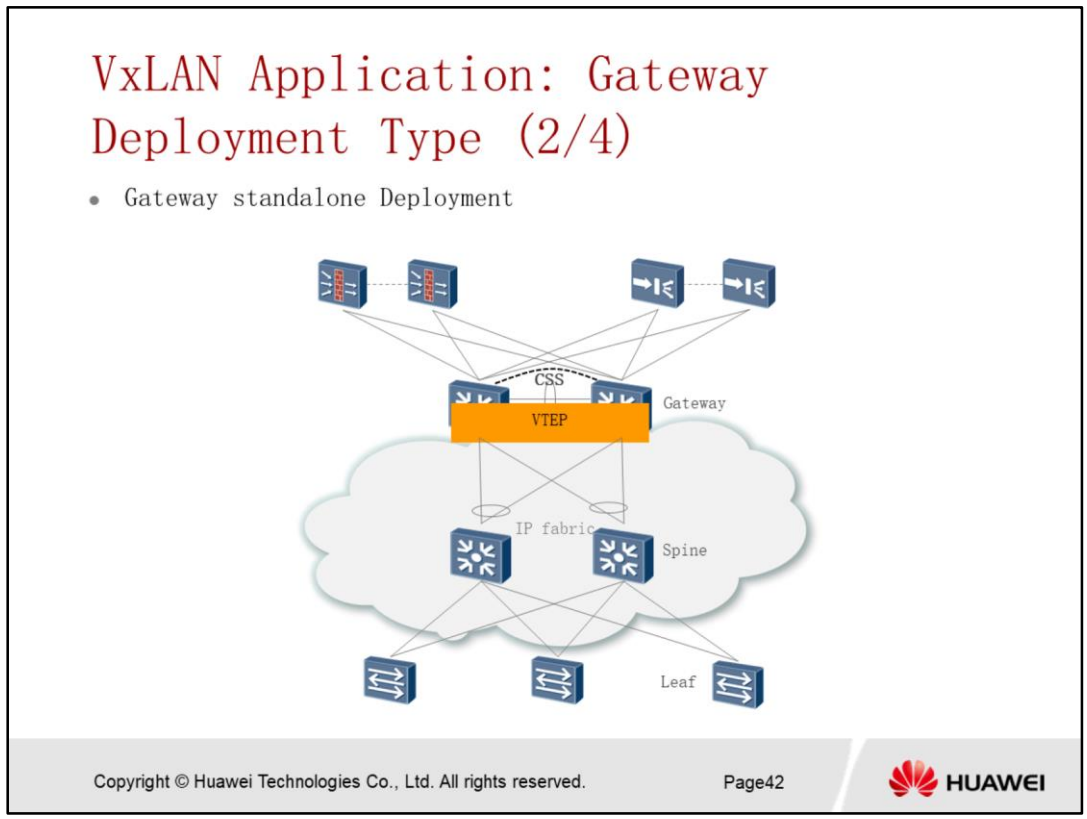


- There are 3 different types of gateway deployments, which are listed below:-

1. **Standalone gateway deployment**
   - A standalone switch or stack system act as the VXLAN gateway. It supports 4K tenants (2K tenants in VPN access scenario), 4K VRFs, 4K subnets, 4K VNIs, 125K VMs (or 25K VMs + physical servers), 32K-1M RIB entries, and 5K ACLs.

2. **Multi-active gateway deployment**
   - Two or four gateway devices set up a DFS group and are configured with the same gateway address and VTEP IP address to act as one logical gateway for VMs.

3. **Multi-group gateway deployment**
   - More subnets can be supported by deploying more gateway groups, but the forwarding capability and reliability of each gateway group remain unchanged.



Requirements

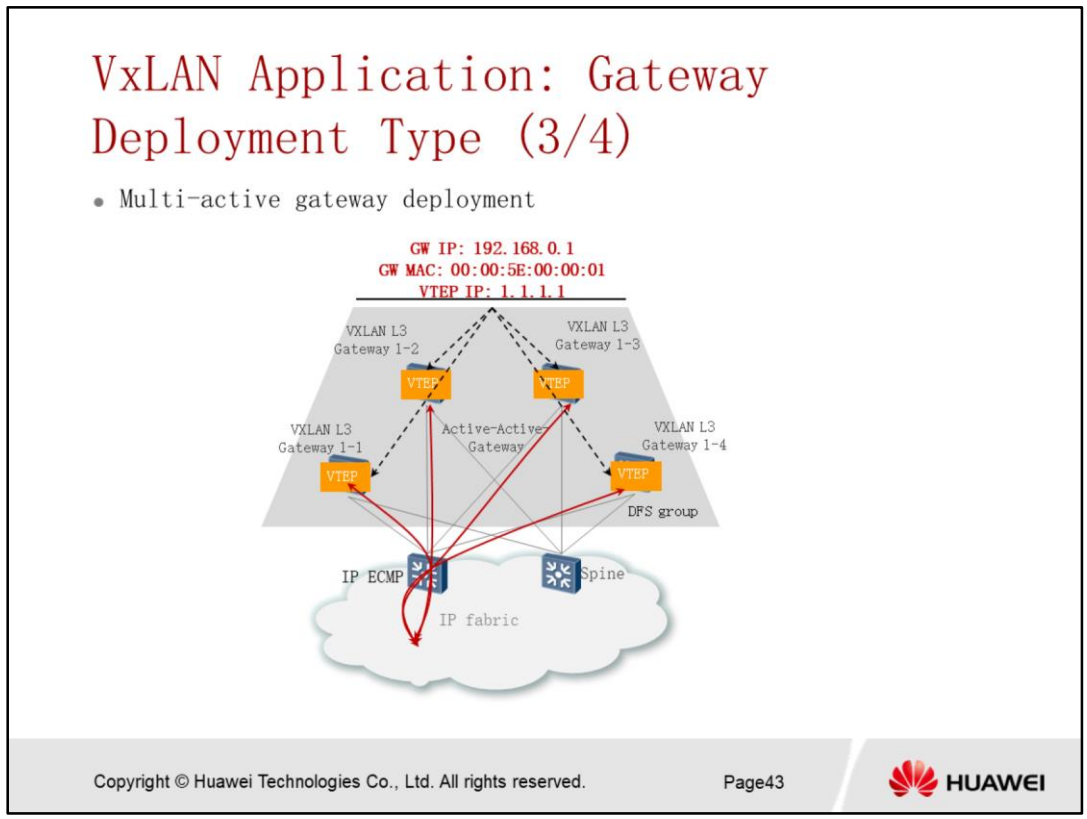
- A CSS system has been deployed in a DC.
- The VXLAN gateway needs to be deployed on the CSS system.

Solution design

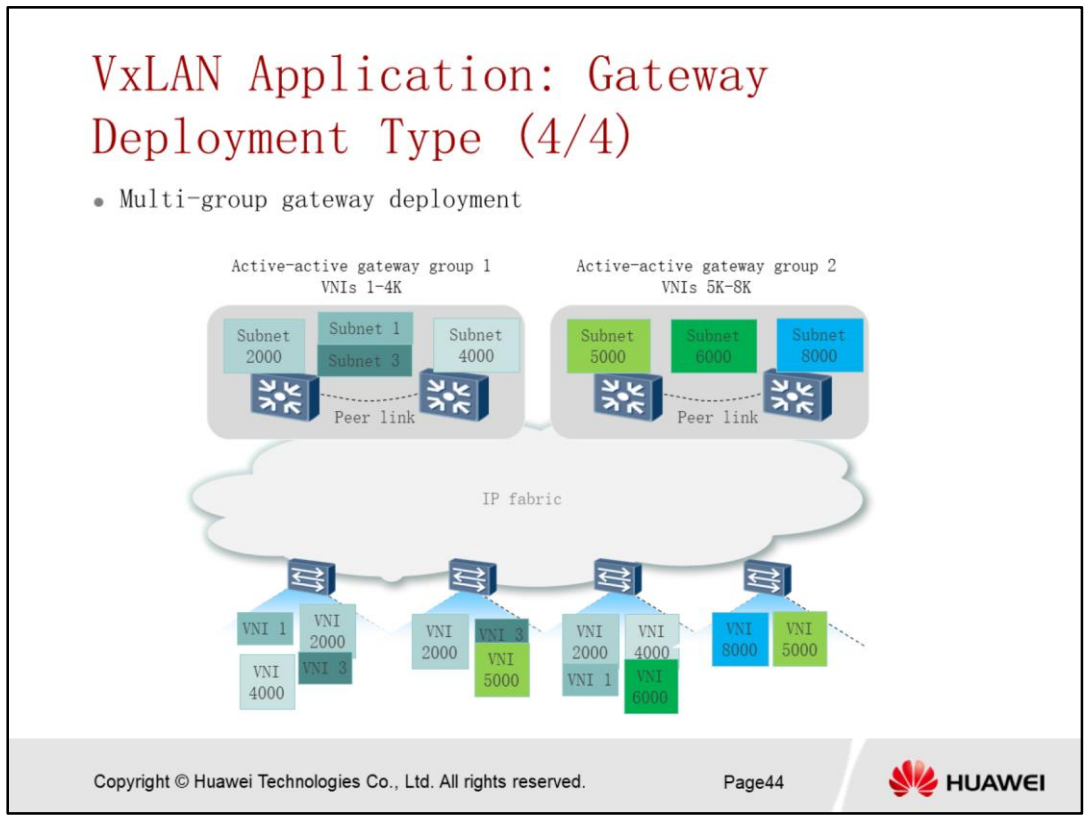
- The CSS system acting as the gateway has a vBDIF IP address, virtual MAC address, and VTEP IP address configured.
- Value-added service devices (FW/LB) are dual-homed to the CSS system in bypass mode.
- Value-added service devices are expanded together with the gateway devices.

Characteristics

- This solution can be used for VLAN-to-VXLAN evolution.
- The CSS system is managed and configured as an independent logical device, which simplifies device management and facilitates network O&M.
- The CSS gateway is more reliable than a standalone gateway device.



- Multiple gateway devices are configured with the **same gateway address and VTEP IP address.** VMs are unaware of locations and number of gateway devices.
- Multi-active gateway devices **set up a DFS group and use the peer link between them to synchronize ARP and MAC entries,** so that they save the same traffic forwarding information.
- Each gateway device in a multi-active gateway group has all forwarding information and can work independently, providing high gateway reliability.
- The underlay network uses IP ECMP to implement load balancing, which enables traffic to be evenly distributed to gateways and improves forwarding performance.
- FYI: In multi-active gateway deployment, ping to a virtual or physical server from the gateway may fail because of inconsistent forward and reverse paths. This is a normal situation.
- This solution cannot increase the number of VRF/Subnet/RIB/FIB/ARP/MAC entries supported on gateway devices.
- This deployment is applicable to private cloud data centers requiring high reliability.



Solution description

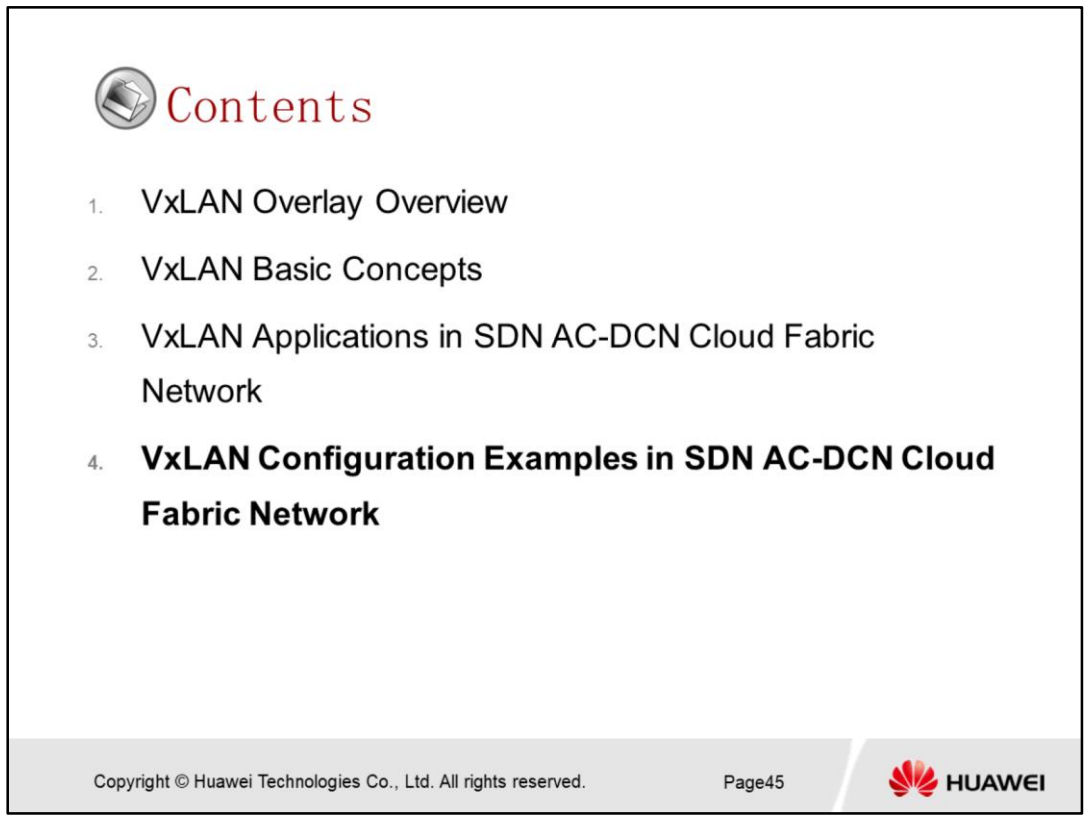
- Deploy new gateway groups to increase the number of resources by multiple times. The new gateway groups run independently and do not affect the original gateway group. VNIs on the same ToR switch can belong to different gateway groups.
- The network scale in a POD expands by multiple times, but the capacity of a single gateway group remains unchanged.

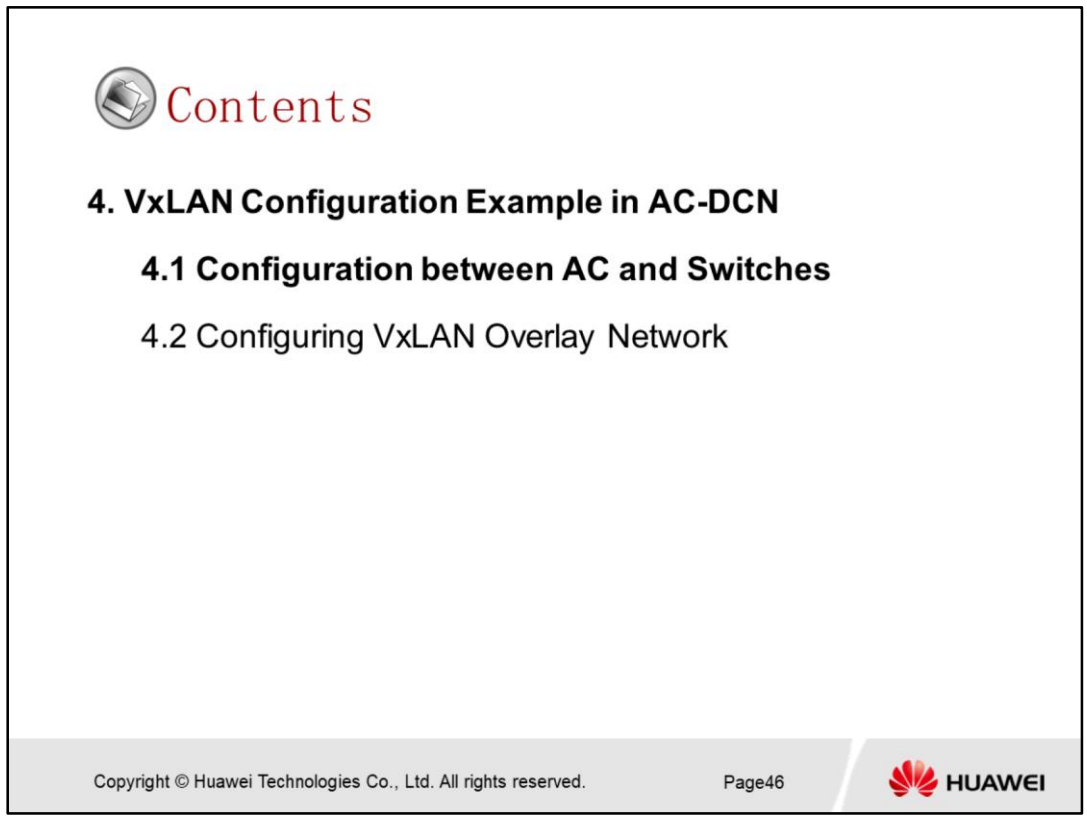
Solution design

- Deploy multiple gateway groups in PODs with a large number subnets.
- A POD supports a maximum of four gateway groups. Each gateway group supports 4K routing domains, 4K subnets, and 25K ARP entries. (This is the specification in a scenario with both virtual and physical servers. 125K ARP entries are supported if there are only VMs.)
- **A gateway group can contain a single gateway, active-active gateways, or quad-active gateways.**
- A tenant can select a gateway group when creating the first VRF. Subsequent VRFs created by the tenant are automatically assigned to the selected gateway group. (Subnets of a tenant must be deployed on the same gateway group.)
- The Agile Controller can assign gateway groups to tenants based on loads of gateway groups.
- Traffic sent from a spine node to a multi-active gateway group is load balanced among IP ECMP paths.

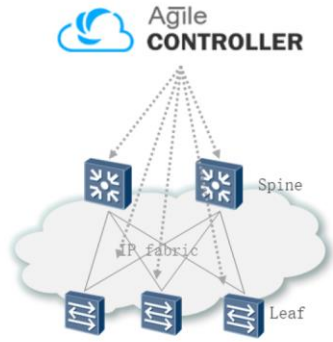
Characteristics

- This solution is applicable to large-scale private cloud DCs.





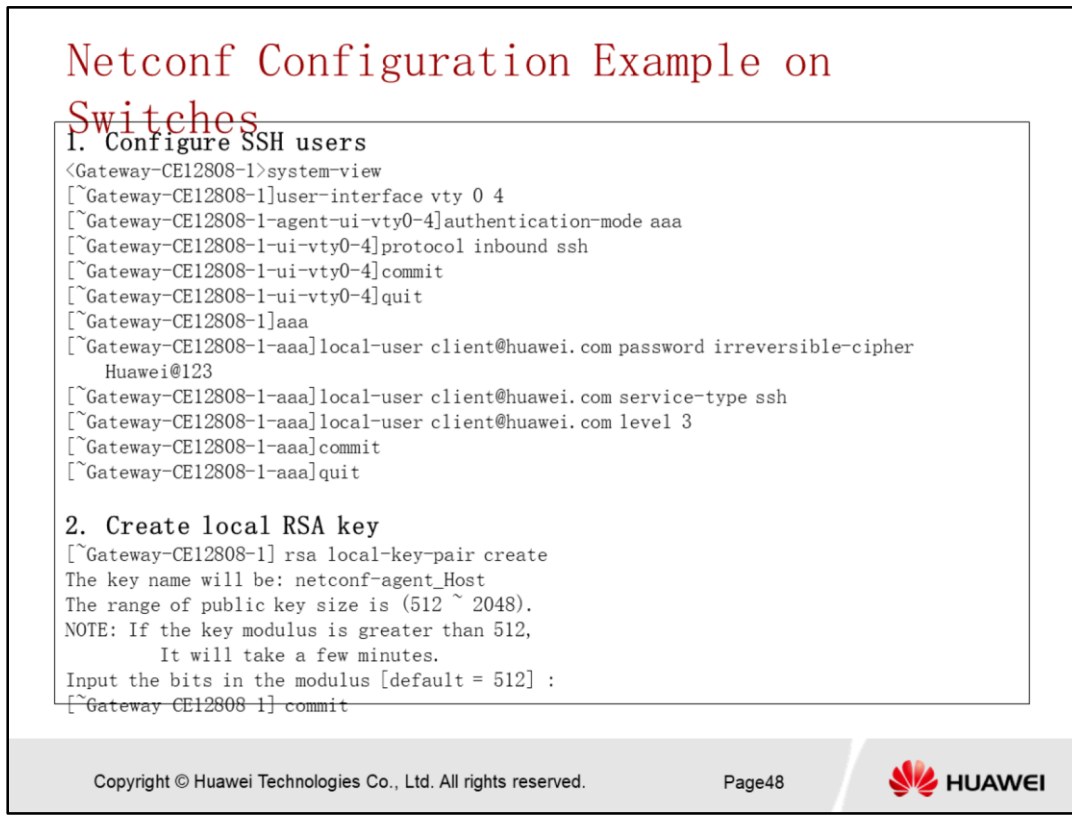
Configuration between AC and Switches



Protocol	Function
SNMP	It is used for AC to be able to add forwarders into topology and manage device status and alarms remotely.
Netconf	For AC to deploy and obtain forwarders' configurations
Openflow	Through Openflow protocol, AC can send and receive VxLAN information and ARP mapping table. After SNMP and Netconf connection is established between AC and forwarders, Openflow configuration on forwarders can be performed through Netconf; no manual configuration on forwarder is needed.

HUAWEI

- There are 3 types of protocol configured between AC and forwarders; all protocols are configured serving for different function.
- SNMP and Netconf configuration must be configured manually on switches while Openflow configuration can be deployed through AC to switches after SNMP and Netconf connection is established.



- Steps above shows an example of the manual configuration needed to be done on forwarders.

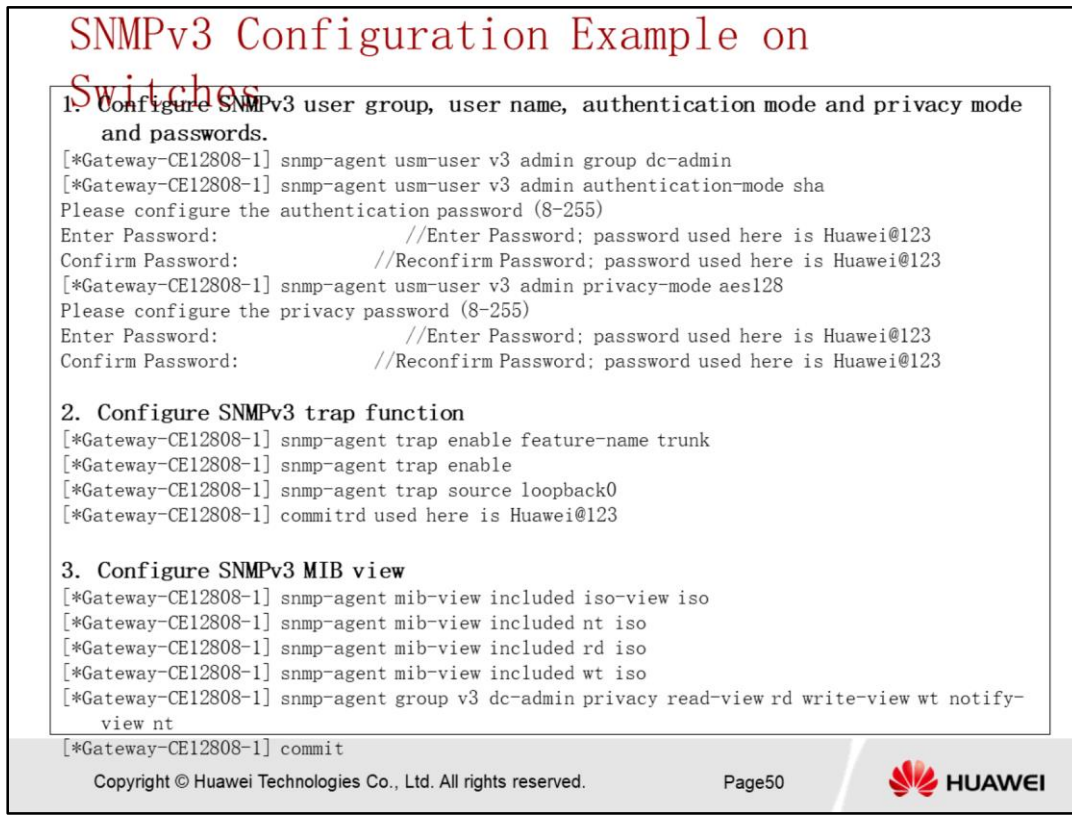
Netconf Configuration Example on Switches

3. Configure SSH user authentication type and service type

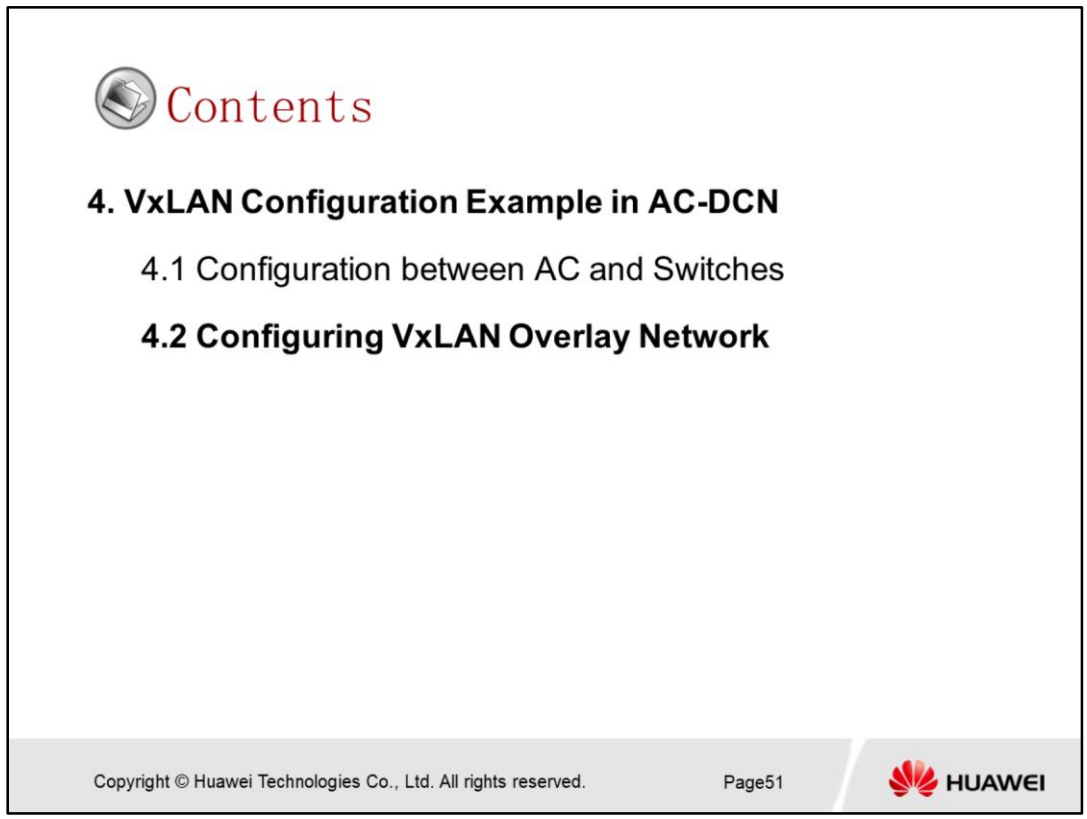
```
[~Gateway-CE12808-1] ssh user client@huawei.com authentication-type password
[~Gateway-CE12808-1] commit
[~Gateway-CE12808-1] ssh user client@huawei.com service-type snetconf
[~Gateway-CE12808-1] commit
```

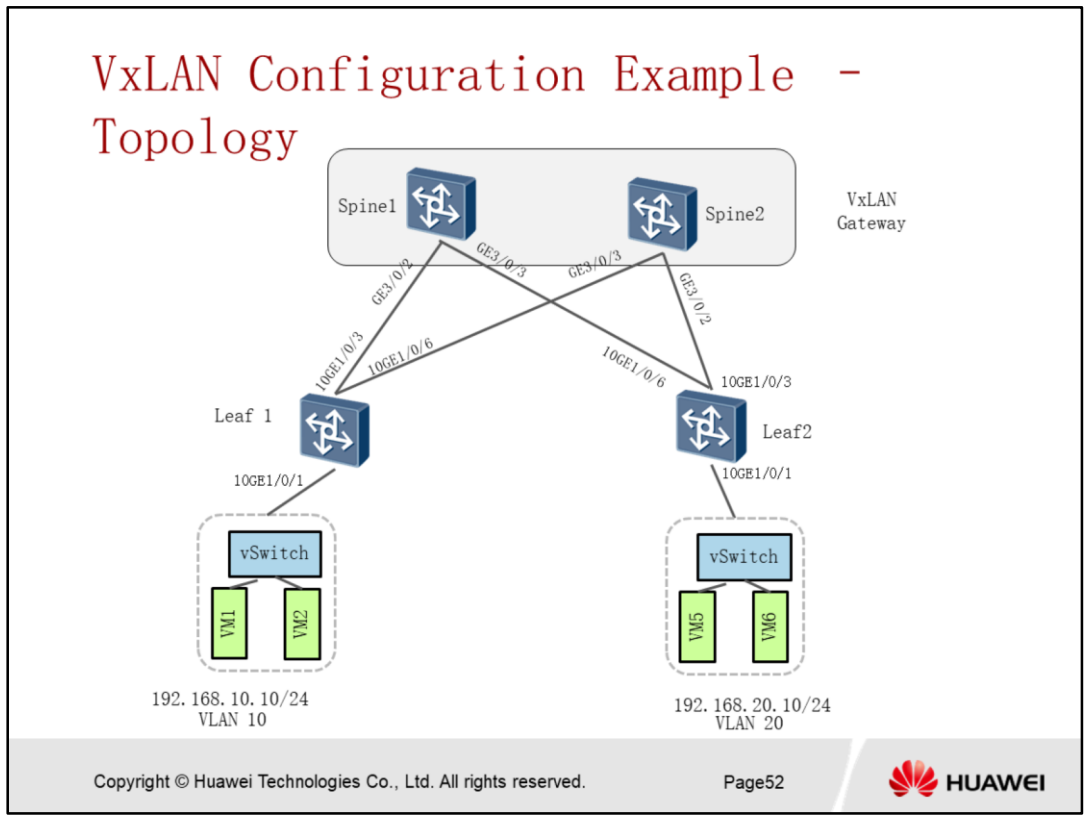
4. Enable Netconf function in global

```
[~Gateway-CE12808-1] snetconf server enable
 [~Gateway-CE12808-1] commit
```



- The AC-DCN obtains LLDP link information from the MIB view specified by SNMP. In this case, the specified MIB view must be iso-view, and the MIB sub-tree of the specified OID must be iso.





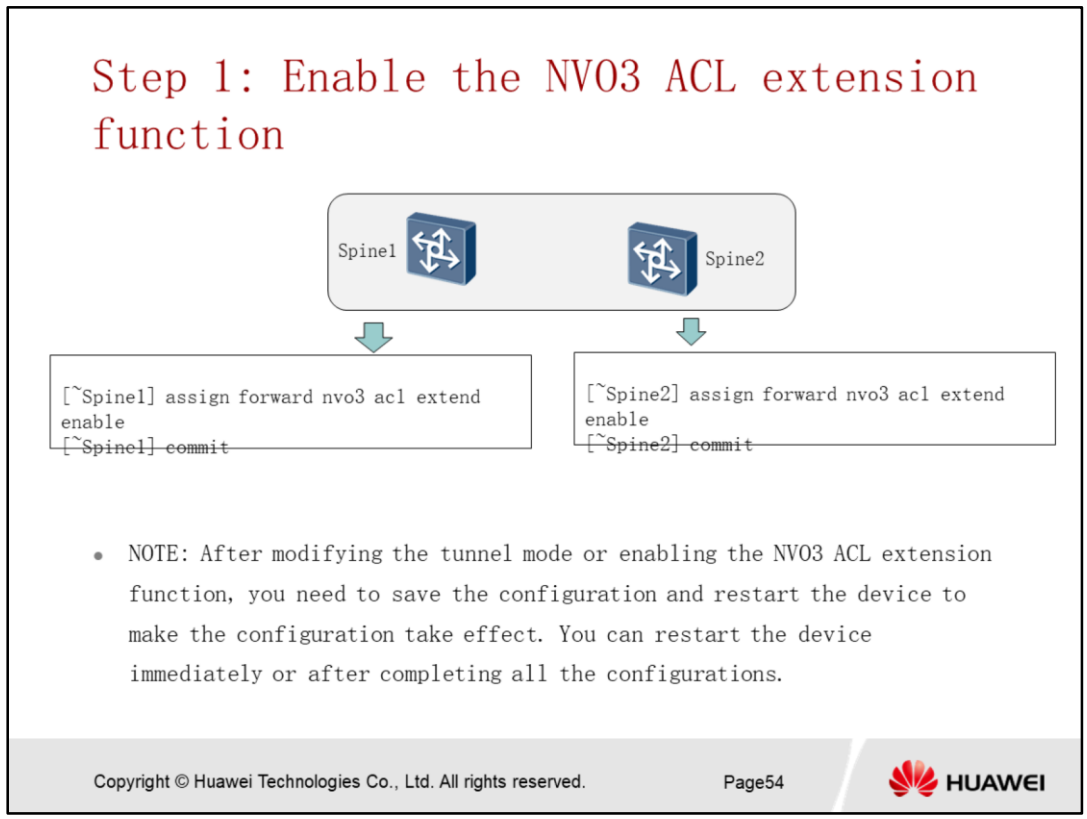
- As shown in the topology above, the DCN shown in the diagram is deploying gateway and spine converged 2 layer DC architecture and centralized multi-active gateway-group. Spine 1 and Spine 2 is located in the core layer while Leaf1, Leaf2, and Leaf 3 are located in the access layer. Full meshed connection is established between spines and leafs, performing ECMP for higher redundancy. No connection between Spines and Leafs.
- VMs are belonged to VLAN 10, 20 and 30 respectively; Bridge domain to be configured are BD10, BD20, and BD30; VxLAN VNI ID are VNI 5000, VNI 5001 and VNI 5002 respectively.

VxLAN Configuration Example - Configuration Roadmap

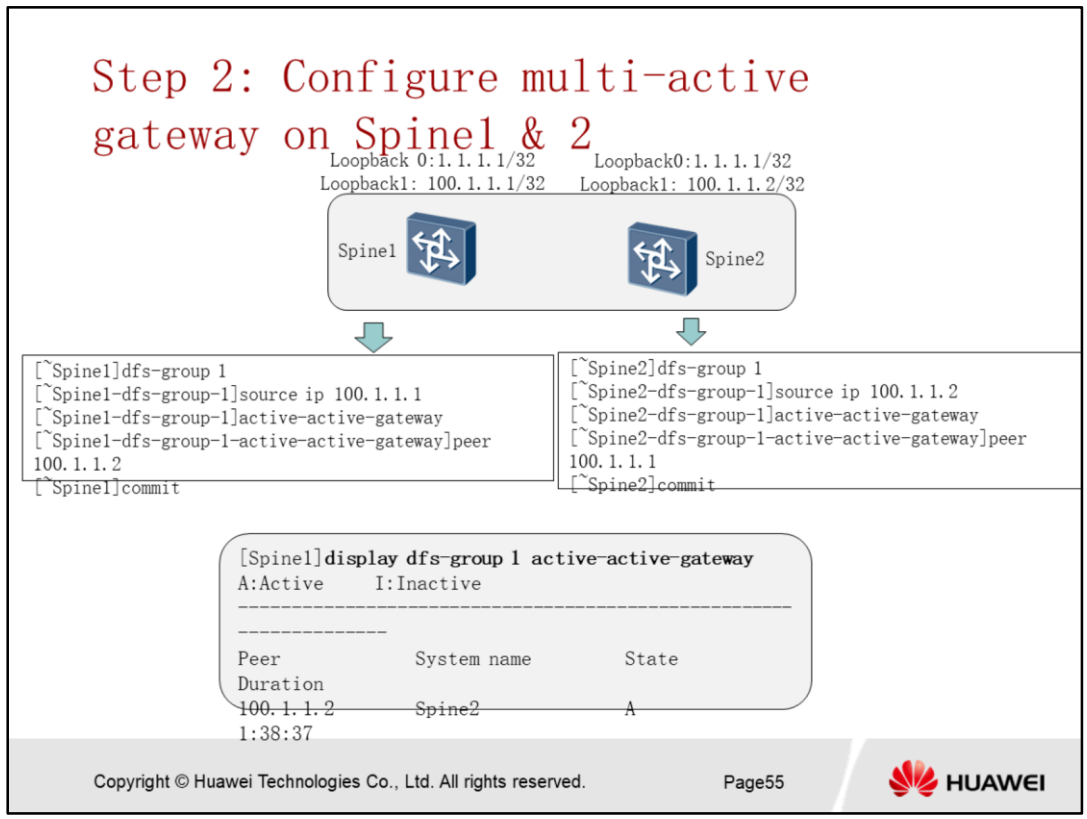
Step	Description
Pre-requisite	Configure OSPF on Leaf1 and Leaf2 as well as Spine1 and Spine2 to ensure Layer 3 network connectivity.
1	Enable the NVO3 ACL extension function on Spine
2	Configure multi-active gateway on Spine1 and Spine 2 by configuring DFS group.
3	Configure VXLAN on Leaf1 and Leaf2 as well as Spine1 and Spine2 to construct a large Layer 2 VXLAN network over the basic Layer 3 network.
4	Configure service access points on Leaf1 to Leaf2 to distinguish traffic from servers and forward the traffic to the VXLAN network.
5	Configure VXLAN Layer 3 gateways on Spine1 and Spine2 to implement communication between VXLAN networks on different network segments and between VXLAN and non-VXLAN networks.

HUAWEI

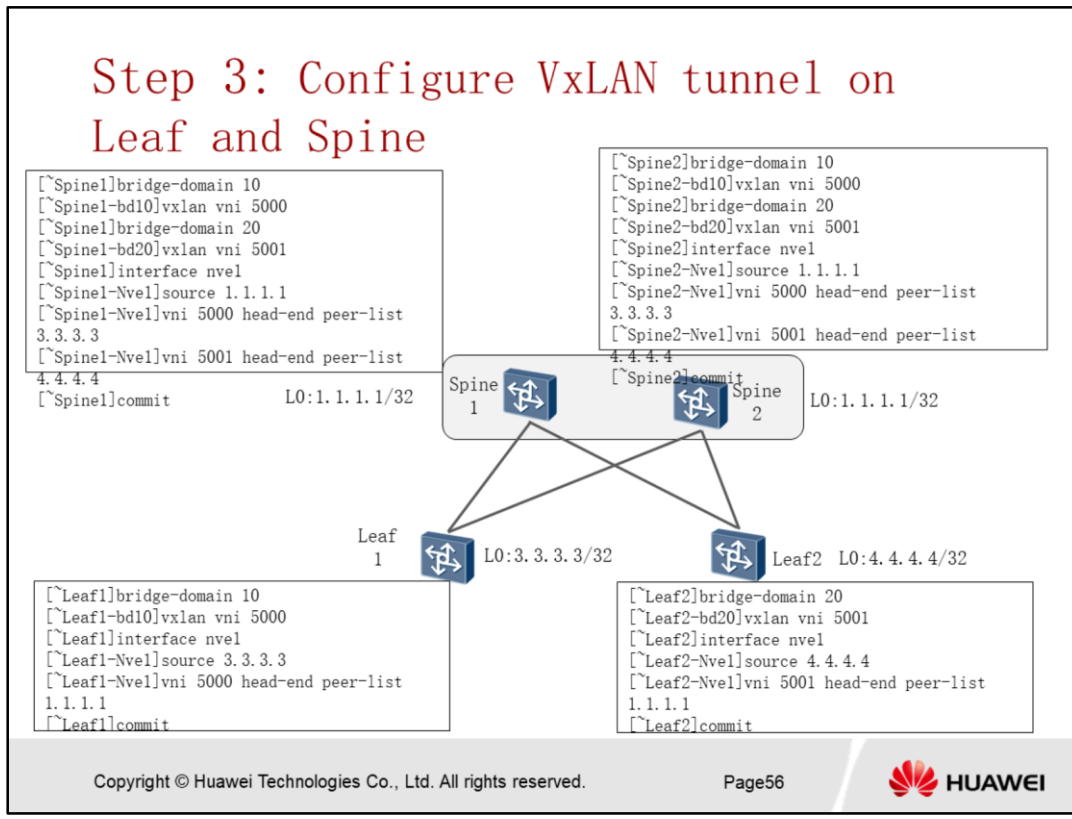
- Step 1 OSPF configuration is omitted.



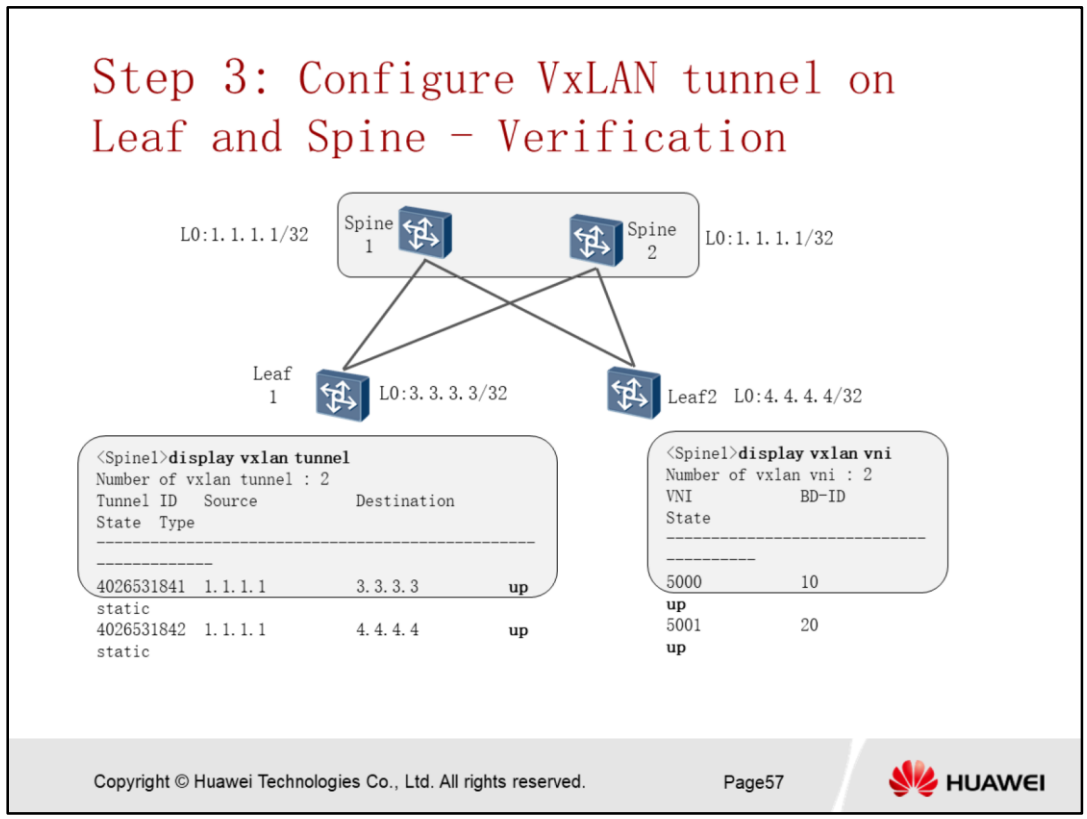
- NOTE: After modifying the tunnel mode or enabling the NVO3 ACL extension function, you need to save the configuration and restart the device to make the configuration take effect. You can restart the device immediately or after completing all the configurations.



- After the configuration is complete, run the **display dfs-group 1 active-active-gateway** command on Spine1 and Spine2. If the state is shown “A:active”, it means that the multi-active gateway connection is established.

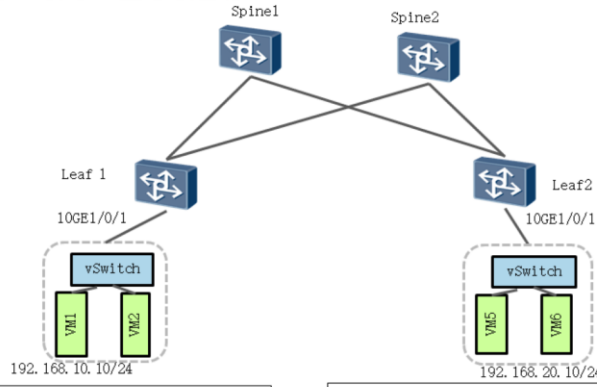


- As Spine 1 and spine2 is working in multi-active DFS group, the loopback 0 configured on both spines must be the same IP. Leaf will see them as 1 device.



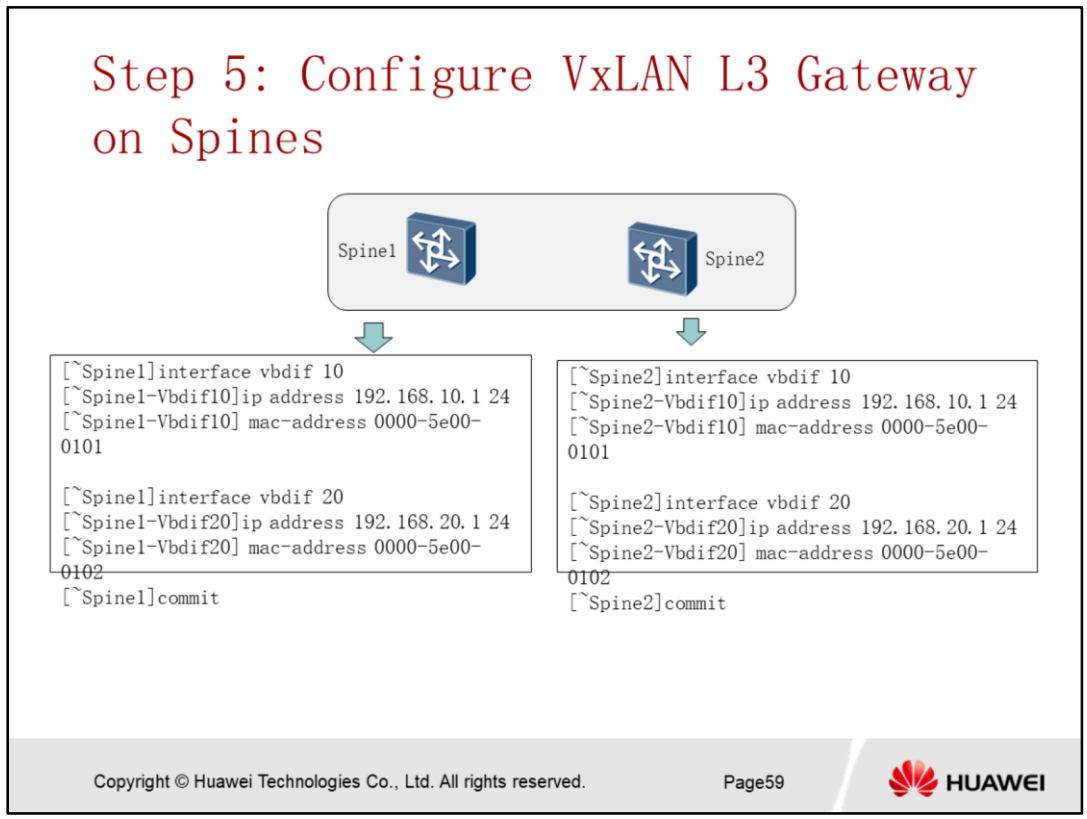
- Example above shows the verification done on Spine1.
- After VXLAN tunnels are established, run the **display vxlan tunnel** command to check tunnel information
- After a VXLAN is configured, to check the VNI status and BD to which the VNI is mapped, run the **display vxlan vni** command. The command output helps you determine whether the VXLAN is correctly configured.

Step 4: Configure Service Access Points at Leafs

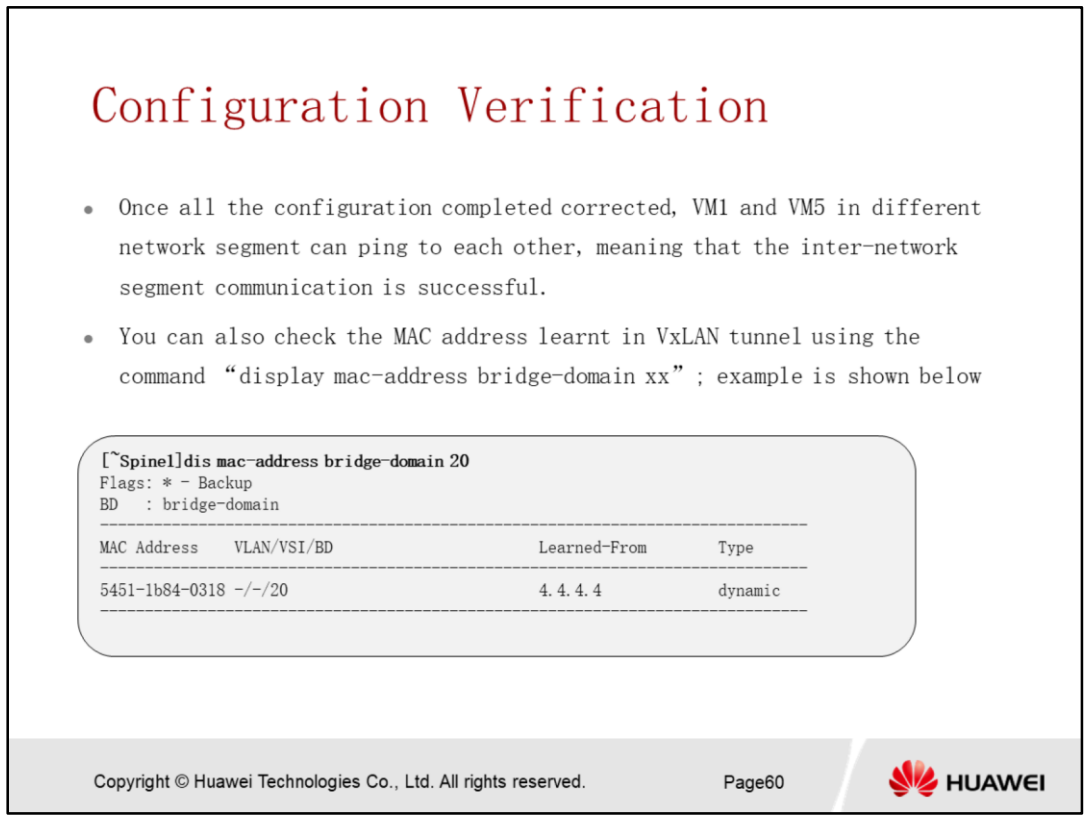



```
[~Leaf1]vlan 10
[~Leaf1]interface 10ge1/0/1
[~Leaf1-10GE1/0/1]port link-type trunk
[~Leaf1-10GE1/0/1]undo port trunk allow-pass vlan
1
[~Leaf1-10GE1/0/1]port trunk allow-pass vlan 10
[~Leaf1]bridge-domain 10
[~Leaf1-bd10]l2 binding vlan 10
[~Leaf1]commit
```

```
[~Leaf2]vlan 10
[~Leaf2]interface 10ge1/0/1
[~Leaf2-10GE1/0/1]port link-type trunk
[~Leaf2-10GE1/0/1]undo port trunk allow-pass vlan
1
[~Leaf2-10GE1/0/1]port trunk allow-pass vlan 20
[~Leaf2]bridge-domain 20
[~Leaf2-bd20]l2 binding vlan 20
[~Leaf2]commit
```



- The configuration on Spine 1 and 2 must be same because they are working in active-active gateway group.



- The **display mac-address bridge-domain** command displays MAC address entries in a specified bridge domain (BD).

Summary

- As a summary for this topic, we have covered:
  1. VxLAN overview including how VxLAN solves issues in traditional DCN networking
  2. VxLAN basic concepts, terms and forwarding models.
  3. VxLAN application in SDN AC-DCN network including VM communication and fabric network
  4. VxLAN configuration examples in SDN AC-DCN Cloud Fabric Network